The Value of Shared Visual Information for Task-Oriented Collaboration

Darren R. Gergle
August 2006
CMU-HCII-06-106

Human-Computer Interaction Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213

Thesis Committee:
Robert E. Kraut (Chair), Carnegie Mellon University
Susan R. Fussell, Carnegie Mellon University
Carolyn P. Rosé, Carnegie Mellon University
Susan E. Brennan, Stony Brook University

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

This work was supported in part by the National Science Foundation under grants IIS #99-80013 and DST #02-08903, and by an IBM Ph.D. Fellowship. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author and do not necessarily reflect those of the funding agencies.

Keywords: Shared visual information, shared visual space, computer-mediated communication, distance collaboration, computer-supported cooperative work, computer-supported collaborative work, collaborative computing, empirical studies, discourse analysis, language use, computational modeling, rule-based computational model, corpus evaluation, pronoun resolution, reference, visual delay, field-of-view, visual salience, linguistic salience, multivariate adaptive regression splines, MARS, sequential analysis, grounding theory, situation awareness theory, centering theory, task awareness, conversational grounding, experimentation, human factors, human performance, and group performance.

Abstract

For several decades, researchers and engineers have struggled with the development of systems to support distance collaboration. The failure of many collaborative technologies is due, in part, to a limited understanding of how groups coordinate in collocated environments and how the coordination mechanisms of face-to-face collaboration are impacted by technology. The major goal of this thesis is to address this deficiency by building a theoretical understanding of the role that shared visual information plays in supporting group communication and performance during task-oriented collaboration. This understanding is developed over three major stages: (1) the development of a paradigm and a series of empirical studies that decompose the features of shared visual information and task structure and explore their interactions in detail, (2) the development and application of a methodology for describing the sequential structure of how visible actions support the understanding of discourse, and (3) the development of a computational model of discourse to further our theoretical understanding of the ways in which shared visual information serves communication in task-oriented collaborative discourse.

Acknowledgments

First and foremost, I would like to thank Bob Kraut for his remarkable thoughtfulness, support, and advice in matters of the academy as well as everyday life. Throughout my tenure as a doctoral student, it was reassuring to know that I could rely on such a brilliant, insightful and gifted mentor. It has been a genuine pleasure.

I would also like to thank my committee members. Sue Fussell has been a tremendous mentor, and I was fortunate to have her serve in a role that is best described as co-advisor. Her boundless energy, shrewd insight, and sympathetic spirit provided me with a great deal of support. My work would be noticeably impoverished without her contributions. Carolyn Rosé introduced me to a new discipline and served as an incredible teacher and resource. She was exceedingly generous with her time and her thoughts, and provided a level of support far exceeds the expected contributions of a committee member. Finally, Susan Brennan provided a refreshing outside perspective on my work. Her expertise was invaluable, and her research innovations and genuine brilliance served as a major source of inspiration. Together, this collection of researchers provided me with astonishing resources and a memorable experience.

I would also like to express a special thanks to Donna Byron and Joel Tetreault for their valuable feedback and support on the modeling portion of this thesis. They each truly encapsulate the meaning of mentor and scholar, and my work benefited greatly from discussions with them. In addition, this work would not have been possible without the hard work and support of several research assistants over the years: Matthew Hockenberry, Rachel Wu, Katelyn Shearer, Gregory Li, Megan Branning, Sajiv Shrivastva, and Lisa Auslander.

A number of other colleagues have contributed to my work and life in the past few years, including: Anne H. Anderson, Roger Bakeman, Ryan Baker, Aaron Bauer, Laura Dabbish, James

Fogarty, Carl Gutwin, Jim Herbsleb, Gary Hsieh, Scott Hudson, Sara Kiesler, Adam Kramer, Gail Kusbit, David E. Millen, Bilge Mutlu, Jeffrey Nichols, Jiazhi Ou, Vincent Quera, Peter Scupelli, A. Fleming Seay, Irina Shklovski, Jane Siegel, Cristen Torrey, Joe Walther, Jacob O. Wobbrock and Jie Yang. I would also like to offer a special thanks to Thi and Daniel Avrahami, not only for their support in my academic endeavors, but also for welcoming me as family when mine was out of reach. Thanks to Charlie, Wally, Anthony, Harry and the rest of the Jitters crew for keeping me caffeinated and happy over the last five years.

My work would not have been possible without the tutelage and inspiration of a number of teachers and professors I have had contact with throughout my academic career: Tom Brinck, George Furnas, David E. Meyer, and Priti Shah all played a central role in my development. Another major source of inspiration in my academic life has been provided by Judy Olson. She is a model researcher whom I hold in the highest regard, both for the manner in which she approaches her research, as well as the way in which she approaches life. Thank you for the lessons.

Finally, I could not have done this without the enduring love and support of my parents Bob and Barb, my sister Tanya, my brother Jim, my grandmother Ruth, and my greatest source of inspiration and companionship, my wife Tracy.

To my parents,

Robert G. and Barbara K. Gergle,

for a lifetime of love and support.

Table of Contents

List of Figures

List of Tables

List of Reproduced Publications

The following presents a list of published works that constitute, in part or in whole, a portion of this thesis work.

Gergle, D., Kraut, R. E., & Fussell, S. R. (2006). The Impact of Delayed Visual Feedback on Collaborative Performance. In Proceedings of the *ACM Conference on Human Factors in Computing Systems (CHI 2006)*, pp. 1303-1312. NY: ACM Press.

Gergle, D. (2006). What's There to Talk About? A Multi-Modal Model of Referring Behavior in the Presence of Shared Visual Information. In Proceedings of *European Chapter of the Association for Computational Linguistics (EACL 2006) Conference Companion*, pp. 7-14.

Gergle, D. (2005). The Value of Shared Visual Space for Collaborative Physical Tasks. In Proceedings of the *ACM Conference on Human Factors in Computing Systems (CHI 2005), Extended Abstracts*, pp. 1116-1117. NY: ACM Press.

Fussell, S. R., Kraut, R. E., Gergle, D., and Setlock, L. D. (2005). Visual Cues as Evidence of Others' Minds in Collaborative Physical Tasks. In B. Malle and S. Hodges (Eds.), *Other Minds* (pp. 91-105). NY: The Guilford Press.

Gergle, D., Kraut, R. E., & Fussell, S. R. (2004). Action as language in a shared visual space. In Proceedings of the *ACM Conference on Computer Supported Cooperative Work (CSCW 2004)*, pp. 487-496. NY: ACM Press.

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*, 491-517.

Gergle, D., Millen, D. E., Kraut, R. E., & Fussell, S. R. (2004). Persistence matters: Making the most of chat in tightly-coupled work. In Proceedings of the *ACM Conference on Human Factors in Computing Systems (CHI 2004)*, pp. 431-438. NY: ACM Press.

Kraut, R. E., Gergle, D., & Fussell, S. R. (2002). The Use of Visual Information in Shared Visual Spaces: Informing the Development of Virtual Co-Presence. In Proceedings of the *ACM Conference on Computer Supported Cooperative Work (CSCW 2002)*, pp. 31-40. NY: ACM Press.

Chapter 1

Introduction

In recent years, structural changes to organizations, such as the rise of large multinational corporations, coupled with technological advances, such as the widespread availability of the Internet, have contributed to increases in distributed work practices mediated by telecommunication technologies. In this time, there has been a growing interest in the design of technologies to support a host of remote collaboration activities such as architectural planning, telesurgery, and remote repair tasks. These activities, when performed in a collocated environment, rely on a number of intricate dependencies between verbal communication and physical actions. However, when designing tools and technologies to support such tasks remotely, we need to understand how the introduction of technological mediation impacts the coordination mechanisms typically relied upon in collocated physical environments.

Consider the following scenarios. An automotive design team develops a 3D model for a new chassis; however, the materials processing engineer is located in Detroit while the structural engineer is in Stuttgart. A team of surgeons performs an operation while a world-renown expert monitors the progress from her office on the opposite coast. An architecture student gets timely help on his mechanical simulation from an engineering tutor across campus. These scenarios are examples of a distributed collaborative task in which at least one person is physically remote from the primary site. However, the literature suggests that such activities are often more difficult and less successful than comparable work in collocated settings (for reviews see Olson & Olson, 2000; Whittaker, 2003). Part of this problem stems from a lack of understanding of how groups

coordinate their activities in real world collocated environments and how the coordination mechanisms of face-to-face collaboration are affected by technology. It is a goal of this thesis to remedy this gap in knowledge by exploring a mechanism often thought to play a critical role in supporting coordination: *shared visual information*.

1.1 Background

Many researchers hypothesize that visual information plays a central role in coordinating collaborative work. While early research posited that seeing other people's faces during conversation was critical for successful coordination (Daft & Lengel, 1986; Short *et al.*, 1976), many empirical studies failed to support this claim (see Nardi & Whittaker, 2002; Williams, 1977 for reviews). In particular, studies on the effect of video-mediated communication systems found that video of the participants' faces and upper bodies provided little additional benefit over the presentation of audio (cf. Veinott *et al.*, 1999; for a review see Williams, 1977). More recently, researchers have shifted their focus to the use of video and visual information in support of dynamic information about the tasks, objects and events that serve collaboration in a visual environment (Kraut *et al.*, 2003; Monk & Watts, 2000; Nardi *et al.*, 1993; Whittaker *et al.*, 1993; Whittaker & O'Conaill, 1997). This approach has identified a range of conditions under which visual information is valuable. 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 a remote helper has of a co-worker's environment influences performance and shapes communication patterns in directed physical tasks (Fussell *et al.*, 2003a).

Yet, as described in several recent reviews (Whittaker, 2003; Whittaker & O'Conaill, 1997), a more nuanced theoretical understanding of the precise functions visible information serves in collaboration is required. How, for example, does seeing a partner's actions alter a person's speech? How does a small field of view affect the ability of pairs to plan subsequent actions? How do delays in the shared view affect grounding processes that rely on temporal precision? How is the generation and comprehension of referring expressions impacted by the availability of shared visual information? A major goal of this thesis is to answer these questions through the development of a detailed theoretical understanding of precisely how shared visual information serves collaboration.

1.2 Thesis overview

The following three stages serve as the basis for the development of a more detailed theoretical understanding of the role of shared visual information in task-oriented collaborations.

Stage I: Empirical Studies of Shared Visual Information. The first stage of this thesis is primarily interested in addressing the question, "Is shared visual information useful?" This stage consists of a theory-based empirical methodology and a coinciding series of rigorously controlled laboratory experiments that decompose the features of shared visual information and examine their influence on communication processes. A primary goal of this portion of the work is to establish quantitative measurements that reflect the benefits of providing access to shared visual information for pairs involved in tightly-coordinated collaborative tasks. A detailed description of the experimental paradigm used in this work is presented in Chapter 2, and the experimental laboratory studies are described in Chapters 3 – 5.

Stage II: Sequential Analyses of Shared Visual Information. The goal of the second stage of this thesis is to answer the question, "Where is the shared visual information useful?" This work involves the application of sequential analysis techniques to provide insight into where in the overall course of the collaborative activity visual information is useful. This methodology supports the investigation of how visible actions support understanding in the discourse and allows detailed statistical examination of the patterns of language use and actions that lead to successful collaborative performance. A detailed description of this stage is provided in Chapter 6.

Stage III: A Rule-Based Computational Model of Shared Visual Information. The results of Stage I and II, as well as prior literature, suggest that a primary area of impact that shared visual information has is on the ability of pairs to efficiently and effectively make use of it to resolve ambiguity and generate efficient referring expressions. It is the goal of this phase of the thesis to answer the question, "How is the visual information useful?" This stage develops a computational model that precisely details how visual information is combined with linguistic cues to enable effective reference-making during tightly-coupled task-oriented collaborations. This work continues the theoretical development from the first two stages that describes how visual information influences language use by expressing this understanding computationally. This stage of work is described in detail in Chapters 7 and 8.

1.3 Research approach and impact

The general approach to this work is to start by understanding—at a broad level—the wide variety of visual factors hypothesized to contribute to successful communication and collaboration. From there, the thesis undertakes a more thorough examination of the process level details of communication and investigates how various forms of visual information impact collaboration. Finally, the thesis presents a detailed and computationally explicit theory of the ways in which visual and linguistic information interact to impact collaborative communication, in the form of a rule-based computational model of referring behavior.

An understanding across these areas impacts the fields of Human-Computer Interaction (HCI) and Computer-Supported Cooperative Work (CSCW) at both theoretical and applied levels. At a theoretical level, it leads to an improved understanding of how features of tasks and media, both alone and in combination, affect communication and coordination. It adds to our knowledge of how task features influence people's use of visual space, and how language and actions are coordinated in team performance. The methodological contributions are primarily in the area of preparing and analyzing behavioral data from multiple parties with multiple channels of expression.

There are also several practical applications of this work. As the opening scenarios illustrate, distributed tasks play important practical roles in medical, educational, and industrial domains. This research builds a theoretical framework that will help maximize the fit between technologies and tasks in these and other critical domains. The findings aim to benefit the public by allowing us to identify technologies that enable specialists to work remotely to the best of their capabilities, and by providing a detailed understanding of how to design new technologies that allow greater numbers of individuals to participate in these domains from a distance. The ultimate goal of this work is to provide a foundation and rationale for the future development, design and deployment of systems to support distributed collaborative physical tasks.

Chapter 2

Theoretical and Experimental Framework

The first stage of this dissertation addresses the question of whether shared visual information, in a variety of forms, facilitates communication and coordination during task-oriented collaborations. However, before doing so, we must first understand how people use specific types of visual evidence for collaborative purposes. This chapter introduces the general theoretical motivation for this work and is followed by a detailed description of the experimental paradigm used throughout the studies.

2.1 Theoretical background

Two theories that provide insight into the impact of shared visual information on collaborative performance are *Grounding Theory* (Clark & Marshall, 1981; Clark & Wilkes-Gibbs, 1986) and *Situation Awareness Theory* (Endsley, 1995; Endsley & Garland, 2000). According to Grounding Theory, visual information provides a means for coordinating language and generating efficient and understandable discourse surrounding a collaborative activity. Visual information also provides evidence of what people are aware of and therefore facilitates the generation, validation, and comprehension of language in conversations based on this knowledge. Situation Awareness has a slightly different focus. It centers primarily on how visual information influences the ability of groups to formulate a common representation of the task state, which in turn allows them to plan and act appropriately. Together these two theories describe the central components required of shared visual information in order to support collaborative activities. The remainder of this section presents a brief introduction to these mechanisms, which will be explored in detail in the following chapters.

2.1.1 Visual information in support of grounding

Grounding Theory states that successful communication relies on a foundation of mutual knowledge or common ground. Visual information can support the formation of some of this mutual knowledge, and thereby improve the conversation surrounding a collaborative task. The process of establishing common ground is what is referred to as grounding or the grounding process.

Throughout a conversation, participants continually assess their degree of shared knowledge and use this to form subsequent utterances (Brennan, 1990; Clark & Marshall, 1981; Clark & Wilkes-Gibbs, 1986). As conversational partners discuss something, they provide evidence of their understanding. This evidence can be exhibited in several ways. In a typical spoken interaction, partners can use explicit verbal statements (e.g., "I got it" or "do you mean the red one?") or back-channel responses (e.g., "uh-huh") to indicate comprehension. Evidence can also be provided through a variety of environmental and social factors. Differences in spatial orientation (Schober, 1993), levels of domain expertise (Isaacs & Clark, 1987), and socio-cultural background (Fussell & Krauss, 1992), have all been shown to shape the effectiveness and fluidity of the grounding process. In environments where visual information is available, the visual feedback itself can be a critical resource for grounding (Brennan, 1990; Kraut et al., 2003).

The work presented in this thesis addresses the central question of how various forms of visual information—particularly those commonly impinged upon by technologies to support remote collaboration—can affect the grounding process. Shared visual information helps conversational partners establish common ground by providing evidence from which to infer another's level of understanding. This evidence can often be deliberate (e.g., as in a pointing gesture) or as a side effect of proper performance of the desired action provided both parties are aware of what the other can see. When a speaker instructs an actor, the actor's performance of the correct action without any verbal communication provides an indication of understanding, while performing the wrong action or even failing to act can signal misunderstanding. In each of these cases, shared visual information plays a crucial role in supporting joint activities by reinforcing the grounding process.

2.1.2 Visual information in support of situation awareness

Visual information can also be valuable for coordinating the task itself. According to Situation Awareness Theory successful collaboration requires group members to maintain awareness of

one another's activities, the status of relevant task objects, and the overall state of the collaborative task (Endsley, 1995; Endsley & Garland, 2000). Situation Awareness Theory aims to capture this by integrating a representation of the current environmental status with a general procedural model of the task.

Visual information supports the formation and maintenance of situation awareness by providing an up-to-date representation of the state of the task and the activities of the partners. This in turn allows group members to plan the next steps toward achieving the task goal, determines what instructions they need to give, and provides a means by which to repair incorrect actions. Nardi and colleagues (1993) describe how a scrub nurse on a surgical team might use visual information about task state to anticipate what instruments the surgeon will need. For instance, if the scrub nurse notices that the surgeon nicks some flesh, she can prepare cauterization and suture materials and have them ready before the surgeon asks for them. The situation awareness needed to facilitate such actions is provided by the availability of a shared visual environment.

In order for visual information to support task awareness and improve collaborative performance, the display itself does not need to be identical for all group members, as long as it allows them to form an accurate view of the current situation and appropriately plan future actions (Bolstad & Endsley, 1999). For example, two fighter pilots can converge on and shoot down another aircraft, even if one of them uses visual line of sight and the other uses radar to "see" the target. However, if the differing displays lead them to form different situational representations, their performance is likely to suffer. For example, if visual sighting allows a pilot to distinguish between friendly and enemy aircraft, but the radar fails to support this discrimination, then the two fighters are unlikely to successfully coordinate their attack purely on the basis of the situation awareness provided by the visual information.

2.1.3 The impact of technological-mediation on the availability of visual information

Although shared visual information will likely improve collaborative task performance by improving situational awareness and grounding, the benefits it provides are apt to depend, in part, on the particular features of the technology and the particular characteristics of the collaborative task. For many engineers and designers developing technologies to provide visual information in distributed settings, the goal is to make a collaborative environment as similar as possible to the gold standard of physical co-presence. In attempting to reach this goal, however, engineers often

must sacrifice technological features that impact the usefulness of the visual information, such as the size of the field of view and who controls it, tolerance for delays, degree of spatial resolution, frame rate, and synchronization with a voice stream. Clark and Brennan (1991) hypothesized that different communication media have features that change the cost of grounding. How do we know which of these features need to be reproduced in order to recreate the benefits of a collocated environment? Is it better to sacrifice field of view for faster visual updates? Are aligned views of a workspace required for efficient performance? Do particular task features depend more or less on the availability of shared visual information?

To investigate these questions, I apply a collaborative online jigsaw puzzle task that can be used to collect data in a controlled laboratory environment (Gergle *et al.*, 2004a, 2004b; Gergle *et al.*, 2004c; Kraut *et al.*, 2002b). This paradigm provides a method for decomposing the visual space in order to better understand how various forms of shared visual information can impact collaborative performance. It also facilitates the collection of quantitative measures and permits a detailed examination of the role played by various technological features, the associated role of task features, and their impact on the hypothesized coordination mechanisms of grounding and situation awareness. This work unites with recent studies to describe the central role shared visual information plays in collaborative task performance (see also Brennan & Lockridge, In preparation; Clark & Krych, 2004).

2.2 Overview of the puzzle study paradigm

The puzzle study paradigm is a referential communication task (Krauss & Weinheimer, 1964, 1966) where a Helper describes a configuration of puzzle pieces to a Worker, who then needs to assemble the puzzle to match the goal state. This task falls into a general category of "mentoring" collaborative physical tasks, in which one person manipulates objects under the guidance of another who usually has greater expertise or knowledge about the task (Kraut et al., 2003).

2.2.1 The puzzle study task

In this task, one participant (the "Helper") instructs another participant (the "Worker") on how to complete a puzzle consisting of four blocks selected from a larger set of eight blocks. The goal is to have the Worker correctly place the four blocks in the proper arrangement in the shortest amount of time so that they match the target solution the Helper is viewing. It is up to the Helper to describe the goal state to the Worker and guide her towards the correct solution.

Figure 2-1 demonstrates a standard view of the screen from the Worker's side (left) and Helper's side (right). The Worker's screen consists of a staging area on the right hand side in which the puzzle pieces are shown, and a work area on the left hand side in which she constructs the puzzle. The Helper's screen shows the target solution on the right and a view (if available) of the Worker's work area on the left. The Helper's view of the Worker's work area can be manipulated in a number of ways to investigate how different features of shared visual information affect communication. For example, the computational implementation of the task allows us to manipulate with a high degree of specificity how much overlap exists between the Helper and Worker views of the workspace. The views between the two displays can be rotated, delayed, or a subset of the work area can be shown. Similarly, the task features can be manipulated by introducing rapidly changing task objects, lexically complex objects (e.g., plaid blocks), or the visual complexity can be manipulated by overlapping objects in the target area.

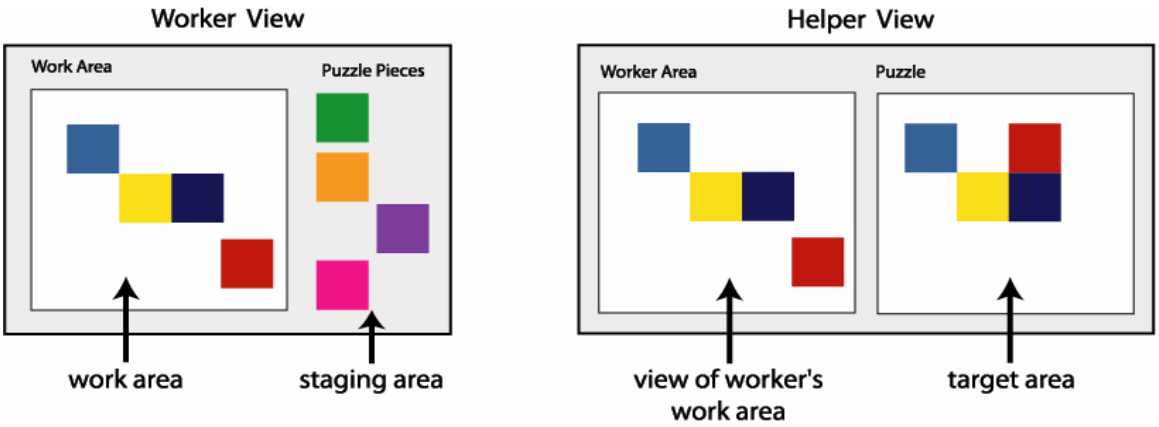


Figure 2-1. The Worker's view (left) and the Helper's view (right).

2.2.2 Collection of empirical studies

I have used this task paradigm to investigate a number of parameterizations of shared visual information and task features. Table 2-1 presents an overview of the studies described in this thesis that investigate different parameters of shared visual information (Gergle et al., 2004a, 2004b, 2006, Under Review; Gergle et al., 2004c; Kraut et al., 2002b).

Table 2-1. Collection of studies using the puzzle paradigm.

Study	Shared Visual Information Features				Task Features		
	Visual Immediacy	*Viewspace Alignment*	*Viewspace Size*	*Viewspace Control*	*Lexical Complexity*	*Temporal Complexity*	*Visual Complexity*
Chapters 3, 6 and 8							
Study 1: Drift Study	**X**					**X**	**X**
Chapters 4 and 8							
Study 2: Continuous Delay Study	**X**					**X**	
Study 3: Continuous Delay / Continuous Drift Study	**X**					**X**	
Chapters 5 and 8							
Study 4: Plaid Study	**X**				**X**		**X**
Study 5: Rotation Study	**X**	**X**			**X**		
Study 6: Viewspace Study			**X**	**X**	**X**		
Chapter 8							
Study 7: Communication Channels Study	**X**				**X**		
Study 8: Chat Persistence Study	**X**				**X**		

2.3 Dissertation organization

In the first two chapters I presented an overview of the thesis topic, the theoretical framework that guides this work, and the experimental paradigm that serves as the foundation for exploring the value of shared visual information for collaborative task performance. The following chapters present a number of studies, evaluations and models that establish a deeper theoretical understanding of the role played by shared visual information in collaborative task performance.

In **Chapter 3,** I present the first study of the thesis and lay the theoretical groundwork for the remaining chapters. The work presented in this chapter serves as a survey study that describes a number of theoretical phenomena and illustrates a range of dependent measurements such as task performance, behavioral patterns, and communicative adaptations that occur when shared visual information is available. It also explores the impact of a delay in the shared visual feedback on task performance and communication patterns.

Chapter 4 is a follow-up study that more closely examines how delay in shared visual information impacts collaborative performance. In particular, this chapter details two studies that examine the form of the function that governs the relationship between visual delay and collaborative task performance at a much finer level of temporal resolution than has been explored in prior studies. The first study precisely demonstrates how a range of visual delays differentially impact performance and illustrates the collaborative strategies employed. The second study describes the ways in which task parameters, such as the dynamics of the objects in the environment, affect the amount of delay that can be tolerated.

The goal of **Chapter 5** is to make a theoretical distinction between the proposed mechanisms that play a role in supporting collaboration when shared visual information is available. In the previous studies, the claim is made that shared visual information supports communication and performance by helping to maintain situation awareness and by supporting conversational grounding. While these are theoretically distinct mechanisms, they are often conflated in research. This chapter presents a series of three studies that empirically isolate the two major mechanisms and describe the independent contributions made by each. The first is a replication study that establishes baseline behavior and illustrates the potential conflations. The second and third studies demonstrate the independent effect of shared visual information on situation awareness and conversational grounding.

Chapter 6 presents evidence from an empirical study that demonstrates how visible actions replace explicit verbal instructions of similar communicative content when shared visual information is made available. This work begins to develop our understanding of how visible actions interact with language, and demonstrate that in order to successfully understand language use in task-oriented collaborations we need to account for both visual and linguistic information. In doing so, it forms the motivation for the remaining chapters which describe the development of a computational model of discourse in the presence of shared visual information.

Chapter 7 describes the development and evaluation of a rule-based computational model that characterizes referring behaviors in the presence of shared visual information. This work demonstrates how a feature-based representation of shared visual information combines with linguistic cues to enable effective pronominal reference. This work continues the development of a theory that describes how shared visual information impacts language use and collaboration. However, this understanding is now expressed computationally, and while it looks at a smaller portion of the task, in particular referring expressions, it provides a much more explicit and detailed description of how this occurs in the presence or absence of shared visual information.

Chapter 8 details the development and evaluation of the computational model. In particular, this chapter presents an empirical evaluation that examines the performance of three hypothesized models of reference resolution using a corpus-based evaluation. The three models consist of a language-only model, a visual-only model, and an integrated model of reference resolution. The results demonstrate that the integrated model significantly outperforms both the language-only model and the visual-only model as a model of reference resolution.

Finally, **Chapter 9** summarizes the work and contributions presented throughout this dissertation and discussion potential avenues for future work.

Chapter 3

The Impact of Shared Visual Information on Collaborative Performance[1]

When collaborators work together on a physical task, seeing a common workspace transforms their language use and reduces their overall collaborative effort. This chapter demonstrates how shared visual information can be used to make communication and collaboration more effective and efficient. Using the puzzle study paradigm, pairs of participants communicated without a shared visual space, using a shared space featuring immediately updated visual information, and using a shared space featuring delayed visual updating. Having the shared visual space helped collaborators understand the current state of their task and enabled them to ground their conversations efficiently, as seen in the ways in which participants adapted their discourse processes to their level of shared visual information. These processes were associated with faster and better task performance. Delaying the visual update reduced the benefits and degraded performance. The shared visual space was more useful when tasks were visually complex or when participants lacked a simple vocabulary to describe their environment.

---

[1] The work presented in this chapter was originally published in Kraut, R. E., Gergle, D., & Fussell, S. R. (2002). The Use of Visual Information in Shared Visual Spaces: Informing the Development of Virtual Co-Presence. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW 2002)*, pp. 31-40. NY: ACM Press; and in 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,* 491-517.

3.1 Introduction

Consider an architect and client working side-by-side to discuss architectural plans for a new corporate headquarters. Communication between them does not merely consist of the words they exchange, produced independently and presented for others to hear. Rather, speakers and addressees integrate and take into account what one another can see (Schober, 1993; Schober & Clark, 1989). They notice where the other's attention is focused (Argyle & Cook, 1976; Boyle *et al.*, 1994; Fussell *et al.*, 2003b), point to objects and use deictic references like "that one" and "there" (Barnard *et al.*, 1996), demonstrate and manipulate objects (Clark & Krych, 2004), make hand gestures, eye contact, facial expressions, and reference prior discourse and behavioral actions. Many of these processes take advantage of shared visual information. Using visual information to infer what another person knows facilitates communication and reduces the ambiguity otherwise associated with particular linguistic expressions.

Shared visual information can be an extremely efficient collaboration mechanism, particularly when behaviors and actions are linguistically complex. As pairs attempt to communicate, the visual information provided in a shared visual workspace can be used in several ways to minimize the overall level of joint effort required. It also serves as a precise indicator of comprehension and may be used to provide situational awareness in regard to the overall state of a joint task. Although these communicative techniques are often critical to successful interaction in the everyday world, technologies designed to support communication at a distance often fail to support them adequately.

A shared visual space occurs when the architect and client are collocated and gathered around the table, looking at architectural plans. It can also occur through technological mediation, for example, when distant collaborators jointly look at documents on yoked computer screens. In either case, a shared visual space enables people to jointly view approximately the same objects at approximately the same time. Designers have many choices about how to technologically construct a shared visual space. For example, they can choose which images are transmitted (e.g., the users or the objects being discussed), the orientation of the images, refresh rates, or the levels of detail that are transmitted between the communicators. As described in Chapter 5, how these decisions are made can be informed by Grounding Theory. Grounding phenomena shape the language and understandings that communicators exchange.

This chapter has two major goals. First, it is designed to examine how a shared visual workspace influences communication in a collaborative work task. The second research goal is to examine how a shared visual space that supports effective communication should be designed.

3.2 Background

Most of the early research examining the utility of visual information in communication focused on the degree to which collaborators were aware of one another, at the expense of visual information about the objects they discussed. This research tradition is derived from work conducted by the Communications Study Group at British Telecom (Short *et al.*, 1976) and in Chapanis' lab in the United States (Chapanis *et al.*, 1972). Studies compared dyads performing a referential communication task (i.e., a task where a speaker communicates information about objects, pictures, directions, etc.) using only an audio channel to dyads performing the same task face-to-face or using an audio/video connection. This research concluded that visual information from a partner's face provides little support for typical referential communication.

More recent research shifts the focus from a view of the participants' faces to a view of the work area. One line of research using realistic work tasks in this new wave has uniformly found that participants in side-by-side settings, in which they share full views of one another and the workspace, perform better than participants using a variety of other communications arrangements (Fussell *et al.*, 2004; Kraut et al., 2003; Nardi et al., 1993).

However, results were initially mixed when the research used video to create the shared visual space. For example, Fussell, Kraut, and Siegel (2000) had "worker" and "expert" dyads repair a bicycle while conversing side-by-side, using audio plus a head-mounted camera transmitting the worker's view of the bicycle to the remote expert, or via audio only. Pairs performed substantially faster when they worked side-by-side than in the audio condition. Although dyads used different techniques to refer to objects in the video-mediated condition than in the audio condition, their overall performance time was no better. In contrast, Fussell, Setlock, and Kraut (2003a) found that pairs performed better when they used video tools that provided views of the workspace than when they used audio or text-based communication alone.

The differences among video configurations may have led to conflicting results. For example, in Fussell, Setlock, and Kraut (2003a), remote communicators could make visible gestures in the video image, whereas in Fussell et al. (2000) they could not. Differences in the quality of the

implementation may also have accounted for different results. For example, in Fussell et al. (2000), technical complications with the field of view, video transmission, and slippage of the camera on the worker's head may have rendered the video-mediated shared visual space inadequate. Thus, there is a need for more tightly controlled laboratory studies of shared visual space to complement these previous efforts.

To address these issues, a second line of work has been exploring more stylized communication tasks in tightly controlled laboratory environments. For example, Clark and Krych (2004) used a stylized communication task in which one participant, a Director, instructed another, a Matcher, on how to construct a simple LEGO® form. When the Director could see what the Matcher was doing, the pair was substantially faster, in part because the pair could precisely time their words to the actions they were performing. Although this work provided initial insight into the ways in which shared visual space led to more efficient conversation, it did not detail the exact mechanisms by which the improvement occurred. Consider the nature of a shared visual space when people are working side-by-side: Voice is synchronized to actions, the parties are mobile, both parties can point to objects in space, each party can see both the work area and each other's face and gestures, and each party sees the workspace from a slightly different angle. Which of these features of the side-by-side setting need to be reproduced to recreate the benefits of proximity through technology-mediated communication? The puzzle study paradigm was developed to address these issues.

3.3 Study 1: The impact of shared visual information on collaborative performance

The study reported here uses the puzzle study paradigm to disaggregate the features of a shared visual space and to observe their effects on performance. The basic methods were described in Chapter 2 and this paradigm was applied to examine how shared visual information (whether the Helper could see the shared visual space) and one of its attributes (the speed with which the shared visual information is updated) interacts with two task attributes (visual complexity and temporal dynamics) to affect communication processes and task performance. Access to shared visual information was expected to be more important for tasks involving difficult-to-describe puzzles or tasks in which the environment rapidly changed. In addition, delays in updating the shared visual information should degrade its usefulness. Krauss and Bricker (1967) had previously shown that auditory delays as small as 250ms could affect both communication

process and efficiency. Do delays in updating a shared visual space, of the sort produced by network congestion and video compression, cause similar problems?

3.3.1 Identifying the critical elements of shared visual information

To identify the important elements of shared visual information—as alluded to in the introductory chapters—we must first understand how people use specific types of visual evidence for collaborative purposes. Clark and Wilkes-Gibbs (1986) observed that collaborative work occurs at multiple levels simultaneously, although the distinction between levels is not crisp. At the highest level, people collaborate on performing the task. In this experiment, they are jointly solving a puzzle. At a lower level, they use language and other communicative behaviors to coordinate actions in order to perform the task. At yet a lower level, pairs use communicative behaviors to coordinate the language they use. For example, pairs jointly determine the names they apply to pieces in the puzzle or indicate whether they understood a description. Visual evidence can be helpful at each of these levels. It can inform the Helper about the next puzzle action that the Worker needs to perform by giving an up-to-date account of the overall state of the task. It can guide the Helper in planning an instruction by indicating when it should be given and how it should be phrased. Finally, it can provide the Helper with evidence about whether the Worker understood an instruction.

3.3.2 Facilitating conversation and grounding

A shared visual space may facilitate the communication that surrounds a joint activity. Successful communication relies on mutual knowledge or common ground (Clark & Marshall, 1981; Clark & Wilkes-Gibbs, 1986): the knowledge, beliefs, understanding, and so on, shared by the speaker and hearer, and known to be mutually available. Shared visual information helps communicators develop common ground by giving them evidence from which to infer what others understand at any moment.

Generally, a speaker would not speak in a non-native language unless he thought a partner understood it, would not suggest “pinging the gateway” unless he thought the partner had telecommunications knowledge, or use a pronoun unless he thought the partner understood the antecedent. Although these inferences about a partner’s state of knowledge may be incorrect, they underlie speech production. As a result, throughout a conversation, participants are mutually assessing what each other knows and then using this knowledge to form their subsequent utterances. Participants are obligated to both assess and give off cues that indicate their

understanding. This method of exchanging evidence about understanding over the course of a dialogue is referred to as the process of grounding.

Clark and Brennan (1991) hypothesize that different communication media have features that change the cost of grounding. For example, when communicating by electronic mail with large delays between conversational turns, participants cannot simultaneously transmit back channel communications—the "uh-huh", "I see", head nods, and smiles—that signal to one another the degree to which they understand the current utterance. In this research, we are interested in how shared visual information affects grounding. Clark and Brennan (1991) and Kraut, Fussell, Brennan, and Siegel (2002a) suggest ways that a shared visual space can be helpful for establishing common ground.

The principle of least collaborative effort asserts that participants in communication will try to minimize their collaborative effort (i.e., the work that they do from the initiation of each communication contribution to its mutual acceptance) (Clark & Wilkes-Gibbs, 1986). Shared visual information can help reduce collaborative effort at two distinct phases in the communication process: the *planning* stage and the *acceptance* stage.

The *planning* stage takes place when a speaker is forming an utterance; it affects the efficiency of expressions. When describing a puzzle, one of the Helpers' goals is to form expressions that succinctly denote to the puzzle's pieces. If the Helper and Worker can see the same work area, the Helper can create efficient referring expressions by relying upon what the Worker sees (e.g., using the phrase "that one" when observing that the Worker is hovering over the correct piece) or anticipating potential ambiguities (e.g., using the phrase "the dark red one" only if he can see that the Worker is likely to be confused by multiple red pieces). If the Helper cannot see the Worker's area, the Helper is likely to provide the wrong amount of information or rely upon the Worker to state explicitly what information she needs. Thus, by the principle of least collaborative effort, we should expect to see shifts in who acknowledges when a task is completed based on the degree of shared visual space.

The *acceptance* stage occurs when the speaker is assessing whether the conversational partner has understood the utterance. It provides comprehension monitoring. According to the collaborative model of conversation, after contributing an utterance to a conversation, a speaker should not move the conversation forward unless speaker and listener believe that the listener has understood

the utterance sufficiently (Clark & Marshall, 1981). After giving instructions about a puzzle, seeing the Worker's consequent behavior provides the Helper information about the Worker's comprehension of the instruction. With shared visual information, the Helper can easily recognize when the Worker performs an incorrect action or appears confused, and use this as evidence that they did not understand the task. For example, in the present experiment, if a Helper noticed that when the Worker put one piece directly above another in response to the instruction, "put the piece kitty-corner" he can assume that "kitty-corner" is not part of their shared language. The Helper can easily remedy this mistake by providing a more meaningful directive such as, "Above and to the right so that the corners are touching." Without shared visual space, the Helper needs to make assumptions about what the Worker understood or rely upon the Worker to explicitly state her level of understanding.

Visual information can provide a more accurate signal of comprehension than a listener's self-assessment of understanding. If the Helper tells the Worker to "position the piece at 2 o'clock" and he can see the Worker's response, he can tell with certainty that the Worker has understood the instruction. However, if there is no shared visual space, then the Worker must state her understanding, for example, "OK, it's above the last piece," to which the Helper might respond, "Above and to the upper right?" Even at this point, the Helper cannot be certain that they are both speaking about the same piece. In this way, visual information can provide a less ambiguous signal of comprehension than can language.

By seeing the partner perform some task, the Helper gets immediate feedback about whether the partner understood a directive. Clark and Krych (2004) demonstrated the temporal precision with which speakers use this visual evidence of understanding. For example, when a shared visual space is available, directors change their descriptions and further elaborate mid-sentence in response to their partner's behavior. They use visual information to determine the precise moment at which to disclose new information. Delays of the sort introduced by video compression or network lags are likely to undercut the value of the visual feedback.

Visual feedback, however, may be less necessary if the task is simple enough (e.g., a game of tic-tac-toe in which the pieces and positions are easily described) or if the partners have an efficient, well-practiced, and controlled vocabulary to describe events (e.g., routine communication between pilots and air traffic controllers). In these cases, a shared visual display provides little new information and its value for communicative purposes is diminished.

3.3.3 Maintaining awareness of task state

In the previous section, we described how shared visual information can be useful in coordinating language during the planning of utterances that a partner can understand, and in monitoring whether that partner does understand. Shared visual information can also be valuable for coordinating the task itself. In particular, if collaborators can see the state of the task as it develops, they know what work remains. This awareness helps them plan how to proceed toward the goal, what instructions they need to give, and how to repair incorrect actions. Shared visual information also provides the ability to monitor specific actions[2].

Imagine a pair performing a typical referential communication task in which a Helper is instructing a Worker on the order in which to place a set of cards (Isaacs & Clark, 1987). If the Worker places a card to the left when it should have been placed to the right, the Helper can intervene with new instructions if he can see the work area. Otherwise, the Helper must query the Worker on the order of the cards and rely upon the Worker to provide an accurate description.

The benefit of the shared visual information should increase as the task grows more visually complex because visual complexity introduces more possibilities of task errors, and because the language is less adequate to describe the task state. For example, in the puzzle task used in the present experiment, the puzzles are two-dimensional (with abutting pieces) or three-dimensional (where one piece may overlap and occlude another), with corresponding levels of complexity. In the simple two-dimensional case, the instruction "Put the red piece on top of the blue one" is unambiguous, whereas in the three-dimensional case, the red piece can either overlap the blue piece or be north of it. If the Helper can see the work area, he can intervene to rectify any misinterpretation. He can also see when the Worker is ready for the next instruction.

---

[2] It should be noted that the distinction between the use of shared visual information for conversational grounding and for maintaining situation or task awareness is a subtle one. Conversational grounding, or knowing what a partner believes and knows, and situation awareness, knowing the state of the task and surrounding environment, often overlap in real world environments. However, maintaining a conceptual distinction between these mechanisms is useful from a theoretical perspective. This chapter considers the impact that shared visual information has with respect to both of these theories; however, Chapter 5 examines the independent effects of each of these mechanisms.

3.4 Hypotheses

This discussion about the influence of shared visual information on conversational grounding and task awareness can be summarized in three sets of hypotheses regarding task performance in the puzzle study paradigm. The first concerns the effect of shared visual information on task performance as measured by completion time. The second and third address the way in which shared visual information changes the content and structure of the communication as the pairs attempt to reduce their collaborative effort.

Performance. Because the shared visual information should help participants maintain awareness of what needs to be done in the puzzle and allows them to communicate more efficiently, we expect that it will lead to improved performance.

> *General Hypothesis 1 (H1): A collaborative pair will perform a referential communication task more quickly when they have a shared view of the work area.*

When the referential task is more visually complex and involves a rapidly changing environment, language alone becomes less adequate for describing the task state, and the likelihood of errors increases. In these cases, the shared visual information should be more useful, and we should expect an interaction effect between the presence of shared visual information and the visual complexity of the task.

> *H1a: A shared view of the work area will have additional performance benefits when the task is more visually complex.*

We would further expect an interaction between the temporal dynamics of the task objects and the fidelity of the shared visual space.

> *H1b: A shared view of the work area will have additional performance benefits when the objects in the task change versus when they are stable.*

However, the shared visual information should be less useful if it is not kept up to date because it will not be synchronized with the state of the task or the language it needs to support. As described by Clark and Krych (2004), spoken language is particularly useful when it can be

precisely timed to physical actions and behaviors. Even a small delay in updating the visual space should be enough to disrupt this precision timing and diminish the value of visual information.

H1c: Delay in transmission will diminish the value of a shared view of the work area.

Communication efficiency. If shared visual information allows pairs to communicate with less collaborative effort, this should be reflected in the efficiency of their language use, that is, the number of words they need to give instructions, refer to objects, or indicate their state of comprehension.

General Hypothesis 2 (H2): A shared visual space will allow collaborators to communicate more efficiently.

H2a: Collaborators will use fewer words to complete their task when they have a shared visual space.

Even though the shared visual information provides new information to the Helper by allowing him to see the Worker's behavior, we expect that the visual tool will primarily influence the Worker's language efficiency. If the pairs are operating according to the principal of least collaborative effort and the Worker is aware that the Helper can see the space, then the Worker can let her actions substitute for words in demonstrating her level of understanding.

H2b: A shared visual space should increase the Worker's communicative efficiency more than the Helper's.

Communication process. To influence communication efficiency, the shared visual information must also affect the strategy collaborators use to form utterances and indicate their level of understanding. Because the Helper forms his utterances on the basis of intuitive hypotheses regarding the information the Worker needs, providing a shared visual space should allow him to rely on more efficient linguistic shortcuts, such as the use of deictic pronouns and spatial deixis, in the formulation of referential statements. Both of these linguistic forms are ways of verbally referencing (or pointing to) a particular object in the display, or in the case of spatial deixis, the spatial relation between a reference object and a to-be-located object. For example, in the phrase "I want that" (pointing to an object), "that" is a deictic pronoun used to linguistically point to the

object. Deictic pronouns are generally efficient, substituting for longer and more linguistically explicit referring expressions. Spatial deictic expressions are an example of longer and more explicit forms. For example, in the expression "It's the one on top of the red block," "on top of " uses the relative spatial position of objects to refer to them. If both Helper and Worker can see the spatial positions of puzzle pieces and know their partner can also see the positions, they should not need elaborated spatial deixis.

H3a: A shared visual space should increase collaborators' use of deictic pronouns.

H3b: A shared visual space should decrease collaborators' use of explicit descriptions of spatial position (spatial deixis).

In addition to the general efficiencies shown in the planning of messages, shared visual information allows pairs to change their strategies for demonstrating and monitoring comprehension and should also reduce the amount of effort needed to monitor comprehension. With shared visual information, the Helper can directly observe evidence of the Worker's comprehension. As a result, the Worker need not explicitly state it. On the other hand, without shared visual information, Workers must frequently indicate verbally whether they have understood utterances.

H3c: The availability of shared visual information should decrease the number of acknowledgements explicitly stated.

A lack of shared visual information should shift the burden of responsibility for verifying comprehension to the person performing the action. In the puzzle study explored here, this means the Worker will need to assume the responsibility of confirming their actions verbally.

H3d: A lack of shared visual information should additionally increase the number of acknowledgements explicitly stated by the Worker.

3.5 Method

These hypotheses are investigated in an experiment that manipulates the fidelity of the shared visual space and the attributes of the task. Participant pairs played the role of Helper and Worker in the puzzle study experiment described in Chapter 2.

3.5.1 Apparatus

The Helper and Worker were each seated in front of separate desktop computers with 21-inch color monitors. A divider positioned between the workstations prohibited the participants from seeing one another. This eliminated the pair's ability to use hand gestures, facial expressions, and so on. The Helper and Worker spoke out loud and each speech stream was captured by microphone and integrated with a time-stamped video capture of the displays. The general structure the displays matched that in Figure 2-1. Pairs were notified before each trial regarding the status of the shared work area for the upcoming trial; for example, they were told whether or not the Helper could see the workspace, and if so, whether or not it was temporally in synch.

3.5.2 Independent variables

We manipulated the extent to which participants viewed the same work area (Immediacy of the Shared Visual Information), the adequacy of lexical tokens to describe the puzzle pieces (Color Drift) and the visual complexity of the task itself (Puzzle Difficulty).

Immediacy of the Shared Visual Information. The Helper could either see a replication of the Worker's work area with no delay, could see the work area with a 3-second delay, or could not see the work area at all. We call these, respectively, the Immediate, Delayed, and None shared visual space conditions.

Color Drift. The temporal complexity of naming the puzzle pieces was varied by manipulating whether the colors of the blocks remained constant throughout the trial in the Stable condition (e.g., red), or constantly cycled in the Drift condition (e.g., red to orange to yellow to…). In the Stable condition, pieces were chosen randomly for each experimental condition from a palette of easily distinguishable colors. In the Drift condition, each piece slowly changed its color, incrementally cycling through the colors in the color palette. The pieces changed at a rate of a major perceivable color change approximately every five seconds. It took roughly one second of continuous observation to notice whether any given piece was changing color. It should be noted that these values fluctuate somewhat due to the fact that people do not perceive change equally across the color spectrum[3].

[3] In this study a single setting was used for the rate of color change of the drifting pieces, a more detailed discussion of change rates and their impact is found in Chapter 4.

Puzzle Difficulty. The visual complexity of the task was manipulated by having 2-D configurations where the pieces simply abutted edges (Easy) or 3-D configurations where the pieces could overlap one another (Difficult). In the difficult condition, a piece could overlap either one-quarter or one-half of another piece. The layout algorithm guaranteed that a single piece was never completely occluded.

3.5.3 Participants and procedures

Participants consisted of twelve pairs of participants selected from the Pittsburgh, PA area. Individuals were randomly assigned to play the role of Helper or Worker. Color Drift was manipulated between pairs of participants, while both Visual Space and Puzzle Difficulty were manipulated within each pair. Each pair participated in six experimental conditions, once in each Visual Space (3) × Puzzle Difficulty (2) combination, counter-balanced. Pairs solved four puzzles within each experimental condition. This resulted in a total of 24 puzzles that were completed in approximately one hour.

3.5.4 Measures

3.5.4.1 Task performance measure

The participants were instructed to complete the task as quickly as possible, so task performance was the time it took to complete the puzzle. Custom software logged and time-stamped all mouse events. Puzzle completion times were extracted from the logs by calculating the time between two events: (1) when both partners pressed buttons indicating that they were ready to proceed with the next trial, and (2) when the Helper pressed a button indicating that the trial was successfully completed. Because the vast majority of the puzzles were solved correctly, differences in error rates were less useful as indicators of task performance.

3.5.4.2 Conversational coding

To investigate the relationship between the availability of shared visual information and dialogue, we employed a coding scheme to identify the speaker (Helper or Worker) and the primary purpose of each utterance and action (see Table 3-1). This method was modified from the coding scheme described in Kraut et al. (Kraut et al., 2003)[4]. The typical cycle of performance involved

[4] The original coding scheme and complete instructions can be found in Appendix A.

the Helper describing one of the puzzle pieces, waiting until he was convinced that the Worker had identified the correct piece, and then telling the Worker its position in the work area. When he was convinced the piece was placed correctly, he would describe the next piece. This would be repeated until the puzzle was completed.

Table 3-1. Types of utterances coded.

Utterance Types	
Referent	References to and attempts to describe a specific piece (e.g., "Take the red one").
Referential context	Information providing the context for identifying a specific piece (e.g., "What colors do you have available?").
Position	Attempts to describe the position of a single specific piece (e.g., "Put that one in the upper right corner").
Positional context	Description of several pieces together (e.g., "The last three blocks should form a triangle like shape").
Acknowledgements of understanding	Responses to statements confirming an understanding (e.g., back-channel responses, "mmm-hmm").
Acknowledgements of behavior	Acknowledgements directly following a behavior indicating whether a partner had made a correct or incorrect move (e.g., "OK, I've done it.").
Deictic expressions	
Deictic pronouns	Utterances that use the deictic pronouns "this," "that," "there," and related terms.
Spatial deixis	Utterances that refer to terms using spatial position, such as "above," "below," "in front of," "on top of," "next to," "behind," "right," "left," "up," "down," "touching."

In this chapter, we are especially interested in the language efficiency and manner in which participants referred to the objects in the puzzle, described the spatial positions of those objects, and verified that they were manipulating the correct pieces and positioning them correctly. To examine these issues in detail, we conducted our analyses using the categories presented in Table 3-1. In particular, the *reference* and *position* categories represent substantive task communication. When spoken by the Helper, they were often instructions telling the Worker what to do. When spoken by the Worker, they were often attempts to clarify an instruction or verify that she had understood it correctly. The *acknowledgement* categories were brief exchanges asserting that the Worker had understood an instruction or performed it correctly. The *acknowledgements of understanding* represent instances of conversational grounding, whereas the *acknowledgements of behavior* primarily represent task awareness. The bottom half of Table 2-1 presents categories that were used to assess the efficiency of the spoken communication taking place by examining the use of deictic pronouns and spatial deictic descriptions.

Two independent coders classified a 12% sample of utterances until they reached 90% agreement on all categories. They then each coded different transcripts, periodically coding a common transcript to ensure that the categories they used did not drift during the duration of the coding. Agreement remained high throughout.

3.5.5 Statistical analysis

Each analysis was a repeated measures analysis of variance where Block (combination of conditions 1-6), Trial (1-4), Puzzle Difficulty (Easy or Hard) and Immediacy of Visual Information (Immediate, Delayed, None) were repeated, and Color Drift (Stable or Drift) was a between-pair factor. All 2-way and 3-way interactions were included in the analysis. Because each pair participated in 24 trials (6 conditions by 4 trials per condition), observations within a pair were not independent of one another. Pairs, nested within Color Drift, were modeled as a random effect. The analysis of performance used time to complete the puzzle, recorded in seconds, as the dependent variable. When conducting analyses of conversational efficiency, the dependent variable was the number of words, and time to complete the task was included as a covariate. The analysis for conversational content examined the number of referents, position statements, acknowledgements and deictic expressions, and included both time and number of words as covariates.

The major interest in this study was in examining how changes to the fidelity of the shared visual information affected task performance, conversational efficiency, and conversational tactics. Although the analyses were full factorial analyses of covariance with up to 3-way interactions, this chapter focuses on the Immediacy of the Shared Visual Information and its interactions with Puzzle Difficulty, Color Drift and Speaker Role.

3.6 Results

3.6.1 Manipulation checks

The manipulation of Puzzle Difficulty had a significant impact on the speed with which the pairs solved the puzzles. The pairs were faster in the easy 2-D condition when the pieces simply abutted edges (LS Mean (and standard error) = 62.5 (3.8)) than when they were in the difficult 3-D condition where the pieces overlapped (70.0s (4.3)), $t_{(258)} = 2.40$, $p = .017$. The manipulation of Color Drift also had a significant impact on performance speed. The pairs were significantly

faster in trials where the colors were stable (54.4s (5.3)) than when they were drifting (78.0s (5.3)), $t_{(258)} = 3.19$, $p = .009$.

3.6.2 Task performance

This experiment was designed to examine the impact of the availability of shared visual information on performance for different types of tasks. The results are shown graphically in Figure 2-1, and additional statistical details are contained in Appendix F[5]. Consistent with *General Hypothesis* 1, the results show that a shared view of the work area benefited performance. The pairs were about a third quicker at solving the puzzles in the Immediate Shared Visual Information trials than in either the Delayed Shared Visual Information, $t_{(258)} = 4.57$, $p < .001$, or the No Shared Visual Information trials, $t_{(258)} = 6.61$, $p < .001$ (Immediate = 52.3s (4.2); Delayed = 69.6s (4.5); None = 76.7s (4.4)). However, consistent with *Hypothesis 1c*, the 3-second delay substantially reduced the benefits of the shared visual information.

Consistent with *Hypothesis 1b*, the Immediacy of the Shared Visual Information × Color Drift interaction demonstrates that a shared view of the work area had greatest benefit in the Drift condition, when the objects being discussed were lexically unstable and difficult to describe $F_{(2, 258)} = 11.41$, $p < .001$ (see Figure 3-1). Decomposition of this interaction reveals that the Immediate Shared Visual Information condition led to substantially faster completion than the No Shared Visual Information condition when colors were changing than when they were stable, interaction $t_{(258)} = 4.33$, $p < .001$. Similarly, the Immediate Shared Visual Information condition was faster than the Delayed Shared Visual Information condition when the colors were drifting than when they were stable, interaction $t_{(258)} = 2.19$, $p = .03$ (see Figure 3-1). Phrased another way, a shared view of the work area was less beneficial when words themselves could easily describe the objects (e.g., they could be referenced by concise color terms such as red, blue, or aqua). Because people precisely time their utterances in the grounding process (Clark & Krych, 2004), temporal synchrony matters a great deal.

It is instructive that the Immediacy of the Shared Visual Information × Puzzle Difficulty interaction, although in the hypothesized direction, was not statistically significant, $F_{(2, 258)} = 1.01$,

[5] Throughout this dissertation, additional statistical details are presented in Appendix F, where the data from each chapter are included under their own sub-heading.

p = .37. Visual complexity itself did not raise the value of a shared view of the work area. Thus, we found no statistical support for *Hypothesis 1a*. It was primarily when the task was dynamic and the environment changing that the shared visual information was most beneficial.

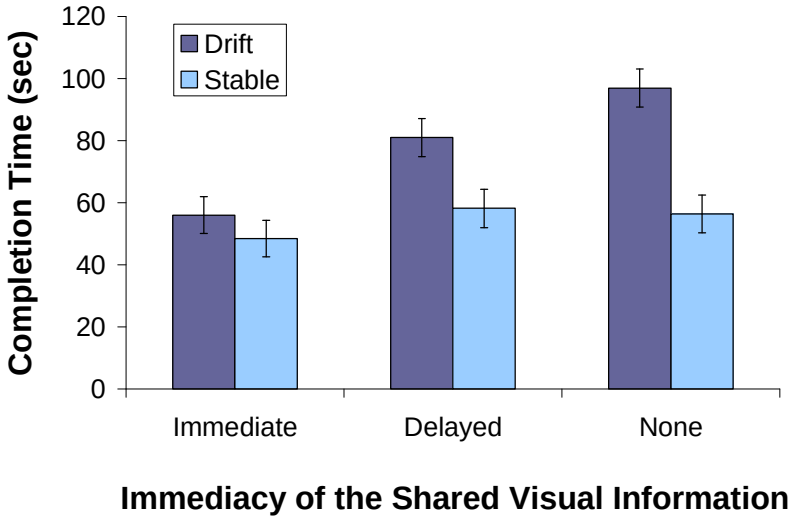


Figure 3-1. Effect of shared visual information and color drift on performance time.

The next stage of the analysis explored the way in which the language use between the Helper and Worker varied when the shared visual information was perturbed.

3.6.3 Communication efficiency

We explored the rate at which the pairs produced words (in the log scale) in order to examine the efficiency with which they communicated. We examined word rate (the number of words, controlling for time) to test this prediction. The model used for the word rate analysis was similar to that used to examine task performance, with a few exceptions. It included Speaker Role (Helper or Worker) as a factor and used time to complete the task as a covariate. Because none of the three-way interactions were significant, with the exception of Block × Availability of Shared Visual Information × Speaker Role, they were removed from the model in subsequent analyses.

Consistent with *General Hypothesis 2* and *Hypothesis 2a*, the pairs produced more efficient speech when they had more immediate shared visual information. They used fewer words to

solve the puzzles, controlling for time, as the shared visual space was more up-to-date (Immediate = 2.97 (.14) words (nLog) per puzzle; Delayed = 3.40 (.15); None = 3.81 (.15)). Using these measures, the Immediate Shared Visual Information condition was more communicatively efficient than both the Delayed Shared Visual Information condition, $t_{(110)} = 2.55$, $p = .01$, and the No Shared Visual Information condition, $t_{(110)} = 4.84$, $p < .001$. In turn, the Delayed Shared Visual Information condition was more efficient than the No Shared Visual Information condition, $t_{(110)} = 5.78$, $p = .017$.

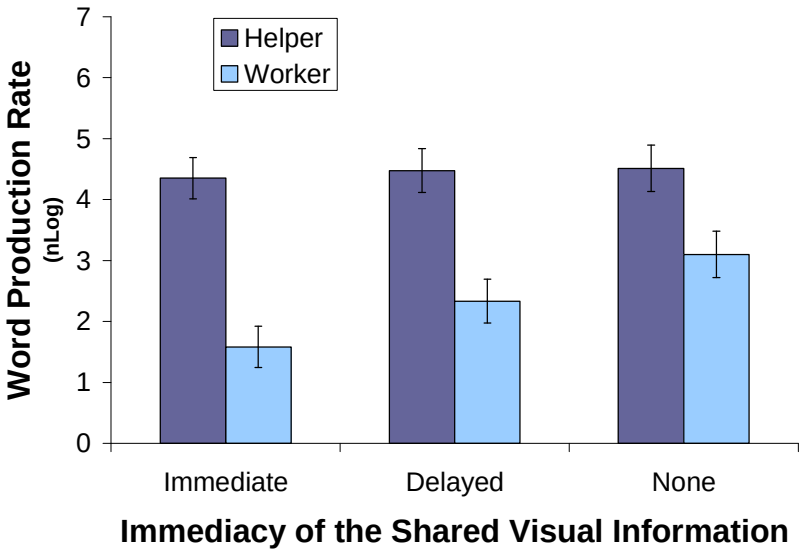


Figure 3-2. Effect of shared visual space and speaker role on word rate.

An examination of the Shared Visual Information × Speaker Role interaction depicted in Figure 3-2 reveals that the immediacy of the shared visual information improved the Workers' efficiency more than it improved the Helpers' (for the interaction $F_{(2,110)} = 10.81$, $p < .001$). Because the Workers could always see the work area, changes in Workers' behavior reflected their accommodation to differences in the Helpers' view of the workspace. This provided support for *Hypothesis 2b*.

3.6.4 Communication processes

We expected that the shared visual space would be useful in allowing the pairs to monitor the state of the task. When the workspace was present, the Helper could monitor the Worker's progress and issue corrections. However, when the shared visual information was not available,

the responsibility of communicating the task state shifted to the Worker. This role switching and the responsibility of contributions was first described by Brennan (1990). Here we extend this work by examining two particular types of verbal acknowledgements that can be produced in describing task state. Acknowledgements of behavior are verbal acknowledgements that occur in response to behaviors or physical actions. Acknowledgements of understanding are verbal acknowledgments that occur in response to verbal statements or questions. The method of coding the transcripts was a modified version of the coding scheme described in Kraut, Fussell, and Siegel (Kraut et al., 2003) and the subset of the codes analyzed here are presented in Table 3-1. The models used to perform the content count analyses were similar to the model used to examine word rate; however, they included the number of words as a covariate, allowing one to view the values described here as proportions of overall word production. These analyses permit the investigation of changes in patterns of language use.

Acknowledgements of behavior. Table 3-2 demonstrates a typical example of the ways in which the pairs acknowledge behaviors with and without access to shared visual information. Consistent with *Hypothesis 3d*, Workers took over the responsibility for assessing and communicating the state of the task when Helpers did not have up-to-date visual information. When the pair had no shared visual space, the Worker indicated explicitly whether she understood an instruction and performed it correctly by reporting on the current task state, (e.g. "OK, so it's like [on the] side of it and you see half of the red block."). The Helper then confirmed the placement was correct with the phrase, "Right of the red, yeah."

Table 3-2. Shifts in responsibility in assessing and communicating correctness of performance.

Immediate Shared Visual Information	No Shared Visual Information
H: The right hand, the top right hand corner of the blue block touches the bottom left hand corner of the first orange block.	**H:** And that's gonna be on top of the red one but only the right side of the red is going to be showing.
W: [Positioned piece correctly]	**W:** [Positioned piece correctly]
W: Like that?	**H:** You know what I mean?
H: Yeah.	**W:** OK, so it's like …
H: All right, that's good.	**H:** Oh, like, put it on the left side of the red.
	W: … side of it and you see half of the red block.
	W: OK.
	H: Right of the red, yeah.

In contrast, when shared visual information was available, the Helper could visually confirm that the Worker understood the instruction (e.g., with the statement, "… that's good") without the Worker having to explain the state of the puzzle environment.

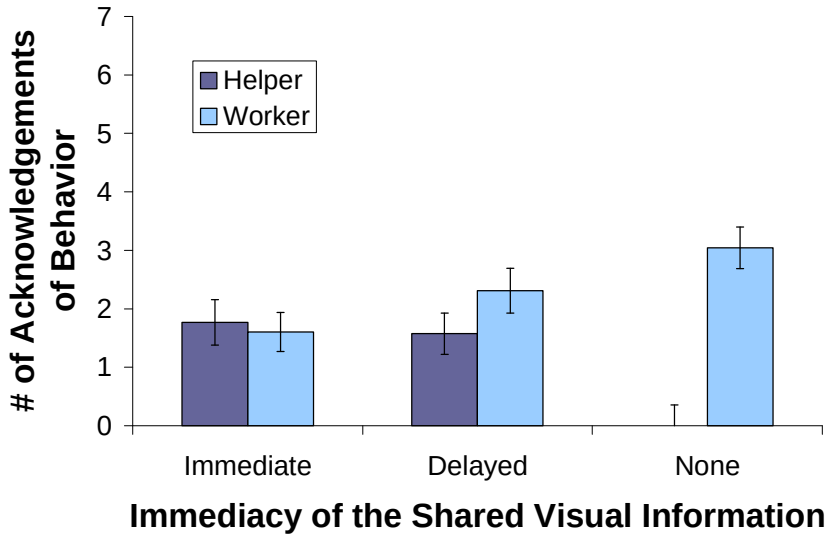


Figure 3-3. Effect of shared visual space and speaker role on the production of acknowledgements of behavior[6].

Consistent with *Hypothesis 3d*, statistical analyses supported the shift in responsibilities. In the Immediate Shared Visual Information condition, the Helper issued nearly as many behavioral acknowledgements as the Worker. That is, the Helper was as likely to tell the Worker that she had positioned a piece correctly as the reverse. However, when the shared visual information was limited, Workers increased their production of acknowledgements (see Figure 3-3), interaction $F_{(2,105)} = 33.56$, $p < .001$. This Availability of Shared Visual Information × Speaker Role interaction is stronger when comparing the Immediate and No Shared Visual Information conditions, $t_{(105)} = 8.10$, $p < .001$, than in comparing the Immediate and Delayed conditions, $t_{(105)} = 2.49$, $p = .014$. *Hypothesis 3c* was not supported for acknowledgements of behavior. Although

[6] The data presented in this figure represent the overall number of acknowledgements of behavior controlling for the total number of words. See Appendix F for detailed model results.

responsibility for acknowledging correct behavior shifted across the shared visual information conditions, the total rate did not appear to change in this study.

Acknowledgements of understanding. The pairs also used visual information to support the conversational grounding process. When shared visual information was available, it was more efficient and easier for them to follow a cycle of the Helper giving instruction and the Worker performing actions. They could reserve speech for clarification when things went wrong. There was little need for Workers to state their understanding of instructions explicitly, since Helpers could infer understanding by observing whether Workers performed correctly. However, when the visual information was not immediately available, Workers had to be more explicit in communicating their understanding.

Consistent with *Hypothesis 3c*, the pairs were most explicit in stating their understanding when they had no shared visual space, $F_{(2,105)} = 12.43$, $p < .001$. They used acknowledgements of understanding more when they had no shared visual display than when it was available, $t_{(105)} = 4.59$, $p < .001$, or when it was delayed, $t_{(105)} = 4.10$, $p < .001$. However, in this study we found little difference between the presence of an immediate display and the presence of a delayed one, $t_{(105)} = .57$, $p = .57$, (LS Means (se): Immediate = 1.30 (.27); Delayed = 1.51 (.27); None = 3.11 (.29)). It appeared as though the pairs were willing to use the delayed visual information to play this role[7].

The Shared Visual Information × Speaker Role interaction provides support for *Hypothesis 3d*, extends work by Brennan (Brennan, 1990, 2005), and demonstrates further support for the notion that pairs act in accordance with the principle of least collaborative effort. Workers were more explicit in stating their understanding when the shared visual information was not immediately available (see Figure 3-4), for the interaction $F_{(2,105)} = 8.66$, $p < .001$, while the Helper's behavior did not change much with variations in the shared visual space.

[7] The next chapter provides additional data that is more temporally explicit regarding the pairs' need to use the shared visual, even when it is delayed, to serve as a mechanism for supporting conversational grounding.

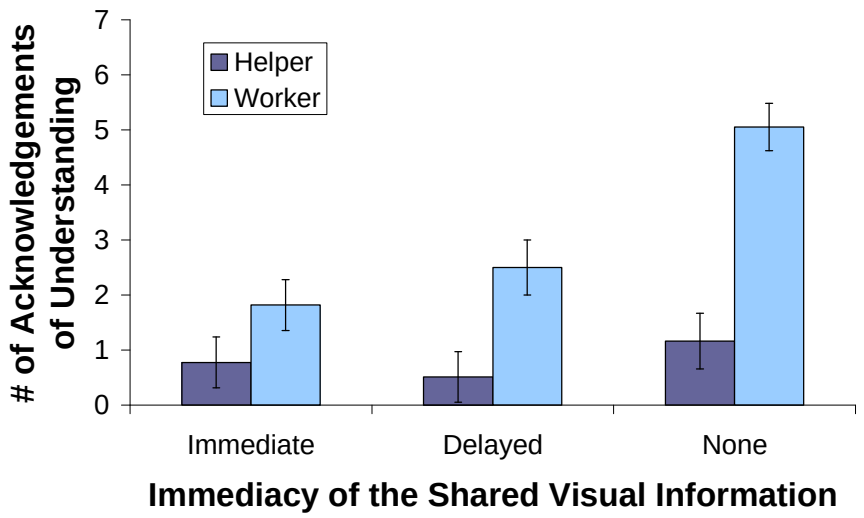


Figure 3-4. Effect of shared visual space and speaker role on the production of acknowledgements of understanding.

The Shared Visual Information × Color Drift interaction showed an additional increase in the use of acknowledgements of understanding when the colors were drifting than when they were stable, for the interaction $F_{(2,105)} = 5.30$, $p < .006$.

3.6.5 Deictic expressions

Deictic pronouns. Since the task in this study required the pairs to identify specific objects and then place them in a spatial arrangement, we expected that they would prefer to use shorthand references to objects as opposed to lengthy verbal descriptions when they could. As demonstrated in Table 3-3, the pairs appeared to take advantage of the shared visual information to support the generation and use of efficient deictic references. The models used to perform the deictic count analyses were similar to the model used to examine word rate; however, they included the number of words as a covariate, allowing one to view the values described here as proportions of overall word production. Statistical analysis of the conversational data was consistent with *Hypothesis 3a*. The pairs differed in their use of deictic pronouns by condition, $F_{(2,105)} = 5.47$, $p = .006$. They used more in the Immediate condition than in the No Shared Visual Information condition, $t_{(105)} = 3.31$, $p = .001$. However, while the difference between the Immediate and

Delayed conditions was in the expected direction, it was not significant, $t_{(105)} = 1.71$, $p = .09$ (Immediate = 1.50 (.20); Delayed = 1.01 (.21); None = 0.512 (.22)).

Table 3-3. Use of deictic pronouns with and without access to shared visual information.

Immediate Shared Visual Information	No Shared Visual Information
H: And that over… put that on top of the red one.	H: The bright blue's, the bright blue's, um, bottom left corner touches the bright red's upper right corner.

Spatial deixis. Spatial deixis is the term used in this study for attempts to refer to an object by describing its position in relation to others, in phrases such as "next to," "below," or "in front of." Spatial descriptions are communicatively expensive. They are less efficient than a simple noun phrase (e.g., "the blue one") or a deictic pronoun (e.g., "that one"). If pairs are trying to minimize collaborative effort, they should use spatial deixis less frequently than when they have access to more efficient shared visual information. Analyses showed a trend of the use of differing proportions of spatial deixis depending on the availability of shared visual information. Although the overall *F*-test did not reach statistical significance, $F_{(2,105)} = 2.67$, $p = .074$, pair-wise comparisons revealed that the pairs tended to use spatial deixis more in the Delayed than in the Immediate Shared Visual Information condition, $t_{(105)} = 2.26$, $p = .02$. However, the difference between the No Shared Visual Information and the Immediate Shared Visual Information did not reach significance, $t_{(105)} = 1.58$, $p = .11$ (Immediate = 2.82 (.29); Delayed = 3.64 (.30); None = 3.41 (.31)).

The shared visual information had less of an impact on spatial deixis when the colors were stable (for the interaction $F_{(2,105)} = 3.21$, $p = .04$), and when the puzzle configurations were easy (for the interaction $F_{(2,105)} = 3.65$, $p = .03$). Thus, if the task was linguistically or spatially difficult, the absence of shared visual information caused participants to resort to costly spatial descriptions to resolve it.

There was also a trend of the shared visual information affecting the Helpers' use of spatial deixis more than that of the Workers'. Although the overall *F*-test did not reach statistical significance (for the interaction $F_{(2,105)} = 2.15$, $p = .12$), pair-wise comparisons indicated that Helpers used spatial deixis more when the fidelity of the display was decreased, whereas Workers tended to produce a consistent number of spatial deixis per puzzle regardless of the view. This interaction was significant for the comparisons between the Immediate and Delayed conditions, $t_{(105)} = -2.01$,

$p < .05$; however, it failed to reach significance for the comparison between the Immediate and No Shared Visual Information conditions, $t_{(105)} = -1.58$, $p = .12$.

3.7 Discussion

In this chapter I have demonstrated the feasibility of the puzzle paradigm for investigating the conditions under which shared visual information improves collaboration, and showed that shared visual information interacts with task features in substantial ways. This work also demonstrates that shared visual information plays a major role in supporting conversational grounding and task awareness.

3.7.1 Facilitating conversational grounding

This research showed that collaborative pairs can perform more quickly and just as accurately when they have a shared view of a common work area. The shared visual information improved task performance and conversational efficiency. Delay in updating the visual information diminished the benefits of having a shared visual workspace in most dimensions.

There are two major ways that the shared view of the work area improved performance by allowing Helpers to accurately ground their instructions. First, the shared work view allowed Helpers to plan and create more efficient referring expressions to describe objects and positions in the work area. Seeing the Workers' behavior allowed Helpers to use deictic pronouns and other compact expressions instead of longer noun phrases to refer to elements in the puzzle. In addition, Helpers could see directly when their partners were ready for the next instruction, reducing the time between their instructions. Similarly, Workers, knowing that their partners could see their moves, could ask for confirmation with compact expressions such as "Like that?", rather than verbally describing the new state of the puzzle.

The second way in which the shared visual information improved task performance was through making conversational grounding more accurate and efficient. The shared visual information provided an important resource that allowed participants to comprehend the degree to which their partners understood an utterance. In particular, when Helpers could see Workers' behavior, they used this information to infer whether the Worker understood the current instruction. Observations of the interactions suggest that when Helpers saw that their partner made a correct move following an instruction, they cut short their descriptions and did not elaborate, but instead

continued to the next instruction. In contrast, if they observed that their partner made an error, they would provide more detail when describing a puzzle piece or its position.

This reasoning is consistent with the finding that Helpers used explicit descriptions of spatial positions (i.e., spatial deixis) less frequently in the trials where they received Immediate Shared Visual Information than those in which they did not. When the Helper could see the Workers' behavior, the Worker's placement of a piece in the correct place was immediate, costless evidence that they understood an instruction. Therefore, they could curtail their more elaborate spatial descriptions. However, without this evidence, the Helpers continued to elaborate the spatial description until the Workers explicitly confirmed their understanding.

The data presented here are broadly consistent with a cooperative model of communication. They provide broad support for Clark's thesis that common ground is crucially important for conversation, and specific support for Clark and Brennan's (1991) hypothesis that different communication features change the cost of achieving common ground. In particular, Workers adapted their communication and behavior to compensate for what the Helper could or could not see. It is important to note that in this experimental design the Worker's view of the workspace was always the same regardless of whether the Helper could see it. If the Workers were using a purely egocentric approach to communication, they would not change their communication behavior in response to variations in the shared visual information, because their view of the space never changed. Instead, they changed their communicative behavior in response to what their partner could see. When the Helper could not see the work area, Workers used more words to complete the task, were more likely to describe the work area after they made moves, and were more likely to indicate explicitly whether they understood an instruction.

The results are also consistent with Clark and Brennan's (1991) framework for analyzing the costs and benefits of different communication technologies. When media provide visual information about what the Worker is doing, the ability of Workers to ground their utterances via actions reduces their need to provide verbal indicators of comprehension. Instead, they let their actions demonstrate their understanding of the Helpers' instructions. In Chapter 6, sequential analysis techniques are used to examine this issue in more detail (Gergle et al., 2004a). In particular, sequential analyses show that the Helper's instructions were more likely to be followed by the Worker's movement of a puzzle piece when the shared visual information was available versus when it was not. In contrast, a Helper's instructions were more likely to be followed by a

Worker's acknowledgement of understanding when there was no shared visual information available.

These results show that people try to compensate for limitations in the communication technologies available to them. However, these compensations often fall short with regard to communication efficiency. For example, as previously discussed, when Workers believe that their partners cannot see their behavior, they are more explicit in indicating their level of comprehension. Yet, acknowledgements of understanding can be inaccurate. As any teacher knows, students can think they understand an instruction without really doing so. When Helpers could view the Workers' behavior, they received more accurate information about Workers' level of understanding, untainted by the Workers' self-assessments.

3.7.2 Maintaining task awareness

This work extends the work of Clark and Brennan (1991) by illustrating how *features of the task* interact with features of the communication setting to influence the grounding process. In this experiment, the value of the shared visual information depended on the task being performed. The shared visual information helped performance and conversational efficiency more when the tasks were dynamic (i.e., in the Color Drift condition).

The interactions between the fidelity of shared visual information and the features of the task demonstrate the importance of understanding task characteristics when determining the value of a shared visual workspace. These findings suggest that the utility of a shared visual workspace depends in part on the visual complexity of the task. In dynamic settings or ones with many objects in a variety of spatial relationships to one another (e.g., for distributed medical teams, aircraft repair), visual space may be particularly important. For less complex visual tasks, especially those in which objects and spatial relationships are static and easily lexicalized, an audio-only connection may suffice. These findings help to rectify the disparity between early and more recent research on the value of visual information in distributed communication.

In this study, task objects changed rapidly in the drift condition, and when they did, temporal delays in visual information rate erased the benefits that the shared visual information otherwise provided. I would expect these results to generalize to other settings with rapidly changing events, such as an operating room. Temporal delays may be less problematic when task objects are relatively static, as they might be in an architectural design task.

3.8 Conclusion

In this chapter, I have argued that shared visual information is essential for complex task-oriented collaborations because it facilitates the ability of the pairs to maintain awareness of the task state, helps them to reduce errors and ambiguities when the environment is visually complex, and facilitates grounding and communication by allowing the use of efficient language and a method for monitoring comprehension. The effects of new communication technology are not superficial, and their developers should not be guided by surface characteristics. By considering the ways that technologies, and the tasks we attempt with their aid, interact with, modify, and rely on language, greater strides can be made in understanding and design. Moreover, these developments illuminate basic principles of conversation and social psychology in profound ways, bringing into focus not only technological but traditional communication processes.

However, one major limitation to the work presented in this chapter is the discrete manipulation to the temporal delay produced in this study. The 3-second delay was unrealistically high for many users of today's technologies. Further research that manipulates delay as a continuous variable is needed to gain additional insight into the specific point at which a temporal breakdown occurs. The subsequent chapters examine the critical temporal nature of the delay (Gergle et al., 2006), and provide a more thorough investigation of the impact of shared visual information on conversational grounding and situation or task awareness (Gergle et al., Under Review).

Chapter 4

The Impact of Delayed Visual Feedback[8]

The previous chapter presented a study that demonstrated that when pairs work together on a physical task, seeing a common workspace benefits their performance and transforms their use of language. The results demonstrated that visual information helps collaborative pairs understand the current state of their task, ground their conversations, and communicate efficiently. It also demonstrated the fact that collaborative technologies often impinge on the visual information needed to support successful collaboration. One example of this was the introduction of delayed visual feedback in the collaborative environment. While the work in the previous chapter explored a constant delay of three seconds, temporal delays in collaborative systems typically occur at much shorter time intervals (Gutwin, 2001a, 2001b). This chapter presents results from two studies that detail the form of the function that describes the relationship between visual delay and collaborative task performance. The first study precisely demonstrates how a range of visual delays differentially impact performance and illustrates the collaborative strategies employed. The second study describes how task parameters, such as the dynamics of the visual environment, affect the amount of delay that can be tolerated.

[8] The work presented in this chapter was originally published in: Gergle, D., Kraut R. E. & Fussell, S. R. (2006). The Impact of Delayed Visual Feedback on Collaborative Performance. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2006)*, pp. 1303-1312. NY: ACM Press.

4.1 Introduction

As previously described, shared visual information is a key element of successful collaboration. However, when mediating an environment, we need to understand how technologies impinge upon the processes that support successful collaborative activity. For example, how does the often unavoidable visual delay that results from video compression or network congestion impact an individual's ability to maintain awareness of their partner's actions? How do visual delays disrupt the critical language processes required for successful communication? And how do these delays impact the task strategies pairs use to successfully collaborate?

This chapter describes a basic function that governs the influence of delayed visual feedback on collaborative task performance. It provides detailed insight into the amount of visual delay that can be tolerated before influencing collaborative performance, and how this range of tolerance depends upon features of the task. Finally, it presents both quantitative and qualitative descriptions of the strategic adaptations that occur across a range of visual delays.

4.1.1 The impact of delay on collaborative task performance

A number of studies have examined the impact *audio delay* has on communication and collaborative task performance. As demonstrated by Krauss & Bricker (1967), small audio delays of 300ms can have detrimental effects on communication processes, and delays as large as 900ms can severely impact a pair's ability to communicate. O'Connaill and Whittaker (1997) found that audio delays between 410ms and 720ms led to reduced use of back-channels, less interactive speech, and fewer instances of overlapping speech. Cohen (1982) described how a simultaneous 705ms delay of audio and video resulted in longer conversational turns and decreased overlap between utterances, and Tang & Isaacs (1993) found that a one-way delay of 570ms severely disrupted turn-taking behaviors. In summary, the work on audio delay and its impact on collaborative performance tends to find that delays below 300ms pose few problems. Delays between 450ms and 700ms can severely impact communication and coordination processes, and delays greater than 700ms drastically impact communication, coordination, and overall task performance.

While these studies examined audio on its own or combined with video, they do not provide insight into the communicative impact of the visual information itself. Do delays in updating *visual information*, of the sort introduced by video compression or network lags, undercut its value in a similar fashion?

Gutwin and colleagues have examined visual delay[9] in a number of tightly-coupled motor coordination tasks performed in a shared workspace (Gutwin, 2001a; Gutwin *et al.*, 2004). In one task (see experiment 2 in Gutwin, 2001a; and study 2 in Gutwin et al., 2004), the participants used their mouse pointers to grab objects from a central repository of on-screen objects, and then drop the objects in their own "bins" located at the edges of the screen. This task required the pairs to coordinate access to the shared objects, and they were allowed to speak to one another while performing the task. The results reported that delays greater than 200ms led to a larger number of coordination errors (as measured by pairs grabbing the same object at the same time) in comparison to conditions with no delay. These errors appeared to result from the fact that the participants had a particularly difficult time understanding the shared state of the task environment. While these studies found evidence of performance differences as measured by errors and error rates, there was little evidence of an influence of visual delay on overall task completion time. In addition, this work primarily focuses on discrete measures of delay, and in doing so, does not address the particular point at which coordination breakdowns occur. Similarly, these results do not provide a description of the functional relationship between delay levels and task performance.

The study presented in Chapter 3 (Gergle et al., 2004a; Kraut et al., 2002b) examined the impact of delay on a collaborative task that had both language and motor components. This work demonstrated that delayed visual feedback impacted performance, in part, due to its lack of synchronization with the state of the task or the language that it needed to support. It found that the delay harmed task performance as well as the communication processes and language patterns of the collaborative pairs, and that the delay had larger effects as the task became more dynamic. However, this work compared a rather long delay of 3000ms with no delay. Recent benchmark studies have suggested that average Internet latencies, across a range of geographic distances, are typically in the range of 100 to 700ms (Gutwin, 2001a, 2001b). This suggests that a more finely-grained temporal investigation of the influence of visual delay on collaboration is needed.

[9] While Gutwin and colleagues distinguish between two forms of delay, latency and jitter. Throughout this paper we use delay to refer to what they call latency.

Finally, Vaghi and colleagues (Vaghi *et al.*, 1999) performed a naturalistic study of a virtual soccer game that took place in a collaborative virtual environment. The game arena consisted of two halves to the field with a net on each side. Participants controlled avatars and attempted to bounce the ball off their avatar and into their competitor's net. The software allowed one player to play locally and the other was connected using a simulated network connection where the visual delay could be precisely controlled. Vaghi reports qualitative evidence that strategic adaptations occur when delays range from 150ms to 1000ms. They report that for delays of up to 150ms, very little changed in performance and strategy. From 150 to 300ms, the delayed player's perception of where his avatar was in relation to the ball became slightly disrupted, and as a result the ball often appeared to follow awkward trajectories. This delay appeared to influence the situation or task awareness of the player. Around 500ms, the game play shifted drastically. Vaghi reports that players demonstrated a strategy shift and began to play in a reactive and defensive fashion. For example, the participants used a move-and-wait strategy instead of using continuous interactions. From their results, the authors argue that delays of 500ms cause major disruptions to collaborative performance.

While these studies provide evidence to suggest that a phenomenon is present, they leave open the range of delays that are tolerable. We do not yet have a firm grasp on the range of visual delays that people tolerate before it influences the quality and processes of collaboration. Nor do we understand how collaborative pairs adapt to delays of different durations.

To address these questions, I developed the following methodology that allows the discovery of the form of the function that relates visual delay to performance. This work provides a much more detailed account of the impact of visual delay on collaborative task performance and communication processes. However, before proceeding with the studies, a brief theoretical discussion will describe the ways in which pairs use shared visual information and how periods of delayed accessibility can inhibit performance.

4.2 Theoretical background

As previously described, our theoretical understanding of the way that pairs use visual evidence for collaborative purposes relies heavily on two psychological theories: Situation Awareness Theory and Grounding Theory (for further discussions see Kraut et al., 2003; Whittaker, 2003; Whittaker & O'Conaill, 1997).

To briefly summarize, Situation Awareness Theory holds that visual information helps pairs assess the current state of the task and plan future actions (Endsley, 1995; Endsley & Garland, 2000). It primarily focuses on how visual information influences the ability of groups to formulate a common representation of the state of the task, which in turn allows them to plan and act accordingly. At another level, having to do more with the language and communication surrounding a collaborative activity, Grounding Theory suggests that visual information can serve as an unambiguous source of evidence of a partner's understanding which allows conversational partners to generate efficient speech and to assess a level of understanding (Clark, 1996; Clark & Marshall, 1981; Clark & Wilkes-Gibbs, 1986). Visual information provides a means of coordinating language and generating efficient and understandable discourse surrounding a collaborative activity.

Together these two theories predict that when groups have access to visual information, they will better coordinate their work: on the one hand, they can monitor the state of the task and plan and act appropriately, on the other they can deliver instructions and clarifications in an efficient and timely manner. Delay in providing the necessary visual information should influence these coordination mechanisms and ultimately have a negative impact on both task performance and communication processes. The following sections review each of these theories in detail and highlight the potential impact that *delayed visual information* may have for each.

4.2.1 Impact of delayed visual information on situation awareness

According to Situation Awareness Theory, visual information is primarily valuable for coordinating the task itself. In order for collaboration to be successful, group members need to maintain an ongoing awareness of one another's activities, the status of relevant task objects, and the overall state of the collaborative task (Endsley, 1995; Endsley & Garland, 2000). This awareness allows accurate planning of future activities, and can serve as a mechanism to coordinate tightly-coupled interactions.

As previously mentioned in §2.1.2, Nardi and colleagues (1993) describe how a scrub nurse on a surgical team can make use of visual information to help assess the current task state and use this information to anticipate the instruments a surgeon will need. If the surgeon nicks an artery and the scrub nurse can see this on an overhead monitor, she can immediately prepare cauterization materials in response to her visual recognition of a change in the current task state. Note that her plan for action occurs regardless of her need to verbally communicate with the surgeon. However,

if the visual information were delayed for some reason, such tight-coordination would not be possible and precious seconds could be lost.

In a similar fashion, but at an even finer temporal level, Gutwin and colleagues (Gutwin et al., 2004) describe how task coordination is supported by the availability of visual information during a tightly-coupled collaborative task in which pairs need to quickly move computational objects within a shared 2D workspace. When the view of the shared workspace is delayed, the pairs have difficulty assessing the state of their partner and the state of the task, and there is an increase in the number of errors they make by grasping the same piece.

To summarize, situation awareness of what is currently happening can influence the next move or action. When pairs are performing tightly-coupled interactions in a distributed environment, a delay in the availability of the visual information may disrupt the formation and maintenance of such awareness, ultimately yielding coordination difficulties.

4.2.2 Impact of delayed visual information on grounding

Grounding Theory suggests that visual information can improve collaborative task performance by supporting the verbal communication surrounding a collaborative activity. It states that successful communication relies on a foundation of mutual knowledge or common ground (Clark, 1996; Clark & Marshall, 1981; Clark & Wilkes-Gibbs, 1986). Speakers form utterances based on their expectation of what a listener is likely to know, and then monitor whether the utterances are understood. In return, listeners have a responsibility to demonstrate their level of understanding. Shared visual information serves to support both the initial *generation* of utterances as well as to provide *evidence of comprehension* (Brennan, 1990, 2005; Kraut et al., 2003).

Throughout a conversation, participants continually assess what one another knows and use this knowledge to generate subsequent contributions (Brennan, 1990; Clark & Marshall, 1981; Clark & Wilkes-Gibbs, 1986). Clark & Marshall (1981) propose that physical co-presence (i.e., visual access to a shared environment and the actions of a partner) allows speakers to anticipate what a partner knows. Hence, a person can point to an object in a shared physical environment and refer to it using the deictic pronoun “that” if she believes her partner can also see the object and her gesture. However, in distributed environments with delayed visual feedback such communicative efficiencies may no longer be available (Gergle et al., 2004b; Kraut et al., 2002b).

Visual information also influences comprehension monitoring in a number of ways. In a typical spoken interaction, partners can use explicit verbal statements (e.g., “I got it” or “do you mean the red one?”) or back-channel responses (e.g., “uh-huh”) to indicate their level of comprehension. As shown in Chapter 3, when visual information is available the visual feedback itself can be a critical resource for comprehension monitoring (Brennan, 2005; Kraut et al., 2003). Evidence can be deliberate (e.g., as in a pointing gesture) or as a side effect of proper performance of a desired action (e.g., by moving the correct object in a workspace), provided both parties are aware of what one another can see (Gergle et al., 2004a).

Recently, Clark and Krych (2004) demonstrated that collaborative pairs use visual information to facilitate the precision timing required when discussants are introducing new entities to a discourse or changing their speech mid-sentence in response to their partner’s actions. However, delays of the sort introduced by video compression are likely to undermine the value of this visual feedback. The study presented in Chapter 3 demonstrated that large delays reduce the communication benefits of shared visual information (Gergle et al., 2004b; Kraut et al., 2002b).

Although immediately available shared visual information generally improves collaborative task performance by supporting situation awareness and conversational grounding, the benefits it provides in any given situation are likely to depend on both the accuracy of the visual information (e.g., whether it is up-to-date or stale) along with the requirements for coordination imposed by the task structure. Any delays in the availability of the visual information are likely to impact these coordination mechanisms in different ways.

4.2.3 Hypotheses

This discussion regarding the influence of delayed visual feedback on collaborative task performance can be summarized in terms of a set of hypotheses that describe expected findings. Study 2 aims to uncover the functional form of the impact delayed visual feedback has on collaboration due to its impact on both lower level coordination tasks as well as higher level communication processes. In particular, from prior studies and our theoretical understanding, we should expect that:

> *H1: A collaborative pair will perform their task more quickly when they share an immediately available shared view of the work space.*

H2: A collaborative pair will perform their task more quickly as the linguistic complexity of the task objects decreases.

H3: An immediately available shared view of the work space will have additional performance benefits as the linguistic complexity of the task objects increase.

In addition to these general hypotheses of the overall impact of visual information, we should expect that the delays will *differentially* impact the coordination mechanisms on a different timescale:

H4: The various benefits provided by the shared visual information will decline as the delay increases.

As the collaborative task becomes more dynamic or tightly-coupled, we should expect the level of tolerance a collaborative pair has for delays to decrease. As the task becomes more tightly-coupled and dynamic, the pairs will experience performance deficits with shorter delays in comparison to less dynamic environments. This proposal is examined in Study 3, when the dynamics of the task environment are manipulated. As a result of this, we should expect that:

H5: A collaborative pair will perform their task more quickly when the objects in the environment are less dynamic.

H6: A dynamically changing environment will reduce the tolerance a collaborative pair will have for delay in the visual feedback.

4.3 Study 2: The impact of visual delay on collaborative performance

This study is primarily interested in assessing the pair's performance over a wide range of visual delays. In addition, it examines the conversational and communication processes adopted by the pairs. The collaborative puzzle task paradigm described in Chapter 2 is used to test the aforementioned hypotheses.

4.3.1 Method

The amount of visual delay present in the Helper's view of the workspace (Visual Delay) along with the amount of conversational grounding that was required to describe the pieces in the environment (Linguistic Complexity) were factors in this experiment.

4.3.1.1 Independent variables

Visual Delay [60-3300ms]: Visual delay times were chosen from a distribution that provided a finer level of granularity at the shorter delays since prior literature suggested that task performance might be more sensitive to times in that range. The times were generated according to the following recursive distribution:

$$f(n) = \begin{cases} T_1 = 60 \\ T_n = T_{n-1} \cdot e^{.05} \end{cases}$$

These times were then grouped into three bins for the sake of balancing participant assignment across three different ranges of delay. The *Low* delay was in the range of [60-230ms], *Medium* delay was in the range of [230-850ms], and *High* delay was in the range of [850-3300ms]. Participants were selected to receive two levels from each bin and these times were crossed with the levels of linguistic complexity[10].

Linguistic Complexity (Primary vs. Plaid): The linguistic complexity of the task was manipulated by providing the pairs with two types of pieces. The pieces were either lexically simple, easily described primary colors (e.g., red, yellow, orange, etc.), or they were more complex tartan plaids that required the negotiation of a naming convention. While the primary colors were likely to be part of a shared lexicon, and therefore required little grounding to name the objects, the plaid pieces were not, and required the pairs to negotiate the terms used to represent the various pieces. Figure 4-1 presents examples of the task objects.

[10] It is important to keep in mind that while we discuss bins here, the variable represents an essentially *continuous* range. The bins were only temporarily used in order to assign each pair to a number of delays that fell somewhere in the low, middle and high ends of the distribution. This was done to balance the pairs across the times and not to conflate any given pairs with the range of times they received. It should also be noted that the number of samples in each of the ranges were equivalent across the three ranges.

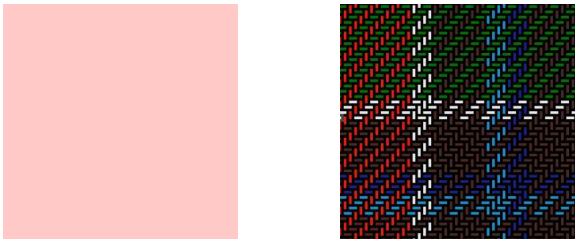


Figure 4-1. Primary pieces (left) and Plaid pieces (right).

4.3.1.2 Participants and procedure

Participants consisted of 27 pairs recruited from the Pittsburgh, Pennsylvania area. They were randomly assigned to the role of Helper or Worker, and the pairs were balanced by gender. *Visual Delay* [60-3300ms] and *Linguistic Complexity* (Primary, Plaid) were manipulated within each pair. Each pair participated in a total of six experimental conditions, three different Visual Delay times for each level of Linguistic Complexity, counter-balanced. Pairs solved four puzzles within each experimental condition. This resulted in a total of 24 puzzles that were completed in approximately one hour.

4.3.1.3 Measures

Task performance measures. The pairs were instructed to complete the puzzles as quickly and accurately as possible. Time to complete the puzzle was the primary measure of task performance. Nearly all puzzles were solved correctly, so error rates were a less useful indicator of performance.

Conversational excerpts. To detail the processes the pairs used at varying rates of visual delay, the interactions were transcribed and representative examples are presented to demonstrate qualitative evidence of the communication patterns witnessed.

4.3.1.4 Statistical methods and analyses

A statistical technique known as Multivariate Adaptive Regression Splines (MARS) (Friedman, 1991) was used to model the influence of visual delay on collaborative performance. This technique describes the effect of the independent variables (e.g., Visual Delay or Lexical

Complexity) on the dependent variable (task completion time) as an optimized sequence of piecewise linear regressions. The algorithm finds optimal breakpoints[11] by examining points within the range of the independent variable where slope changes are most likely to occur. The data are then modeled by constructing a series of dummy variables that allow for slope changes at the thresholds set by the breakpoints.

The equation (*eq.*4.1) and the illustration in Figure 4-2 present a simplified example of the piecewise linear regression approach (adapted from Gujarati, 1995). This example uses a single independent variable (*X*) regressed on a single dependent variable (*Y*). It assumes knowledge of the value for the breakpoint (*X**) (which will be learned using the MARS method) and uses the dummy variable (*D*) technique to allow for two different slopes on alternate sides of the breakpoint.

Formally, assume the following function:

$$(eq.4.1) \qquad Y_i = \beta_0 + \beta_1 X_i + \beta_2 (X_i - X^*) D_i + u_i$$

where

$$Y_i = Dependent\ Variable$$
$$X_i = Independent\ Variable$$
$$X^* = Threshold$$
$$D = \begin{cases} 1\ if\ \ X_i > X^* \\ 0\ if\ \ X_i \le X^* \end{cases}$$

Note that when $D = 0$, the third term falls out of the equation, leaving only the slope coefficient, β_1. However, when $D = 1$, the third term remains in the equation and represents the additional influence that occurs at levels of X greater than X^*. This is presented graphically in Figure 4-2.

[11] I refer here to breakpoints which are also known in the literature as "knots," particularly in the more general class of models known as spline functions (i.e., piecewise polynomials of order *k*).

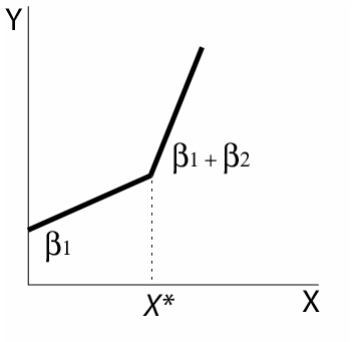


Figure 4-2. Demonstration of the line segments and their slope coefficients using a piecewise linear regression with a learned breakpoint at point *X*.*

The resulting coefficients for β_1 and β_2 can be examined to ascertain the slopes of the segments. The β_1 coefficient provides the slope of the first segment, and if the β_2 coefficient is significant, $\beta_1 + \beta_2$ determines the slope of the second segment. Applying this technique to the puzzle study data, these slopes can explain, for example, how much delay can be tolerated before group performance begins to suffer or describe the rates at which collaborative performance is impacted over a particular range of delays.

The use of this technique to uncover where the visual delays lead to performance breakdowns requires a method for learning the breakpoints (*X**). This study uses the MARS method, a two-stage process for learning optimal cut points. The first stage begins with a forward selection process that adds functions (i.e., variables capturing the breakpoints) to the model. As more functions are added, the model begins to account for non-linear trends in the data. This cycle continues until a pre-defined number of functions have been added. At this point, the algorithm enters a second stage where it prunes the functions until it achieves an optimized tradeoff between the number of functions and the goodness of fit. The Generalized Cross Validation (GCV) measure is used as the model measure of goodness of fit (originally described in Craven & Wahba, 1979; modified by Friedman, 1991). The GCV measure strikes a balance between model complexity and quality of fit in a fashion similar to that of the Akaike Information Criterion (AIC) commonly used in parametric regression models (Akaike, 1978).

In the models presented in this paper, the algorithm was permitted to construct up to 100 functions for inclusion. Each model was evaluated using a 10-fold cross validation technique. That is, each model was created over 10 trials, with each trial using 90% of the data to train, and

the remaining 10% to test the model's performance. Performance is optimized based on the best fit as assessed by the GCV error measure.

4.3.2 Results

The first stage of the analysis used the MARS method to discover the optimal partitioning of the continuous Visual Delay factor. Two major breakpoints were found at delays of 939ms and 1798ms (the circles seen in Figure 4-3). These results were then used to construct an appropriate random effects piecewise linear regression model where [*Visual Delay*], [*Visual Delay – 939ms*], [*Visual Delay – 1798ms*], Linguistic Complexity (*Primary, Plaid*), Block (*1-6*), and Trial (*1-4*) were repeated factors. All 2- and 3-way interactions were included in the analysis for each Visual Delay segment. Because each pair participated in 24 trials, observations within a pair were not independent of one another and were modeled as a random effect. The final model achieved a good fit to the data (*Adj-*R^2 = .532; *GCV-*R^2 = .497; $p < .001$).

4.3.2.1 Task performance

Linguistic Complexity. Consistent with *H2*, the manipulation of linguistic complexity had a large impact on the speed with which the pairs could solve the puzzles. Overall, the pairs were consistently faster in the trials in which the puzzle pieces were Solids than when they were Plaids (38.0sec vs. 61.7sec; $F_{(1,610)} = 270.6$, $p < .001$).

Visual Delay. Consistent with *H1*, the more quickly the visual feedback was provided, the faster the pairs were able to complete the puzzles. However, this result was not consistent across the entire range of delays. Similarly, the results addressing *H3* were only found for delays greater than 1798ms.

For delays between 60ms and 939ms, we found no evidence to indicate any impact of delayed visual feedback on task performance ($\beta = 0.48$, *SE* = (2.87), $F_{(1,610)} = .028$, $p = .87$). As can be seen in Figure 4-3, the slope for this segment is relatively flat. In this range of delay there was no impact for either the *Primary* or *Plaid* pieces. In other words, there was no evidence of a *[Visual Delay]* × *Linguistic Complexity* interaction ($F_{(1,610)} = .71$, $p = .40$).

However, for delay rates between 939ms and 1798ms there is a significant impact on task performance ($F_{(1,610)} = 13.57$, $p < .001$). This can be seen in Figure 4-3, where the slope for this segment is rather steep. In this range, every 100ms increase in visual delay slowed the pair's

completion time by an additional 2.3 seconds (holding constant at the mean all other variables in the model). The impact of delay was equally important for both the *Primary* and *Plaid* pieces, as evidenced by the fact that there was no statistical evidence of a [*Visual Delay – 939ms*] × *Linguistic Complexity* interaction ($F_{(1,610)} = 1.74$, $p = .19$).

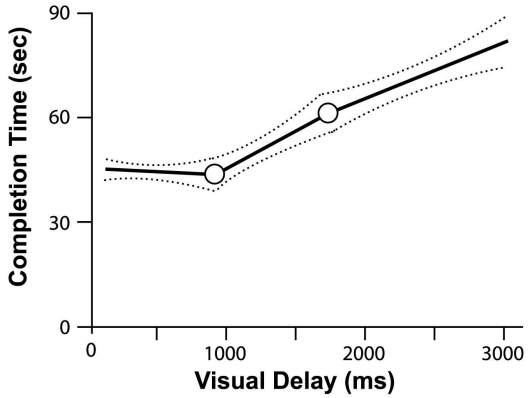


Figure 4-3. Effect of Visual Delay on Task Completion Time. Main effect graph of piecewise linear regression fit line (solid) with learned breakpoints (circles) and corresponding 95% confidence intervals (dashed).

Delay rates greater than 1798ms also demonstrated a significant impact on task performance ($F_{(1,610)} = 15.28$, $p < .001$). While Figure 4-3 illustrates the mean increase across the two levels of linguistic complexity, there was a significant [*Visual Delay – 1798ms*] × *Linguistic Complexity* interaction ($F_{(1,610)} = 10.46$, $p = .001$). In support of *H4*, decomposition of this interaction reveals that the slope for the Plaid pieces remains strong and positive, while for the Primary pieces it is flat to slightly negative. In the higher range of delays, the impact of the delay appeared to additionally affect the Plaid puzzles. This suggests that when the delays were greater than 1798ms, it appeared to impact the conversational grounding processes required to talk about the plaid pieces, while having little additional impact on the primary colored pieces which were already a part of the pairs' shared lexicon. The following qualitative descriptions of the pairs' performance detail these differences.

4.3.2.2 Conversational excerpts

Figure 4-4 presents an example of the types of problems that arose for both Primary and Plaid puzzles when the delay was in the range of [939-1798ms]. In this range, the pairs demonstrated a number of coordination errors that signified misaligned awareness of one another's task state. In this example, the Helper describes a piece and where to put it (line 1). However, the delayed

visual feedback causes him to reiterate his directive (line 3), since he assumes his partner did not hear or did not understand. However, when the Worker hears this, her puzzle state already indicates the correct move (line 3), and therefore she interprets his reiteration as a clarification and incorrectly adjusts the piece to the lower left of the workspace (line 4). The Helper then sees the delayed view, believes everything is fine, and confirms the placement (line 5). Unfortunately, the Worker believes this confirmation refers to her new placement. Shortly thereafter, the Helper sees the incorrect move and they begin a repair sequence. This example demonstrates how the delay led to misaligned views of the task state, ultimately resulting in coordination problems that harmed task performance.

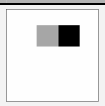
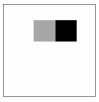
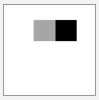
Worker View	Helper View	Speaker	Speech / [*Actions*]
		*Helper*	Now take that one right to the left of it and put it in the bottom left hand corner
		*Worker*	[Correctly moves piece to bottom left of previous piece]
		*Helper*	Move it to the bottom left corner
		*Worker*	[Incorrectly re-positions piece to the bottom left of the workspace]
		*Helper*	OK, now…
		*Helper*	…no, no, no

Figure 4-4. Excerpt demonstrating a coordination error resulting from a lack of shared situation awareness (at a delay of approximately 1100ms).

In the highest range of delays [1798ms – 3300ms], the differences between views becomes readily apparent, and the pairs demonstrate a strategic shift whereby they exhibit fewer behaviors that rely on tight integration between speech and visual information. At these higher levels of delay, the pairs tended to try to complete the puzzles simply using spoken language. This was evidenced by the relative lack of use of deictic pronouns such as “that one” and “this.” Instead, the pairs relied on lengthier verbal descriptions to describe the objects and their arrangement. However, this posed a greater problem for Plaids than it did for the Primary pieces. When

describing the solid pieces, where names were a part of their shared lexicon, pairs could quickly describe the colors and where to place them while using verbal acknowledgements to keep on track, reserving use of the visual information for delayed confirmation that the task had been performed correctly (see Figure 4-5). However, when negotiation of the names of the pieces was required, as was the case with the Plaids, the inefficiencies of the linguistic medium became much more evident. As can be seen in Figure 4-6, the Helper and Worker both became much more active in negotiating their descriptions. However, when the visual evidence was needed for disambiguation or confirmation, they had to wait to receive the information. This shift in strategy likely led to the additional impact of delay on task performance for the Plaids over the Primary colors in the high range of delays.

Primary Pieces
H: then there's like a-a red **W:** okay **H:** make it touch the corner **W:** okay **H:** then there's another red but it's more rosy... **W:** rosy, okay **H:** make it touch the red one on the left hand side...

Figure 4-5. Excerpt demonstrating grounding difficulties in the Plaids pieces at a delay of approximately 2700ms.

Plaid Pieces
H: um horizontal white stripe... **W:** any blue? **H:** and two and two...and two like hor- vertical gray stripes **W:** horizontal white stripe with two vertical gray stripes? **H:** yeah **W:** this one? [pause 2.5 sec] **H:** uh...no... **W:** oh... **H:** it's horizontal white stripe and two vertical stripes... **H:** yeah that one

Figure 4-6. Excerpt demonstrating grounding with the easier Primary pieces at a delay of approximately 2700ms.

4.4 Study 3: The impact of task dynamics and visual delay

While the results in Study 2 are consistent with the findings in the earlier work presented in Chapter 3 (Gergle et al., 2004b; Kraut et al., 2002b) they do not necessarily align with the timings presented by Gutwin and colleagues (Gutwin, 2001a; Gutwin et al., 2004). The work by Gutwin and colleagues found that collaboration difficulties occurred at delays of much shorter duration. For example, they began to find increased errors in some conditions with delays as short as 200ms, and at 400ms they tended to find a significant impact of visual delay on task coordination. However, one must keep in mind that the types of tasks investigated in these studies tended to focus less on the language generated around a shared visual environment and more on the task awareness afforded by the displays for tightly-coupled motor-based coordination activities. This study aims to clarify this seeming inconsistency by examining how the dynamics of the task objects interact with the visual delay to impact the coordination mechanisms required for successful collaborative performance. It also demonstrates how the dynamics of the task interact with the amount of visual delay that can be tolerated before impacting task performance.

4.4.1 Method

The amount of visual delay present in the Helper's view of the workspace (Visual Delay) and the dynamics of the task objects by providing puzzle pieces that changed colors at different rates (Object Dynamics) were manipulated.

4.4.1.1 Independent variables

Visual Delay [100-3119ms]: This factor used a similar distribution as the one described in Study 2. However, the initial delay was set at 100ms and the times were generated according to the following distribution:

$$f(n) = \begin{cases} T_1 = 100 \\ T_n = T_{n-1} \cdot e^{.05} \end{cases}$$

These times were temporarily slotted into three sub-ranges for assignment. *Low* delay was the range of [100-306ms], *Medium* delay was [319-977ms], and *High* delay was [1020-3119ms]. Participants were selected to receive two levels from each bin and these were crossed with similarly binned levels of the Object Dynamics.

Object Dynamics (Moderate, Fast, and Very Fast): The dynamic complexity of the task objects was manipulated by allowing the colors of the blocks to cycle. Each piece changed its color,

smoothly moving through the color palette. At the *Moderate* cycle rate, the pieces experienced a major perceivable color change (e.g., from "red" to "orange", or "blue" to "purple") approximately every 6-8 seconds. It took roughly one second of continuous observation to notice whether any given piece was changing. In the *Fast* cycle rate, the pieces achieved a major perceivable color change approximately every 2-3 seconds. While at the *Very Fast* cycle rate, the pieces rapidly changed color at a rate of approximately one perceivable change every second or less. It should be noted that these values fluctuate somewhat due to the fact that people do not perceive change equally across the color spectrum.

4.4.1.2 Participants and procedure

Participants consisted of 27 pairs recruited from the Pittsburgh area. They were randomly assigned to play the role of Helper or Worker and the pairs were balanced by gender. *Visual Delay* [60-3300ms] and *Object Dynamics* (Moderate, Fast, Very Fast) were manipulated within each pair. Each pair participated in a total of nine experimental conditions that varied across a range of delays crossed with a range of object cycle rates. Pairs solved four puzzles within each experimental condition. This resulted in a total of 36 puzzles that were completed in approximately an hour and a half.

4.4.1.3 Measures

This study used the same measures of task performance as in Study 2, once again choosing task completion time over errors, since the number of final errors was very low.

4.4.1.4 Statistical methods and analyses

The analyses in this study are the same as those described in Study 2, with one exception. We ran separate models for each level of the Object Dynamics (Moderate, Fast, and Very Fast) in order to discover the optimal breakpoints for each object cycle rate.

4.4.2 Results

For ease of exposition, the results focus on a description of the overall model fits, breakpoints, and the slopes of the initial two segments[12]. This clearly demonstrates how the dynamics of the environment shift the range of tolerable delays when a more dynamic environment is in play (as previewed in Figure 4-7).

[12] Detailed results from the piecewise linear regression models are included in Appendix F.

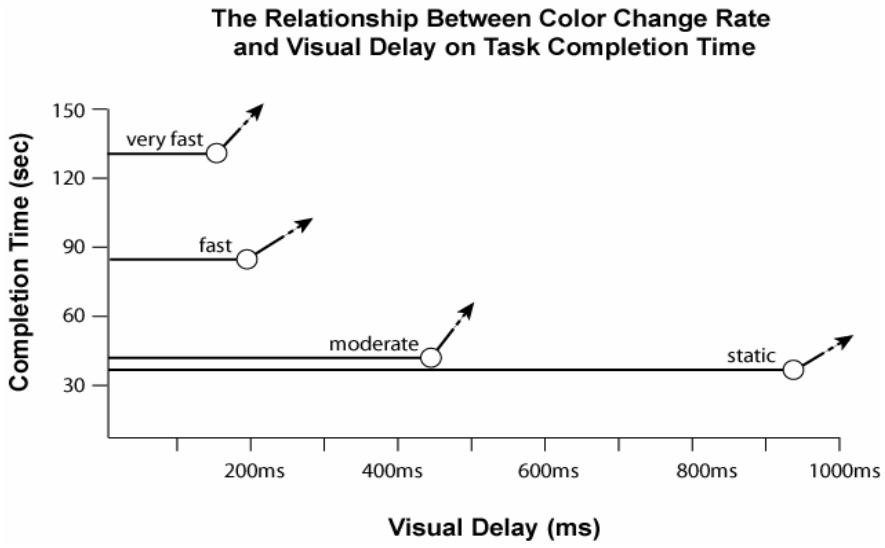


Figure 4-7. This illustration presents a stylized view of the data. It shows the initial breakpoints (circles) across a range of color dynamics. Lines up to the breakpoints are slopes not significantly different from zero, and the subsequent trajectories represent slope changes. From top-to-bottom the lines represent the three speeds at which the colors changed: Very Fast, Fast, Moderate (Study 3), and Static (Study 2).

4.4.2.1 Model of the moderate change rate

The model of the *Moderate* change rate revealed two optimal breakpoints in delay at 431ms and 558ms. These results were then used in a random effects piecewise linear regression model where [*Visual Delay*], [*Visual Delay – 431ms*], [*Visual Delay – 558ms*], Block (*1-9*), and Trial (*1-4*) were repeated factors. We included all 2- and 3-way interactions in the analysis. Because each pair participated in 36 trials, observations within a pair were not independent of one another and were modeled as a random effect. The model achieved a good fit to the data (*Adj-*R^2 = .522; *GCV-*R^2 = .498; $p < .001$).

Examination of the impact of visual delay in the *Moderate* condition revealed no influence on task performance when the delay was below 431ms (β = -2.16(13.9), $F_{(1,265)}$ = .024, p = .88). In other words, the slope for the range of delays between 100ms and 431ms was essentially flat. Note, however, that the dynamics of task objects did substantially shorten the range of tolerable delays in comparison to those found in Study 2, when the pieces were non-changing solid colors (see Figure 4-7).

However, when the delay reached 431ms there was a significant impact on task performance ($F_{(1,265)} = 8.26$, $p = .004$). At this point, there was evidence of a drastic impact of the visual delay on task performance, with every 100ms increase in visual delay increasing the pairs' completion time by approximately 14 seconds.

4.4.2.2 Model of the fast change rate

For the Fast paced changing objects the optimal breakpoints were found to be at 191ms and 1783ms. We used the previous model with the following changes: [*Visual Delay*], [*Visual Delay – 191ms*], [*Visual Delay – 1783ms*]. The model fit the data well (*Adj-*$R^2 = .528$; *GCV-*$R^2 = .442$; $p < .001$).

An examination of the influence of visual delay at the *Fast* level showed that the range of tolerable delays was greatly reduced in comparison to the *Moderate* change rate (see Figure 4-7). When the delay was under 191ms, there was no evidence of an influence of delay on task performance (β = -68.4(60.6), $F_{(1,278)} = 1.288$, $p = .26$). Once again, the slope for this initial segment was essentially flat.

Once the delay reached 191ms, the trend towards an impact on task performance appeared in the hypothesized direction ($F_{(1,278)} = 2.22$, $p = .14$). However, while the plots appear to indicate an upward swing consistent with the other models, indicating an impact of delay on task performance, the slope of this shift was not significant, as it was in all the other models. This may be due in part to an increase in the amount of noise in the data given the increasing complexity of the task.

4.4.2.3 Model of the very fast change rate

For the *Very Fast* paced changing objects the optimal breakpoints were found to be at 154ms and 450ms. The above model was used with the following changes: [*Visual Delay*], [*Visual Delay – 154ms*], [*Visual Delay – 450ms*]. The model achieved a reasonable fit to the data (*Adj-*$R^2 = .443$; *GCV-*$R^2 = .367$; $p < .001$).

Here, the range of tolerable delays was smaller than that of any other condition (see Figure 4-7). For delays under 154ms, the slope was again flat and there was no evidence of the impact of delay on task performance (β = -341.2(215.9), $F_{(1,254)} = 2.50$, $p = .12$). This suggests that there was no impact of the delay on performance in the range between 100 and 154ms.

However, once the delay reached 154ms, there appeared to be a marginal impact on task performance ($F_{(1,254)} = 3.20$, $p = .08$). At this point, there was evidence of a drastic impact of the visual delay on task performance.

4.5 Discussion

These studies demonstrate the application of a statistical method that allows the examination of collaborative task performance over a continuous range of visual delays. This method provides detailed insight into the range of delays within which collaborative task performance is not affected, as well as uncovers the points at which performance begins to break down. In addition, examination of the corresponding slope coefficients provides an indication of the relative impact of additional delays on performance. This method allows us to extend the findings of earlier work that examined discrete levels of delay but could not pinpoint the precise time at which collaborative performance broke down in the presence of delayed visual information (e.g., see Gergle et al., 2004b; Gutwin, 2001a; Gutwin et al., 2004; Kraut et al., 2002b).

Study 2, using static primary colored task objects, found that the amount of visual delay had no impact on task performance when it was less than 939ms. However, in the range between 939 and 1798ms the delay impacted the Primary and Plaid puzzles equally. The conversational transcripts suggest that these deficits in performance may be due in part to the fact that the coordination processes supported by shared situation awareness are disrupted. For example, the ability of the Helper to successfully plan an utterance based on an assessment of the current state of the puzzle, appears to be disrupted. This is dramatically demonstrated when the information leads to misalignments in a pair's model of the current state of the shared task (as was shown in Figure 4-4). Such misalignments, or inaccurate mental representations of task state, can severely impact coordination on the part of the pairs.

At delays greater than 1798ms, the impact of the delay seemed to shift to conversational grounding processes. This was evidenced by the fact that the [*Visual Delay – 1798*] × *Linguistic Complexity* interaction was significant, and that there remained an increasing slope for the Plaid pieces while the slope for the Primary pieces leveled off. Similar to the findings presented in Chapter 3 , the transcripts revealed that this may be due to the fact that the pairs simply resorted to using linguistic terms to describe the primary objects and their placement and only used the visual information for delayed confirmation. However, when attempting to use this strategy to

describe the Plaid pieces, the pairs suffered a much greater penalty for not being able to use the efficiencies of the visual space to support grounding on the object terms. Instead, they had to use rather inefficient linguistic descriptions in an attempt to ground on terms that represented the Plaid pieces. In this case, the pairs relied much more on the visual information to play a role in disambiguation and comprehension monitoring.

In Study 3, as can be seen in Figure 4-7, and in support of *H5* and *H6*, when the dynamics of the task objects increased, the visual delay began to have an impact at much shorter time intervals. This was demonstrated by the tendency of the first breakpoints to move closer to the 100ms lower bound. Together, these results provide evidence that the dynamics of the task objects and environment have a major impact on the range of delay that can be tolerated before collaborative task performance begins to suffer. In the moderately dynamic environment the pairs could accommodate up to a 431ms delay. However, as the dynamics of the task approached the fast rate, the pairs appeared to suffer performance deficits once the delay reached 191ms, and at the very fast dynamic rate the pairs could only tolerate delays up to 154ms before their performance degraded.

The range of tolerable delays found in Study 3 appear to be much more in line with those described in Gutwin's work (Gutwin, 2001a; Gutwin et al., 2004). This is likely due, in part, to the nature of Gutwin's tasks. As previously described, the tasks used in his work were primarily motor or physical tasks with a strong coordination component. In theoretical terms, these tasks require a precise knowledge of the current state of the task in order to be successfully performed. Therefore, it is likely that the disruption caused by the latency primarily impacts the pairs' ability to maintain an accurate model of the shared state of the task, similar to the way the dynamic task objects impact performance in Study 3. However, the Gutwin tasks require little use of language, and as such, the impact that delay has on conversational grounding is not seen.

Together, these results suggest it is not as simple as picking a single number to serve as a hard threshold for dictating whether or not a given delay is tolerable for collaborative task performance. Instead, a detailed task-analysis needs to be performed in order to establish the collaborative requirements of the task.

4.6 Conclusion

This chapter examined the effect that delayed visual feedback has on collaborative task performance. The results demonstrate that a number of factors come into play when assessing a tolerance for visual delay. An understanding of the complexity and dynamics of the task environment, of the degree to which the collaborative pairs rely on situation awareness to perform their tasks, and of the amount of visual and domain context they share, are all keys to determining how well a given technology may serve a particular group.

Up to this point, two theories of collaborative behavior have been used to inform our understanding of the ways in which shared visual information supports collaborative performance. However, each of these theories claims that the advantages of shared visual information come about for different reasons. The next chapter takes a more detailed look at the theoretical coordination mechanisms that play a role in successful collaborative behavior, and how various parameterizations of shared visual information independently impact these coordination mechanisms.

Chapter 5

Shared Visual Information for Grounding and Awareness

As demonstrated in the prior chapters, when pairs work together on a physical task the ability to see a common workspace facilitates communication and ultimately benefits their performance. However, when mediating such activities, the choice of technology can transform the visual information in ways that impact critical coordination processes. This chapter explores two coordination processes that are impacted by visual information: situation awareness and conversational grounding. While these coordination mechanisms are theoretically distinct, they are often confounded in empirical research.

The following presents three studies that demonstrate how shared visual information supports collaboration through the independent mechanisms of situation awareness and conversational grounding. In addition, the studies address how particular features of visual information interact with features of the task to influence situation awareness and conversational grounding. Study 4 replicates the findings in previous chapters and clarifies how immediate visual feedback facilitates collaboration by improving both situation awareness and conversational grounding. In Study 5, misaligning the perspective through which the Worker and Helper see the work area disrupts the ability of visual feedback to support conversational grounding but not situation awareness. The results demonstrate that visual information supports the central mechanism of conversational grounding. Study 6 impedes the ability of visual feedback to support situation awareness by reducing the size of the common viewing area.

5.1 Introduction

Previous chapters have described a decompositional framework for understanding the ways in which visual information affects collaboration. This work suggested that the degree to which visual information will improve performance in any particular situation depends both on *technological choices* and *the task the group is performing*. Technological choices influence the amount and quality of visual information exchanged. For example, instructors provide better guidance on a robot construction task when using a scene-oriented camera with a wide-angle view of the work area than when using a head-mounted camera that shows a narrow, dynamic view of the work area (Fussell et al., 2003a). Task features also influence whether visual information improves performance (Whittaker et al., 1993). For example, visual feedback helps collaborators more when they are working with objects that are difficult to describe than when they are working with objects that are easy to describe (Gergle et al., 2004b; Kraut et al., 2002b). This work, and others like it (Clark & Krych, 2004; Velichkovsky, 1995), demonstrates the need for a more nuanced theoretical understanding of the precise functions served by visible information in collaboration (for further discussions see Whittaker, 2003; Whittaker & O'Conaill, 1997).

Our framework is based on two psychological theories that help explain the role of visual information in collaborative work. First, *Situation Awareness Theory* (Endsley, 1995; Endsley & Garland, 2000) holds that visual information helps pairs assess the current state of the task and plan future actions. For example, a teacher watching over a student's shoulder might intervene to provide timely instructions because she can see from the calculations that the student has not mastered necessary algebraic equations. *Grounding Theory* (Clark, 1996; Clark & Marshall, 1981; Clark & Wilkes-Gibbs, 1986) maintains that visual information can support the conversation surrounding a joint activity by providing evidence of common ground or mutual understanding. For example, a teacher may clarify her instruction after seeing the student's calculations because she can see that the student had misunderstood something she said. Together these theories predict that when groups have access to the correct visual information they are better able to coordinate their work because they can monitor the state of the task, deliver instructions and clarifications in a more timely fashion, and refer to objects and actions more efficiently.

However, while visual information is thought to influence both situation awareness and conversational grounding, most empirical research has failed to distinguish these conceptually

distinct coordination mechanisms. The distinction is important for developing accurate theoretical models of the role of visual information in collaborative work and for building systems that provide the right type of visual information for the task at hand.

A major goal of this chapter is to provide empirical evidence that visual information can improve collaboration through these *two distinct routes*. A secondary goal is to answer the pragmatic question regarding how particular features of visual information interact with features of the task to influence situation awareness and conversational grounding.

This chapter begins by further detailing the theoretical foundation for this work, and describing the necessity of visual information for supporting both situation awareness and conversational grounding. Three studies using the puzzle task paradigm attempt to disentangle the independent effects of situation awareness and conversational grounding. Study 4 manipulates the immediacy of the visual information and shows that immediate visual feedback facilitates collaboration by improving both situation awareness and conversational grounding. Study 5 disrupts the ability of visual feedback to support conversational grounding, but not situation awareness, by misaligning the perspective through which the Worker and Helper see the work area. The results demonstrate that conversational grounding is a central mechanism supported by visual information. Finally, Study 6 impedes the ability of visual feedback to support situation awareness by selectively reducing the size of the common viewing area. The findings suggest that visual information independently supports both situation awareness and conversational grounding. The chapter concludes with a general discussion of the results and their implications for theory development and system design.

5.2 The role of visual information in supporting collaboration

In this section we present a brief overview of Situation Awareness Theory (Endsley, 1995) and Grounding Theory (Clark & Marshall, 1981; Clark & Wilkes-Gibbs, 1986), focusing on the ways that visual information improves collaborative performance via these mechanisms.

5.2.1 Situation awareness

According to Situation Awareness Theory, visual information improves collaborative performance by giving actors an accurate view of the task state and each others' activities. This awareness allows the accurate planning of future activities (Endsley, 1995). However, as long as the visual information allows them to form an accurate view of the current situation and

appropriately plan future actions, it does not need to be identical for all group members in order to support situation awareness (Bolstad & Endsley, 1999). For example, two fighter pilots can converge on and shoot down another aircraft, even if one of them uses the visual line of sight and the other uses radar to "see" the target. However, if the differing displays lead them to form different situational representations, then their performance is likely to suffer. For example, if visual sighting allows one pilot to distinguish between friendly and enemy aircraft, but the radar fails to support this discrimination for the other pilot, then the two fighters are unlikely to successfully coordinate their attack purely on the basis of the situation awareness provided by visual information (Snook, 2000).

5.2.2 Conversational grounding

According to Grounding Theory, visual information can improve coordination by supporting the verbal communication surrounding a collaborative activity. Grounding Theory states that successful communication relies on a foundation of mutual knowledge or common ground. Speakers form utterances based on their expectation of what a listener is likely to know and then monitor that the utterance was understood, while listeners have a responsibility to demonstrate their level of understanding (Clark & Marshall, 1981; Clark & Wilkes-Gibbs, 1986). Conversational grounding is the *process* of establishing common ground.

Throughout a conversation, participants are continually assessing what other participants know and using this knowledge to help formulate subsequent contributions (Clark & Marshall, 1981; Clark & Wilkes-Gibbs, 1986). Clark & Marshall (1981) propose three major factors that allow speakers to anticipate what a partner knows: *community co-membership*, *linguistic co-presence*, and *physical co-presence*. Because of community co-membership, members of a professional group, for example, can use technical jargon with each other that they could not use with outsiders. Because of linguistic co-presence, one party in a conversation can safely use a pronoun to refer to a person previously mentioned in the conversation. Because of physical co-presence, one person can point to an object in their shared physical environment and refer to it using the deictic pronoun "that" if she believes the other can also see the object and her gesture.

Shared visual information helps communicators to establish common ground by providing evidence from which to infer another's level of understanding. This evidence can be demonstrated deliberately (e.g., as in a pointing gesture) or as a side effect of proper performance of the desired action, provided both parties are aware of what one another can see. When

responding to an instruction, performing the correct action without any verbal communication indicates understanding, while performing the wrong action or failing to act can signal misunderstanding.

Visual information can support conversational grounding at two distinct phases of the communication process: the *planning stage* and the *acceptance stage* (Clark & Schaefer, 1989). During the *planning* phase, in which speakers formulate their utterances (Levelt, 1989), visual information provides cues to that which a listener is likely to understand. In the puzzle paradigm, Helpers need to refer to puzzle pieces so that Workers can identify them easily. If Helpers can see the work area and are aware that the Worker can also see it, they can use the mutually available visual information to help describe the piece. For example, when describing a plaid piece they can use efficient expressions such as, "the one on the left" rather than lengthier descriptions of the patterns contained within a particular piece. Similarly, they can reduce verbal ambiguity by using the phrase, "the dark red one," when they can see that both dark and light red pieces are visible to the Worker.

During the *acceptance* phase, speakers and hearers mutually establish that they have understood the utterance well enough for current purposes (Clark & Wilkes-Gibbs, 1986). In the puzzle paradigm, Helpers can use visual feedback from the Worker's performance to monitor whether the Worker has understood the instructions. This visual feedback is efficient because with it the Worker does not need to explicitly state his or her understanding (see for example, Doherty-Sneddon *et al.*, 1997; Gergle et al., 2004a). It is also less ambiguous than verbal feedback, because listeners may not know they have misunderstood an utterance. Clark and Krych (Clark & Krych, 2004) demonstrated that when shared visual information was available, pairs spent approximately 15% less time checking for comprehension (see also Doherty-Sneddon et al., 1997).

In most real-world settings, visual feedback provides evidence of both the current state of a task and a listener's degree of comprehension. As a result it is often difficult to *empirically distinguish* the routes through which visual information improves collaborative performance. The experiments reported below are designed to empirically demonstrate that visual information improves performance on collaborative tasks by supporting both situation awareness and conversational grounding.

5.2.3 The impact of technological mediation on the availability of visual information

Although visual information in general can improve collaborative task performance by improving situation awareness and conversational grounding, the benefit it provides in any particular situation will likely depend on the technology used and the characteristics of the collaborative task. For designers and engineers creating technologies to provide visual information through telecommunications, the goal is to make a collaborative environment as similar as possible to the gold standard of physical collocation. In attempting to reach this goal, however, they must trade off features that shape the usefulness of the visual information, such as field of view and who controls it, delays, alignment of perspective, degree of spatial resolution, frame rate, and level of synchronization with a voice stream. These different features of the communication media change the costs of grounding and situation awareness (Clark & Brennan, 1991; Kraut et al., 2002a). How do we know which of these features need to be reproduced in order to provide the benefits of a collocated environment?

The puzzle study paradigm provides a method for decomposing the visual space to better understand the impact of various features of the visual information on collaborative performance. The experiments reported here examine the impact of particular media features such as delay, perspective, field of view and view control, in addition to distinguishing between the coordination mechanisms of situation awareness and conversational grounding.

5.2.4 Overview of experiments

The following work presents a series of three experiments that are intended to disentangle the effects of visual information on conversational grounding and situation awareness. As shown in Table 5-1, the three experiments manipulate different features of the visual environment. Study 4 manipulates the immediacy of the visual information, with the Helper seeing the Workers' work area either immediately, after a delay, or not at all. The results are consistent with the hypothesis that immediate visual feedback helps collaborative performance by improving both situation awareness and conversational grounding. However, this manipulation does not distinguish between these two mechanisms, because delay can disrupt both situation awareness and grounding as seen in Chapter 3 and Chapter 4.

Table 5-1. Overview of studies and manipulations presented in this chapter.

Study	Features of Shared Visual Information				Task Feature
	Immediacy	*Perspective Alignment*	*Field of View*	*Field of View Control*	*Lexical Complexity*
Study 4: Replication Study	X				X
Study 5: Rotation Study	X	X			X
Study 6: Field of View Study			X	X	X

Study 5 impedes the ability of visual feedback to support conversational grounding by misaligning the perspective through which the Worker and Helper see the work area. This misalignment makes it difficult for pairs to describe the puzzle pieces and puzzle layout using a common spatial vocabulary. If visual feedback improves collaborative performance in this case, it does so primarily through situation awareness.

Finally, Study 6 impedes the ability of visual feedback to support situation awareness by reducing the size of the available common viewing area. As a result, the Helper has difficulty keeping track of the puzzle layout as it is being constructed. This manipulation, however, does not greatly interfere with the pairs' ability to develop a common vocabulary for identifying and describing the pieces in the shared environment. If visual feedback improves collaborative performance when the Helper can only see a small section of the field of view, it does so through conversational grounding, by supporting the pairs' ability to easily refer to puzzle pieces. Study 6 also manipulates whether the Worker or Helper has manual control over the work area or whether the field of view automatically tracks the Workers' actions. The implications of these manipulations are detailed in the subsequent study descriptions.

5.3 Study 4: Replication study

Study 4 is a replication study used to take a closer look at the way shared visual information impacts both conversational grounding and situation awareness. Since the visual information improves both situation awareness and conversational grounding, pairs who have visual feedback should perform better in the puzzle experiment, completing the puzzles more quickly, and the work in this section allows a more detailed qualitative examination of its impact on these coordination mechanisms.

Prior work demonstrated two facets of task objects that have an impact on the collaborative referring procedure. These are *discriminability*, how easy it is to linguistically differentiate an object from other available objects based on its visual features, and *codability*, how easy it is to initially describe or name an object (Hupet *et al.*, 1991). Visual information should have the most benefit when codability is low. Without visual feedback, collaborators must use language to describe task objects. When discriminability is low, the referring experience will be less efficient and more ambiguous, leading to problems in the initial planning of an utterance and to more opportunities for misunderstanding. In this experiment, we manipulate the overall lexical complexity of the task objects by using either simple primary colors that have high codability and high discriminability, or by making the task objects tartan plaids, which have low codability and low discriminability.

At the technological level, this experiment examines how delays in the availability of the visual feedback—of the sort introduced by video compression or network lags—are likely to undercut its value. As seen in previous chapters, delay in visual updating reduces the value of shared visual information. Collaborators in face-to-face settings use visual information to precisely time when they will provide new information and to change speech in mid-sentence in response to their partner's gaze (Boyle et al., 1994) or behavior (Clark & Krych, 2004). The study presented in Chapter 4 varied the availability of the visual feedback on a continuous range between 60ms and 3300ms. Breakdowns in grounding and situation awareness tended to occur when the delay was greater than 950ms. This study used a delay of 3000ms, a number well above the previous threshold found for disruptive collaborative performance and discourse.

Along with differences in task performance, we expect to see differences in the ways that pairs adapt their discourse structure to make use of the visual information provided to complete the task. If visual information benefits task performance through situation awareness, Helpers who can receive visual feedback should more quickly introduce instructions for a step after a Worker has completed a prior instruction. In addition, they should more readily identify errors or deviations from the optimal solution path and efficiently correct these problems.

If visual information benefits task performance by facilitating conversational grounding, participants should spend less time requesting and giving confirmation that they have understood their partners' utterances (Brennan, 1990, 2005; Clark & Krych, 2004; Fussell et al., 2000). In addition to this, the principle of least collaborative effort (Clark & Wilkes-Gibbs, 1986) suggests

that pairs should change the structure of their discourse in order to expend the least amount of effort for the group as a whole (Kraut et al., 2002b). Therefore, both the Helpers and Workers should be influenced by the presence of visual feedback, even though only the Helpers see it.

The following hypotheses summarize this discussion:

> *H1: A collaborative pair will perform their task more quickly when they have a shared view of the work area.*
>
> *H2: A collaborative pair will perform their task more slowly as the linguistic complexity of the task increases.*
>
> *H3: A shared view of the work area will have additional performance benefits when the linguistic complexity of the task objects increases.*
>
> *H4: Delay in transmission of the shared visual information will weaken the value of a shared view of the work area.*

5.3.1 Method

Pairs of participants played the puzzle study described in Chapter 2 and were randomly assigned to play the role of "Helper" or "Worker". Each participant was seated in a separate room in front of a computer with a 21-inch monitor. They communicated over a high-quality, full-duplex audio link with no delay.

5.3.1.1 Independent variables

This study manipulated whether the Helper viewed the same work area as the Worker, and if so, how quickly the visual information was delivered (*Immediacy of Visual Feedback*). *Lexical Complexity* manipulated the adequacy of lexical tokens to describe the puzzle pieces.

Immediacy of Visual Feedback (*Immediate* vs. *Delay* vs. *None*): In the immediate visual feedback condition (*Immediate*), the Helper's view of the work area was identical to the Worker's work area, with no perceptual delay. In the delayed condition (*Delay*), the Helper saw the Worker's work area with a 3-second delay. In the no visual feedback condition (*None*), the Helper's view was black.

Lexical Complexity (*Primary* vs. *Plaid*): The lexical complexity manipulation provided pairs with different types of puzzle pieces. The colors of the pieces were either lexically simple, easy to describe primary colors (e.g., red, yellow, orange, etc.), or they were more complex visual patterns (e.g., tartan plaids), that required the negotiation of a common naming convention for the pieces (i.e., they were not initially part of a shared lexicon) (see Figure 4-1).

5.3.1.2 Participants and procedure

Participants consisted of 12 pairs of Carnegie Mellon University undergraduate students. Participants received $10.00 for their participation. They were randomly assigned to the role of Helper or Worker. The immediacy of the visual feedback and the visual complexity were manipulated within the pairs, while the lexical complexity was a between-pair factor. Each pair participated in six blocks of four trials. They completed a total of 24 puzzles in approximately one hour.

5.3.1.3 Measures and statistical analysis

The pairs were instructed to complete the task as quickly as possible, so task performance was the time it took to properly complete the puzzle. Because the vast majority of the puzzles were solved correctly and differences in error rates among conditions were minor, we focus on completion time as our primary measure of task performance.

The analysis is a mixed model analysis of variance in which Block (*1-6*), Trial (*1-4*) and Immediacy of the Visual Feedback (*Immediate*, *Delayed*, *None*) were repeated within-subject factors, and Lexical Complexity (*Primary* or *Plaid*) was a between-pair factor. We included 2-way and 3-way interactions in the analysis. Because each pair participated in 24 trials (6 conditions by 4 trials per condition), observations within a pair were not independent of each other. Pairs, nested within Lexical Complexity, were modeled as a random effect.

5.3.2 Results and discussion

5.3.2.1 Task performance

Immediacy of visual feedback. Consistent with *H1*, a shared view of the work area benefited performance. The pairs were approximately 30% faster at completing the puzzles when they were

in the Immediate Shared Visual Space condition (M = 51.27s, SE = 4.12) than in the No Shared Visual Space condition (M = 74.63s, SE = 4.03), $F_{(1, 266)} = 47.43$, $p < .001$[13]. Consistent with *H4*, a 3-second delay in updating the shared visual information considerably reduced its benefits. The Delayed Shared Visual Space condition (M = 69.04s, SE = 4.12), was only 7% faster than the No Shared Visual Space condition, $F_{(1, 266)} = 2.71$, $p = .10$.

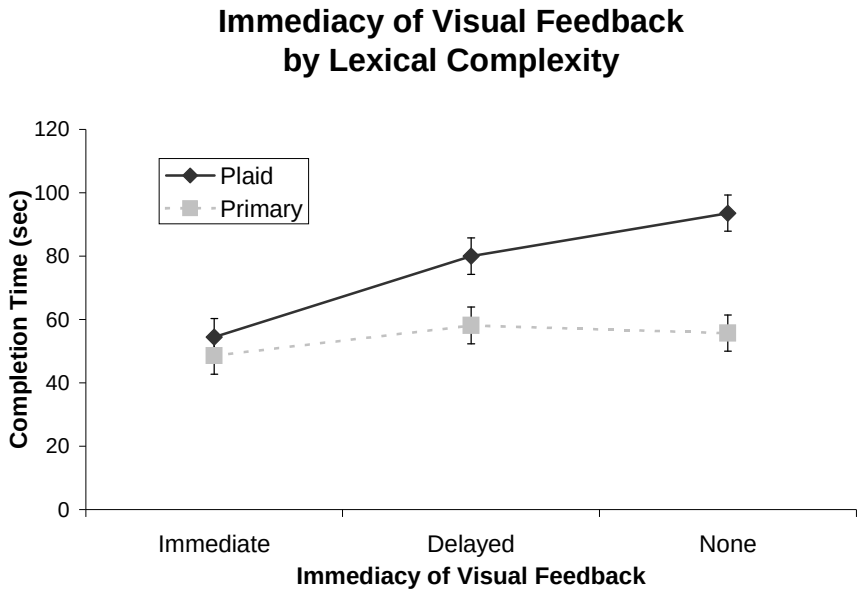


Figure 5-1. Shared Visual Space by Lexical Complexity on task completion time (all figures show LSMeans ±1 SE)

Linguistic complexity. Consistent with *H2*, linguistic complexity substantially increased completion times. The pairs were approximately 30% faster in trials where the pieces were easy-to-name primary colors (M = 53.95s, SE = 5.04) than when they were more complex plaids (M = 76.0s, SE = 5.04, $F_{(1, 10)} = 9.62$, $p = .011$).

Consistent with *H3*, the visual information had the greatest benefit in the plaid condition (see Figure 5-1), when puzzle pieces were linguistically complex and difficult to describe. The Immediacy of Visual Feedback × Linguistic Complexity interaction, testing whether the linear

[13] Full statistical details for the models in this chapter can be found in Appendix F.

effect of the immediacy of the visual feedback was greater for plaid pieces than for primary colored pieces, was highly significant, $F_{(1,266)} = 66.40$, $p < .001$. A detailed examination of this interaction and effect sizes reveals that the pairs took much longer to complete the task for plaids than for primary colors in the No Shared Visual Space condition, $F_{(1,266)} = 22.05$, $p < .001$, $d = .58$. They also took longer to complete the task for plaids than for primaries in the Delayed Shared Visual Space condition, however, the effect size was substantially smaller, $F_{(1,266)} = 7.26$, $p < .008$, $d = .33$. This difference all but disappeared in the Immediate Shared Visual Space condition, $F_{(1,266)} = 0.64$, $p = .426$, $d = .10$.

5.3.2.2 Communication processes

Previous work has detailed how discourse structure changes when shared visual information is available. Immediately available visual information about the work area yields lower rates of spoken discourse since communicators rely instead on more efficient visual information (Boyle et al., 1994; Brennan, 1990, 2005; Clark & Krych, 2004; Daly-Jones et al., 1998; Fussell et al., 2000; Kraut et al., 2002a). Visual information is also useful for supporting efficient referring expressions (Brennan & Lockridge, In preparation; Clark & Krych, 2004; Fussell et al., 2000; Kraut et al., 2002b). It also provides evidence of understanding (conversational grounding) as well as unambiguous information about the state of the task (situation awareness). This was shown in Chapter 3 by demonstrating that visual information is used in place of verbal acknowledgements of understanding—a vital part of the grounding process—as well as in place of explicit verbal acknowledgement that a particular task had been completed in order to support situation awareness (Gergle et al., 2004a, 2004b; Kraut et al., 2003; Kraut *et al.*, 1996). The following excerpts from the current study demonstrate many of these processes:

Immediate Visual Feedback, Plaid Pieces
1. **H:** the first one is gray, gray lines on the top and brown lines on the left
2. **W:** *[moves correct piece]*
3. **H:** put it on the right middle corner
4. **H:** yeah perfect
5. **H:** uh take it up slightly
6. **H:** and the second one is uh two blue vertical bands
7. **H:** a lot of light gray err light blue lines
8. **W:** *[moves correct piece]*
9. **H:** take it half a block down
10. **H:** to… yeah.

Figure 5-2. Immediate visual feedback and Plaid pieces.

No Visual Feedback, Plaid Pieces	
1.	**H:** the last one is
2.	**H:** the, it has two light blue …ah… big stripes going up the sides with…
3.	**H:** with a like vertical royal blue up the middle like
4.	**W:** it just has…
5.	**H:** the background is royal blue
6.	**W:** does it just have one, one
7.	**H:** just one royal blue up the middle
8.	**W:** *[moves correct piece]*
9.	**W:** I got it
10.	**H:** and it has two hash marks going through the middle horizontally
11.	**W:** yeah, I got it
12.	**H:** yeah, that goes directly to the left of the the, that last one I just told you
13.	**W:** ok, done

Figure 5-3. No visual feedback and Plaid pieces.

Figure 5-2 provides a snippet of the interaction that takes place when the pairs have immediately available shared visual information in the lexically complex, plaid condition. In line 1, the Helper begins by generating a referential description of a puzzle piece. The Worker demonstrates her understanding of the intended referent by moving a piece into the shared view (line 2). Contrast this interaction with that shown in Figure 5-3, where there is no shared visual information available to the pair. In this case, the Worker becomes much more active in the negotiation of a shared understanding (lines 4 and 6), and he provides explicit confirmation that he understands the intended referent by declaring, "I got it" (line 9 and again in 11). Together these excerpts demonstrate how the pairs use the shared visual information for efficient task performance and to support the coordination mechanism of conversational grounding.

These same excerpts also demonstrate the use of visual information to support situation awareness. When visual information is available, the Helper uses it to determine when one subtask is completed and the next should be started. In Figure 5-2, once the Worker moves the correct piece into the shared display area, the Helper instantly provides the next instruction describing where to place the active piece. This same trend can be seen again at lines 6, 7, 8 and 9, when the Helper describes the piece, the Worker places it in the work area, and the Helper immediately instructs the Worker on where to place it. In contrast, without the visual feedback, the Helper must rely upon the Worker's explicit declaration that he has finished a subtask, and this determination may require negotiation before the Helper is convinced that a subtask is done. In Figure 5-3, with no shared visual space, the Worker explicitly declared that he had completed the instruction (line 13). Note here that the linguistic evidence is more ambiguous than the visual

information. For example, the first “I got it” on line 9 could indicate that the Worker believed that he had understood the Helper or that he literally had obtained the piece. The Helper continues to describe the piece, until the Worker follows up again and says, “Yeah, I got it” on line 11. Only at this point does the Helper describe where to place the piece.

These excerpts provide qualitative demonstrations of visual information being used to support both conversational grounding and situation awareness. Either or both of these mechanisms could account for the performance benefits found in this experiment and in prior studies. It is the goal of the remaining experiments in this chapter to help understand the effect of each of these coordination mechanisms in a more controlled fashion.

5.4 Study 5: Rotation study

Study 4 suggested that visual information could potentially serve two separable roles in the collaborative task. First, visual information supports situation awareness, allowing Helpers to monitor and determine the state of the task and to instruct and intervene at appropriate and useful times during the process. Second, visual information supports conversational grounding, helping speakers to construct efficient referring expressions that a partner is likely to understand and then monitor whether it was understood. Study 5 was designed to differentiate the use of visual information for situation awareness and grounding by manipulating the display so that the Helper and Worker no longer saw the visual information from the same perspective (see Figure 5-4).

In order for the visual information to be useful in providing support for grounding, the Helper and Worker must have similar views of the task and environment so that they can use the same language to describe it. In the puzzle task the support for grounding occurs at two major task levels. The first is in the initial reference to the puzzle piece under consideration (e.g., “get the red piece with a white cross in the upper left”) and the second is in describing the spatial positioning of the selected piece in the overall environment (e.g., “place it above the last one”). These relative spatial references may use the speaker, the listener, or some object as the frame of reference. This type of referent is easier, however, if the speaker and listener share a common perspective. While research has not definitively established whether there is a preferred or default reference frame or whether some reference frames are easier than others in group settings (Levelt, 1982, 1989; Miller & Johnson-Laird, 1976; cf. Schober, 1993), it has been established that shifts between frames of reference can harm group communication and performance (Schober, 1995).

In this study, the Helper's display and target area were rotated so they were different from the spatial orientation the Worker saw. After the rotation the Helpers and Workers no longer had a common reference point from which to describe object locations and discuss spatial features of the objects. The rotated views forced the pairs to negotiate their shared spatial perspective and shift their reference frame (Schober, 1993, 1995). Whether they are using a speaker-centric reference frame or an object-centric frame, rotating the Helper's view of the work area will cause difficulties for the Helper and Worker to agree upon a description of some of the objects and their relative positions. For example, in the rotated condition, the Helper's use of a description such as "the white cross in the upper left," may no longer accurately correspond to the Worker's view. Similarly, in the rotated condition it is more difficult to use efficient speaker-centric spatial descriptions such as, "to the left."

While rotation of the Helper's view is likely to degrade the Helper and Worker's ability to ground their conversation, it should not degrade the Helper's ability to maintain situation awareness. Because we rotated the Helper's view of both the work area and the target area they are describing, she could still compare the work area to the target and assess whether the Worker has performed actions correctly or not. For example, the Helper could easily assess when the Worker had placed a piece in the correct *relative position* and could still accurately gauge when the Worker needed the next instruction. The hypotheses below summarize this reasoning:

> *H5: If visual information is primarily used for conversational grounding, then collaborative pairs will perform their task more quickly when they share spatial perspective (i.e., when their view of the work area is aligned rather than rotated).*
>
> *H6: If visual information is primarily used for situation awareness, a shared spatial perspective will have little additional influence on task performance.*

Study 5 also included an immediacy of visual feedback manipulation parallel to that of Study 4, by providing continuous, immediate updates of the visual information or updating only when the Worker sent "snapshots" of the current state. Although Study 5 was primarily designed to test the impact of perspective shifts on the value of visual information and to differentiate situation awareness and conversational grounding, it was also designed to replicate key hypotheses from Study 4. In addition to replicating the hypotheses examined in Study 4, we expected several

additional interactions between the alignment of the visual space, immediacy of the visual feedback, and linguistic complexity.

Because rotating the Helpers' view of the work area was likely to harm conversational grounding by limiting their ability to use and monitor spatial descriptions, it would be especially important for the Helper to have rapid visual feedback in order to remedy any misunderstandings the pairs might develop. Therefore, if the rotation is degrading grounding and not simply situation awareness, we should expect:

> *H7: An immediately available view of the work area will have additional performance benefits when the views between the partners are rotated.*

However, the degree of similarity between viewpoints should not impact performance equally for both levels of lexical complexity. Rotation is especially likely to interfere with pairs' ability to agree upon referring expressions for the plaid pieces compared to the primary-colored ones, since rotation requires the pairs to first establish a shared perspective from which to make reference to piece attributes. Most participants did not know the pre-existing names for the tartan plaid patterns used in this experiment (e.g., Old Sutherland tartan) and instead described the plaids by describing detailed features (e.g., "white stripe on the right"). When the plaids are rotated, some of these spatial descriptors were no longer the same for the Helper and Worker. In contrast, when describing the solid pieces, the visual information could easily be used to confirm the object referent. Therefore, we expected an interaction whereby the rotated views would impact performance more for the plaid than the solid pieces.

> *H8: An identical perspective on the work space will have additional performance benefits when the linguistic complexity of the objects increases.*

5.4.1 Method

Study 5 consists of *Immediacy of the Visual Feedback*, *Viewspace Alignment* (the spatial symmetry between the Helper and Worker views), and *Lexical Complexity* manipulations. We manipulated the immediacy of visual feedback by instituting a snapshot command, which allowed the Worker transmit a view of the current work space to the Helper.

5.4.1.1 Experimental manipulations

Immediacy of Visual Feedback (*Immediate* vs. *Snapshot*): The immediacy of visual feedback was either presented continuously or only when the Worker pressed a button on the display to transmit a static image of the current state of the work area to the Helper. The shared view was either immediately available to the Helper (*immediate condition*), or the Worker had to manually choose when to send back an image of the work area to the Helper (*snapshot condition*). Study 5 dropped the no-feedback condition that was present in many of the previous studies.

Viewspace Alignment (*Aligned* vs. *Rotated*): The Helper and Worker either had identical or misaligned views of the task. The views in the aligned condition were identical between the Helper and Worker displays, similar to previous studies. However, in the *rotated* condition, when the Worker moved a puzzle piece, the view that the Helper saw was randomly flipped in the vertical or horizontal direction and then randomly rotated 45, 90 or 135º. The target puzzle the Helper saw was transformed in the same way, so that the Helper's view of the work area and target had the same orientation. For example, with a 90º rotation, when the Worker placed a puzzle piece to the right of another, the Helper might see the two pieces as aligned one on top of the other in his picture of the target puzzle and his view of the Worker's actions (see Figure 5-4). The same geometric transformation was used for all trials for a single pair of subjects.

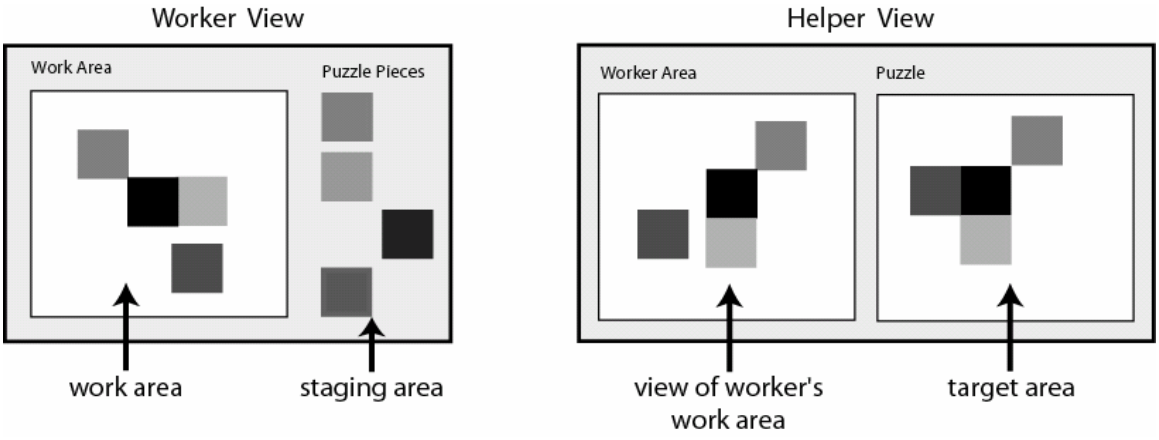


Figure 5-4. Rotated View. The Helper's view of the work area and the target are rotated 90° clockwise when presented in the Helper's view of the Worker's work area (right).

Lexical Complexity (*Primary* vs. *Plaid*): As in Study 4, the pieces were either solid primary colors (e.g., red, yellow, orange), or they were more complex visual patterns (tartan plaids).

5.4.1.2 Participants and procedure

Participants consisted of 32 pairs of Carnegie Mellon University undergraduate students. Participants received an hourly payment of $10 for their participation in the study and were randomly assigned to play the role of Helper or Worker. The Immediacy of Visual Feedback and the Viewspace Alignment were manipulated within the pairs, while the Lexical Complexity was a between-pair factor. Each pair participated in four blocks of six trials each.

5.4.1.3 Measures and statistical analysis

The analysis of performance uses time to complete a puzzle as the dependent variable. The analysis is a mixed model analysis in which Block (*1-4*), Trial (*1-6*), Field of View Alignment (*Aligned* or *Rotated*) and Immediacy of the Visual Feedback (*Immediate* or *Snapshot*) were repeated, and Lexical Complexity (*Primary* or *Plaid*) was a between-pair factor. All 2-way and 3-way interactions were included in the analysis. Because each pair participated in 24 trials (4 conditions by 6 trials per condition), observations within a pair were not independent of each other. Pairs, nested within Lexical Complexity, were modeled as a random effect.

5.4.2 Results and discussion

5.4.2.1 Task performance

Immediacy of Visual Feedback. As in Study 4 and consistent with *H1*, an immediate shared view of the work area benefited performance. As expected, the pairs were approximately 30% faster at completing the puzzles when they had an immediately available shared visual space (M = 50.45s, SE = 2.85), than when the Worker had to send back snapshots of the space (M = 72.53s, SE = 2.85), $F_{(1, 721)} = 118.80$, $p < .001$.

Linguistic Complexity. Consistent with *H2* from Study 4, Lexical Complexity significantly increased completion time. The pairs were over 35% faster in the trials where the colors were easy-to-name primary colors (M = 46.67s, SE = 3.76) than when they were more complex plaids (M = 76.0s, SE = 3.76), $F_{(1, 30)} = 31.0$, $p < .001$.

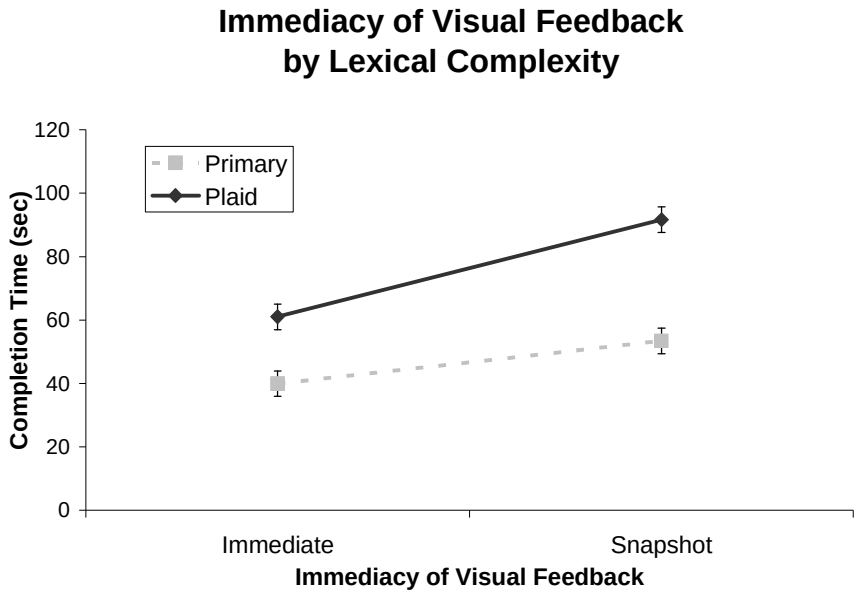


Figure 5-5. Immediacy of the visual feedback by lexical complexity (LSMeans ±1 SE).

Also consistent with Study 4 and supportive of *H3*, immediate visual feedback had the greatest benefit in the plaid condition, when the puzzle pieces were difficult to describe (see Figure 5-5), for the Immediacy of Visual Feedback × Linguistic Complexity interaction, $F_{(1, 721)} = 17.89$, $p < .001$. A detailed examination of this interaction and the effect sizes revealed that while immediate visual space improved performance when the pieces were easy-to-describe primary colors, $F_{(1, 721)} = 22.25$, $p < .001$, $d = .35$, it improved performance much more when the pieces were linguistically complex plaids, $F_{(1, 721)} = 114.44$, $p < .001$, $d = .80$.

Field of View Alignment. In support of *H5*, but inconsistent with *H6*, manipulation of the field of view Alignment had a significant impact on performance. Pairs were over 55% faster when the views were aligned ($M = 37.07$s, $SE = 2.85$) than when they were reflected and rotated ($M = 85.91$s, $SE = 2.85$), $F_{(1, 721)} = 581.44$, $p < .001$. The pairs took longer when their ability to describe the spatial arrangement of the pieces was reduced. These results suggest that the visual information was supporting conversational grounding.

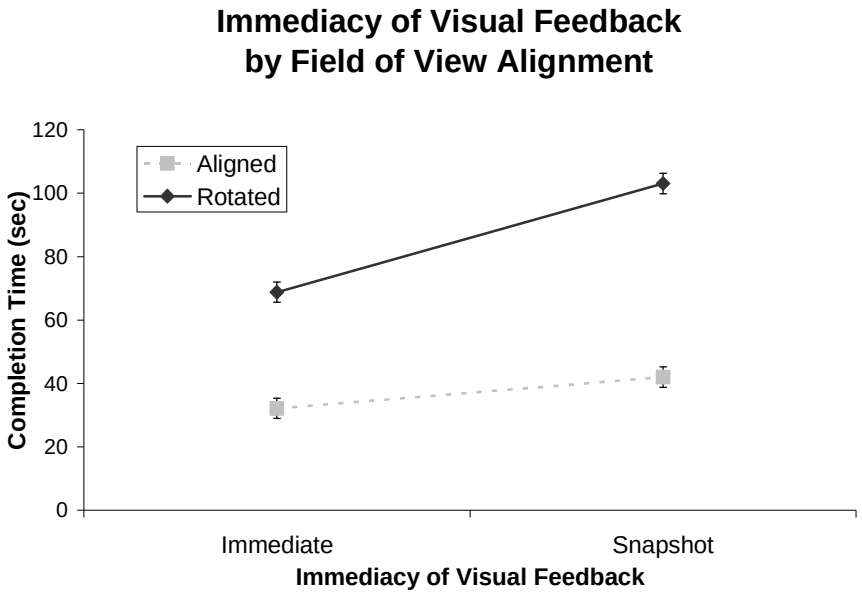


Figure 5-6. Immediacy of the visual feedback by field of view alignment (LSMeans ±1 SE).

Also consistent with the reasoning that visual information was supporting conversational grounding (*H7*), the Immediacy of Visual Feedback × Field of View Alignment interaction demonstrated that the immediate visual feedback had the greatest benefit in the rotated condition (see Figure 5-6), $F_{(1, 721)} = 36.30$, $p < .001$. A detailed examination of this interaction and effect sizes revealed that while the availability of the shared visual space improved performance when the environments were aligned, $F_{(1, 721)} = 11.88$, $p < .001$, $d = .25$, it improved performance much more when the workspaces were rotated, $F_{(1, 721)} = 143.23$, $p < .001$, $d = .89$.

Although the Field of View Alignment × Lexical Complexity interaction was in the expected direction (the difference between the plaids was greater than that of the primaries in the visual feedback condition), this difference was not significant, $F_{(1,721)} = 2.32$, $p = .13$. Therefore, we found no support for *H8*. However, it is unclear whether this was an issue of experimental power or a true lack of finding. This needs to be investigated in future research.

To summarize, the pattern of results is consistent with the interpretation that visual information improved task performance primarily by supporting conversational grounding. The visual information needs to be both temporally and spatially synchronized between people performing the task to achieve this result. If the Helpers were simply using the visual information for

situation awareness, rotations should not have made the task more difficult and the snapshot manipulation should not have accentuated the performance drop in the rotated condition. In order to explore this interpretation, we examined the transcripts at length.

5.4.2.2 Communication processes

As demonstrated by the performance data in previous chapters and in Studies 4 and 5, when a simple shared vocabulary exists to discuss the task, its objects, and the environment, there is little added benefit to having immediately available shared visual information. However, as suggested by the performance results supporting *H5*, once the views are misaligned the pairs begin to exhibit difficulties grounding spatial descriptions. Although the pairs can still easily generate and comprehend object referents, they have difficulty describing the spatial arrangement of the pieces. Simple relative spatial terms are no longer sufficient to describe the space. These problems are illustrated in

Figure 5-7, lines 3-5. Here the Helper compensates for the lack of alignment in the views by using the ambiguous term "diagonal," rather than a more concrete spatial description such as, "the upper-right corner." The Helper then uses situational awareness, by comparing the current work area to the target, to identify whether the piece has been positioned correctly. When he discovers the positioning is incorrect, he tells the Worker to try a different corner (lines 4 and 5). In this example, the Helper has no trouble producing distinct initial references to the pieces (e.g., "the red"), yet he has difficulty describing the spatial arrangement in a way that is efficient and unambiguous for the Worker. So while the rotated visual display harmed the Helper's ability to ground his descriptions, it did not seem to harm his ability to track the overall progress of the task.

Immediate, Primary and Rotated
1. **H:** put the red down somewhere
2. **H:** ok, move it down and to the right
3. **H:** now put the blue down, diagonal from the red
4. **H:** ok, try a different diagonal
5. **H:** ok, opposite diagonal
6. **H:** I want yellow adjacent to blue
7. **H:** umm, how about yellow diagonal from green
8. **H:** perfect
9. **H:** now, blue diagonal from yellow
10. **H:** perfect
11. **H:** red diagonal from yellow
12. **H:** other side
13. **H:** perfect, done

Figure 5-7. Immediate, Primary and Rotated.

Snapshot, Primary and Rotated	
1.	**H:** put the blue diagonal to the green on some side
2.	**W:** diagonal on some side
3.	**H:** yeah
4.	**W:** any side?
5.	**H:** pick a side
6.	**W:** ok
7.	**W:** *[shows]*
8.	**H:** ok, umm, put it adjacent to the green on some side
9.	**W:** *[shows]*
10.	**H:** ok, it's gonna be adjacent on a different side
11.	**W:** on the opposite side, or?
12.	**H:** it's gonna be adjacent, no, not on the opposite side
13.	**W:** ok, let's try this one
14.	**W:** *[shows]*
15.	**H:** that's right

Figure 5-8. Snapshot, Primary and Rotated.

Comprehension monitoring becomes even more difficult when visual feedback is delayed. In Figure 5-8, line 1, a different Helper uses the intentionally vague description "some side" to say where the blue piece should go. The Worker (lines 7, 9, and 14) must explicitly show the current position of the piece for the Helper to verify if it has been position correctly. Because visual feedback was not continuously available, the pairs proceeded in a lock-step fashion when trying to ground their spatial descriptions. Because the visual information was delayed through the snapshot mechanism, the Helper had to wait for the Worker to transmit an image before he could confirm or elaborate, leading to slower descriptions.

As seen previously in Figure 5-2, when puzzles contained linguistically complex plaids, continuous visual feedback helped pairs confirm which piece was being talked about. When the views were aligned, pairs could make efficient use of spatial descriptions (e.g., "…top right," "…bottom right," etc). Because the tartans were hard to describe, they often described them through efficient pronominal referents (e.g., "this," "that," "it") or other indirect references (e.g., "the block that you were just touching") rather than describing the intrinsic features of the pieces (e.g., "the one with the yellow stripe on the left"). They used visual information to facilitate referent identification and track which piece is currently at the center of attention.

Rotation and delays in visual feedback make this strategy more difficult. Figure 5-9 and Figure 5-10 provide examples of interactions from puzzles with plaid pieces and rotated views, when the visual feedback was either immediate Figure 5-9 or delayed Figure 5-10. In the beginning of

Figure 5-9, the Helper and Worker are having difficulty identifying the correct block to move, but the quick visual feedback allows the Helper to know which block the Worker is currently moving (line 3). However, the rotation prevents them from efficiently describing the relative location of the pieces (lines 6-12). The Worker tries positioning a block, which reassures the Helper he has identified the correct block (lines 6 and 7), but moves it first to the wrong place (line 11, "the other corner") and then to the correct place (line 12, "yep"). In Figure 5-10, the lack of immediate visual feedback compounds the additional grounding problems posed by having both the plaid pieces and rotation. On line 3, the Helper asks the Worker to update the display, to understand if the Worker understands her descriptions.

Immediate, Plaids and Rotated
1. **H:** that's good where that one, yeah
2. **H:** ok start, leave that one that has the four stripes right there
3. **H:** the one that you're moving right now is good
4. **H:** you want to move it to the other side
5. **H:** yeah that corner, ok
6. **H:** start with the one that's at the bottom of the screen
7. **H:** yeah
8. **W:** the top you mean
9. **H:** the top of the screen
10. **H:** and move it down, to the corner of the one that you just moved
11. **H:** the other corner
12. **H:** yep

Figure 5-9. Immediate, Plaids and Rotated.

Snapshot, Plaids and Rotated
1. **H:** you're getting close, but on my screen it has to go down
2. **H:** and that's connected to the wrong corner
3. **H:** is your, do your show
4. **W:** *[shows]*
5. **H:** no don't move that block
6. **W:** oh
7. **H:** that was in the right position
8. **W:** ok
9. **H:** ok, now we have to work on this last block. um.
10. **W:** wait, which one is the last one?
11. **H:** the one that has the cross in it

Figure 5-10. Snapshot, Plaids and Rotated.

5.5 Study 6: Field of view study

Study 6 was designed to determine whether visual information improves task performance by supporting situation awareness. In this experiment, the Helpers could see either the entire work area, a subset of it (large or small), or none of it. Compared to the full display, partial fields of view should degrade the Helpers' situation awareness (i.e., their knowledge of the overall puzzle layout), but should not interfere as much with conversational grounding (i.e., their ability to use efficient vocabulary to describe puzzle pieces in ways their partner can understand and to monitor understanding).

Field of view size. As demonstrated in Study 4, when shared visual information is provided, the pairs benefit from an increased ability to conversationally ground their piece descriptions. A small field of view that provides a view of the puzzle piece should suffice for grounding a piece description. As a result, the pairs should complete the puzzle more quickly as they go from having no shared visual information to having a small shared viewing area, in part due to the benefits of conversational grounding. This benefit should be increased when the pieces are linguistically complex plaids. However, a narrow field of view, in comparison to a wider field of view, decreases the ease with which the Helper can track the overall progress of the puzzle and the surrounding pieces. Therefore, if the pairs also get faster as they go from a small shared field of view to a larger one or from a larger one to a full shared view of the work space, the most likely explanation for these performance improvements is the additional impact on situation awareness. Figure 5-11 illustrates the various levels of the field of view.

One can use the magnitude of the performance benefit from each increment in the size of the view space to estimate the performance benefits the pairs receive from using visual information for grounding and situation awareness. We should see a greater benefit to the availability of visual information for the linguistically complex plaid puzzle pieces when going from no shared visual information to a small amount of shared visual information centered on the puzzle pieces. However, as the field of view grows larger the benefits should be equal for the plaids and primary colors, provided that the visual information is primarily supporting situation awareness. Therefore, to the extent that the pairs gain from situation awareness, the performance improvement gained with larger fields of view should not be greater for more linguistically complex tasks.

The following hypotheses summarize this reasoning for the additional conditions tested in this study:

H9: If pairs are using the visual information primarily for situation awareness, larger fields of view should improve their ability to maintain task awareness and allow them to complete the puzzles more quickly.

H10: If pairs are using the visual information primarily for situation awareness, the benefit of a larger field of view should not interact with the linguistic complexity of task.

However, if the pairs are using the visual information primarily for conversational grounding, we would alternatively expect:

H11: If pairs are using the visual information primarily for conversational grounding, larger fields of view should not improve their ability to identify objects nor allow them to complete the puzzles more quickly.

Field of View Control. When designing a system that gives its users only a narrow field of view of the work area, designers must decide who controls the view. For example, a video-mediated communication system using a camera with a small field of view could automatically track the actions of the conference attendees, allow attendees at the local site to position the camera, or allow the remote attendees to control the view. Commercial and experimental video conferencing systems have tried each of these alternatives (e.g., see Wang & Chu, 1997) and these choices are likely to have implications for both situation awareness and conversational grounding.

This study also examined three ways of controlling the Helper's field of view—Automatic, Worker-Controlled and Helper-Controlled. With Automatic view control, the field of view was centered on the Worker's mouse pointer. In this case, when the Worker grabbed a piece it was guaranteed to be in the Helper's view. Automatic view control should facilitate conversational grounding, because the Helper could always get feedback on the piece that the Worker was manipulating. However, automatic control should interfere with situation awareness, because it requires the Worker to scan the full work area with her cursor in order for the Helper to see the current state of the puzzle. The other two conditions featured manual control. In the Worker-controlled condition, the Worker used the mouse to grab an outlined window frame, indicating the area of shared view, and then either manually positioned the frame within the work area or moved the pieces over the frame to "show" them to their partner (see Figure 5-12). Worker

control should harm situation awareness for the same reason that it is harmed by automatic control. This control technique requires explicit Worker action for the Helper to see the current state of the puzzle layout. In addition, Worker control should harm conversational grounding, because the Helper does not receive feedback on the success of his utterances until the Worker has explicitly shown him what piece she was working on. This interference with feedback should be especially problematic if the pieces being described are linguistically complex. In the Helper-controlled condition, Helpers controlled the window with their cursor. This allowed Helpers to refresh their awareness of the puzzle layout at their own pace, by moving the window around the work area. However, Helper control could make grounding difficult, because the Helper and Worker might be looking at different objects or work areas. The following hypotheses summarize this reasoning:

> *H12: If the pairs are using the visual information primarily for situation awareness, they will perform their task most quickly when the Helper manually controls the view of the work area, followed by an automated view, and least quickly when the Worker needs to manually control the field of view.*

> *H13: If the pairs are using the visual information primarily for conversational grounding, they will perform their task most quickly with an automated view, followed by when the Helper manually controls the view of the work area, and least quickly when the Worker needs to manually control the field of view.*

5.5.1 Method

Study 6 manipulated the proportion of the Worker's work area viewed by the Helper (Field of view size), which partner controlled the view when only a partial field of view was available (Field of view control), and the adequacy of lexical tokens to describe the puzzle pieces (Lexical complexity).

5.5.1.1 Independent variables

Lexical Complexity (*Primary* vs. *Plaid*): The same pieces were used as in the prior two studies. The colors of the pieces were either lexically simple, easy to describe primary colors (e.g., red, yellow, orange, etc.), or they were more linguistically complex tartan plaids.

Field of View Size (*Full* vs. *Large* vs. *Small* vs. *None*). The Helper could either see the *Full* area, a *Large* area (equivalent to the size of four puzzle pieces), a *Small* area (equivalent to the area of a single puzzle piece), or nothing (*None*). Figure 5-11 shows the corresponding levels. For the small and large levels, the partner that controlled the view of the work area was varied.

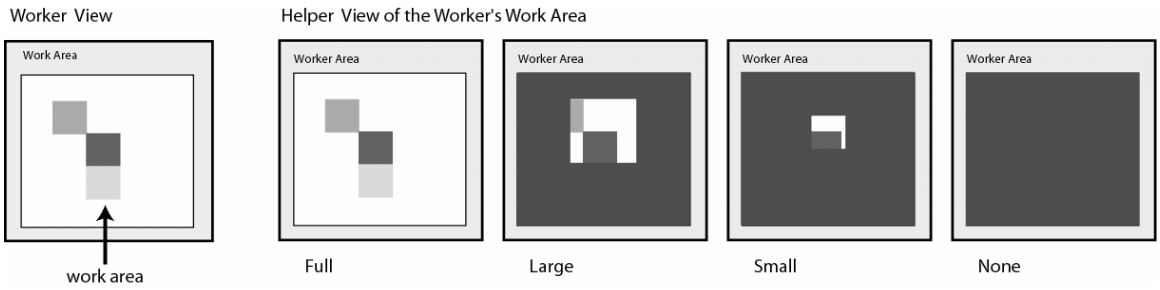


Figure 5-11. Field of View. Given the Worker's view on the left, the four Helper views on the right demonstrate the corresponding view onto the work area (Full, Large, Small and None).

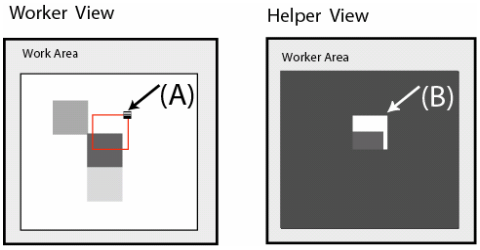


Figure 5-12. Field of View Control in the Manual Worker condition. In this condition the Worker had to manually select the shared view indicator by clicking on its corner as shown in (A) and position it within the work area, while (B) presents the corresponding Helper view.

Field of View Control (*Automatic* vs. *Manual Helper* vs. *Manual Worker*). There were three types of view control available when the Helper saw only a partial field of view (i.e., in the Large and Small view conditions). In the *Automatic* condition, the sub-view automatically followed the Worker's cursor when it was in the work area. In the *Manual Helper* control condition, the Helper controlled where they wanted to look by moving their cursor to the appropriate space. In the *Manual Worker* control condition, the Worker had to position the view over the work area (as shown in Figure 5-12).

5.5.1.2 Participants and procedure

Participants consisted of 24 pairs of Carnegie Mellon University undergraduate students who were randomly assigned to play the role of Helper or Worker. They received variable payment of $15.00 to $20.00 based on their performance. Field of View Control was manipulated between pairs, and Field of View Size and Lexical Complexity were manipulated within pairs. Each pair participated in eight blocks of four trials (32 puzzles) in an hour and a half session.

5.5.1.3 Statistical analysis

The primary performance measure was the time to complete a puzzle. The analyses also examined the type of errors made. In an error of identification, the final puzzle solution contained some wrong puzzle pieces. In an error of position, the selected pieces were correct, but they were positioned in an incorrect final position.

The analysis was performed in two stages. The first stage tested the influence of the Field of View Size and Linguistic Complexity, using a repeated measures analysis of variance in which Block (*1-8*), Trial (*1-4*), Field of View size (*None*, *Small*, *Large*, *Full*), and Lexical Complexity (*Primary* or *Plaid*) were repeated within-pair factors. Pairs were modeled as a random effect.

Field of View Control was a between-subjects factor. Because the No Visual Feedback and Full Visual Feedback conditions did not require control of the field of view, Field of View Control was only manipulated in the Small and Large Field of View conditions. Analyses examining the impact of Field of View Control used a subset of the data in which Block (*1-8*), Trial (*1-4*), Field of View (*Small* or *Large*), and Lexical Complexity (*Primary* or *Plaid*) were repeated within-pair factors, and Field of View Control (*Auto*, *Manual Worker*, or *Manual Helper*) was a between-pair factor. Pairs, nested within Field of View Control condition, were modeled as a random effect.

5.5.2 Results and discussion

5.5.2.1 Task performance

Field of View Size. As the proportion of viewable space increased, there was a strong decrease in the time it took the pairs to complete the task, (for the linear contrast, $F_{(1,707)} = 340.11$, $p < .001$). In addition, all pairwise contrasts between adjacent levels (Full = 54.5s (4.5), Large = 59.1s (4.4), Small = 67.6s (4.4), and None = 92.4s (4.5)) were significant at $p < .05$. The 27% performance

improvement in going from the no visual space condition to the small view condition was substantially larger than the 12.6% improvement going from the small to large field of view or the 7.8% improvement in going from the large to the full field of view. This pattern suggests that the use of visual information for conversational grounding may have a greater impact than its use for situation awareness. However, the fact that performance improved when going from the small to large and the large to the full field of view suggests that situation awareness improved task performance as well. These results are consistent with *H9* that the visual information was useful for supporting situation awareness. They are not consistent with *H11*, which proposes the field of view would have little impact on performance if conversational grounding was the only mechanism at play.

Lexical Complexity. As in the first two studies, the manipulation of Lexical Complexity had a significant impact on completion time. The pairs were approximately 40% faster in the trials where the colors were primary colors (M = 50.91s, SE = 4.33) than when they were plaids (M = 85.92s, SE = 4.32), $F_{(1, 707)} = 593.14$, $p < .001$.

The Field of View Size × Lexical Complexity interaction shows that the benefit received by a larger view space was greater when the pieces were lexically complex ($F_{(1,707)} = 16.21$, $p < .001$). Figure 5-13 shows that the differential benefit of increasing the size of the field of view for linguistically complex puzzles was greatest for small fields of view. To look at the interaction in more detail, a series of Field of View Size × Linguistic Complexity interactions was computed that contrasted adjacent pairs of sizes.

Similar to *H3* in Study 4, we should expect a linguistic complexity interaction with the contrast between the None versus Small field of view conditions. This would indicate that the availability of a small visual field of view was primarily aiding conversational grounding—the ability to agree on names for puzzle pieces. However, while this interaction was in the expected direction, it was not significant ($F_{(1,707)} = 1.19$, $p = .27$). As expected in *H10*, the interaction between linguistic complexity and the contrast between the Large versus the Full field of view conditions was not significant ($F_{(1,707)} = .001$, $p = .99$). Contrary to our expectations, however, the linguistic complexity interaction with a contrast between the Small and Large field of view conditions was significant ($F_{(1,707)} = 5.39$, $p = .02$). One interpretation of these findings is that the large field of view provides support for grounding as well as makes it easier for the Helper to maintain situation awareness.

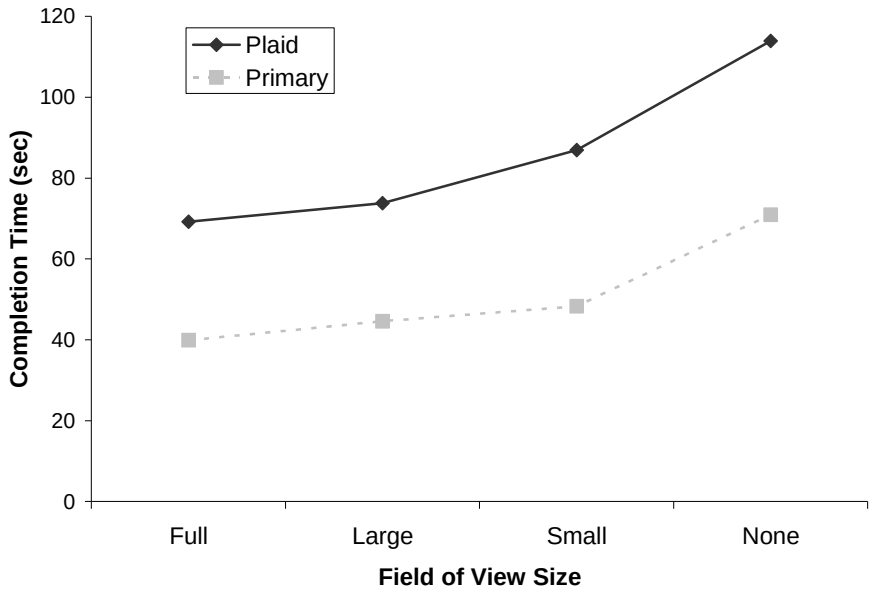


Figure 5-13. Field of View Size by Lexical Complexity on Completion Time.

Field of View Control. There were no main effects of the type of control on time to complete the puzzle (and hence no support for *H12* and *H13*). There was, however, a Field of View Control × Lexical Complexity interaction, $F_{(2,343)} = 5.58$, $p = .004$, consistent with the interpretation that the visual information is improving conversational grounding. The one degree of freedom comparison of the Automatic condition to the two Manual conditions revealed that the Lexical Complexity harmed task completion times to a lesser extent in the Automatic condition than in the two Manual conditions, $F_{(1,343)} = 10.32$, $p < .001$ (see Figure 5-14). This is in part because an automated view of what the Worker is currently working on—as a side effect of the view being yoked to their cursor—provides visual information about which piece the Worker had just selected (or not selected). When, for instance, the Helper had to manually position the work area viewer, they could miss critical information about what piece the Worker was actually working with.

The Field of View Control × Field of View Size interaction was not significant, $F_{(2,343)} = 0.73$, $p = .48$. In this case, the differences between the Field of View Control did not change dependent on the proportion of the shared work space available.

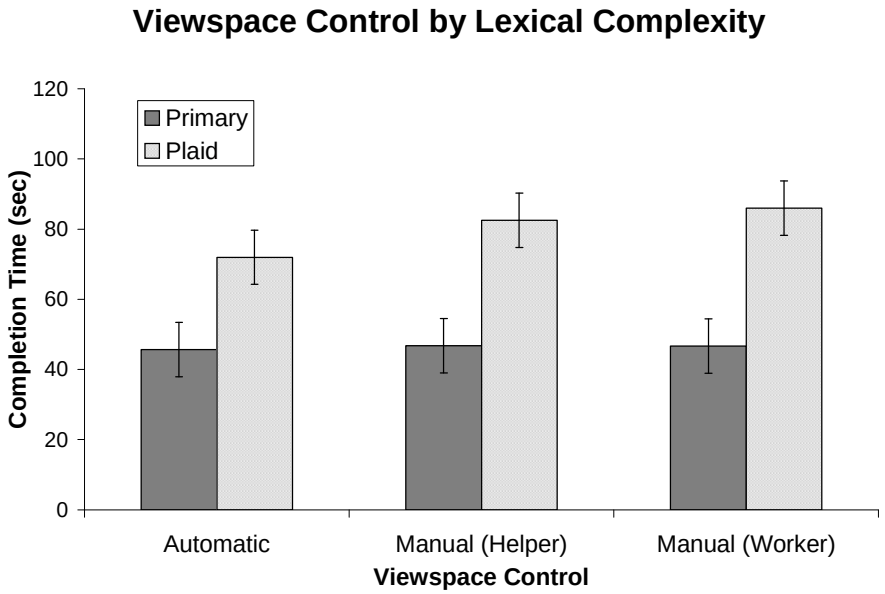


Figure 5-14. Field of View Control by Lexical Complexity (LSMeans ±1 SE).

5.5.2.2 Errors of identification and positioning

The way the visual space is used should have implications for the types of errors made when the size of the field of view is reduced. We distinguish between errors of piece identification and errors of piece positioning. Identification errors occur when the Worker moves the wrong piece. They are likely to result from failures in conversational grounding. In contrast, position errors occur when the Worker places a piece in the wrong relative position. These errors can result either from failures in grounding or failures in situation awareness such that the Helper fails to match the current puzzle layout with the target. In order to gain additional insight into the types of benefits provided by having a larger proportion of the viewing area and control over where to look in the shared space, we examined how field of view size and control influenced errors of identification and positioning.

Field of View Size. When the pairs shared a larger field of view, they committed fewer errors of both identification and positioning (for the linear contrast of field of view size on errors of identification, $F_{(1,632)} = 25.28$, $p < .001$ and on errors of position, $F_{(1, 632)} = 24.63$, $p < .001$). Thus, more visual information aided participants both in identifying pieces, an indicator of successful grounding, and in positioning them correctly, an indicator of situation awareness.

Field of View Control. There was no main effect of Field of View Control for the errors of identification or position. However, the Field of View Control × Lexical Complexity interaction for errors of position was significant, $F_{(2,304)} = 3.25$, $p = .04$. A detailed examination of this interaction reveals that there were fewer placement errors for the lexically complex pieces in the automatic control condition than for the two manual conditions, $F_{(1,304)} = 11.54$, $p < .001$. In the two manual control conditions, pairs made more errors of positioning with plaids than with primary colors, but this difference disappeared in the automatic control condition. In the manual control conditions, Helpers had difficulty verifying the correct placement of the plaids. When giving a complex instruction, such as, "So the piece with the yellow stripe at the left and the red in the upper right is to the right of the first piece," the complexity of the language needed to describe the relative positions led to more errors than when describing the placement of solid colored pieces. The manual control conditions made it difficult for the Helper to verify that the Worker had placed the pieces correctly.

5.5.2.3 Communication processes

When the pairs had a smaller view of the work area, their ability to track the state of the task and intervene at appropriate times was diminished. It is likely that this stems from an inability to gather information about the surrounding environmental context and assess the current state of the task in a timely fashion.

Small, Plaids and Automatic
1. **H:** next one we have umm, like a vertical blue stripe and then crossing it is like three red stripes wide
2. **W:** vertical, vertical blue stripe
3. **W:** ok this one
4. **W:** *[moves correct piece into workspace]*
5. **H:** yeah, that goes in, two three
6. **W:** *[positioned piece next to the wrong piece]*
7. **W:** alright
8. **H:** so the next one
9. *…[pair corrects error near the end of the trial]*

Figure 5-15. Small, Plaids and Automatic.

Figure 5-15 and Figure 5-16 contrast two examples that demonstrate how the lack of visual information about the surrounding work space minimized the amount of situation awareness available and negatively impacted performance. In both of these examples, the Helper incorrectly aligns a piece next to an existing piece in the workspace. With the larger field of view (Figure

5-16), this is immediately recognized by the Helper and remedied; however, with a smaller field of view (Figure 5-15), this error went undetected and left the pairs with inconsistent models of the state of the task.

Large, Plaids and Automatic	
1.	**H:** now there's one that's almost like that but its two
2.	**W:** this one
3.	**W:** *[moves correct piece into workspace]*
4.	**H:** yeah, it goes straight up
5.	**H:** umm, over a little bit, right above that second
6.	**W:** *[positioned piece next to the wrong piece]*
7.	**H:** or on, on the right, the red plaid one on the right
8.	**W:** *[moved piece over]*
9.	**H:** over one yep, nope, up
10.	**W:** here?
11.	**H:** up above that
12.	**W:** *[moved piece to correct position]*
13.	**H:** yeah, right there

Figure 5-16. Large, Plaids and Automatic (right).

In Figure 5-15, the Worker incorrectly positions a piece within the workspace (line 6). However, since the Helper can only see an area the size of the selected block, she does not receive enough visual feedback about the surrounding puzzle pieces to notice that it has been incorrectly positioned. As a result, she takes for granted the Worker's acknowledgement (line 7) and continues with a description of the next piece (line 8). Near the end of the trial the Worker moves his cursor over the area in the puzzle where the mistake occurs, revealing the corners of both pieces at the same time, and only then does the Helper recognize an error in the state of the solution. Contrast this with the example in Figure 5-16, in which the Worker makes a nearly identical mistake (line 6). However, the larger field of view reveals the corners of surrounding pieces, providing the Helper with visual confirmation that the task is not proceeding correctly. At this point, the Helper begins to immediately correct the positioning of the piece (line 7, 9, 11 and 13) using the immediate visual feedback to guide subsequent descriptions until the piece is correctly positioned.

5.6 General discussion

This chapter presented a series of three experiments that investigated the theoretical proposal that visual information serves as a resource for collaboration. The studies established broad support for a cooperative model of communication and demonstrated detailed support for the notion that

visual information is a critical resource for both conversational grounding and situation awareness. In addition, we examined how particular features of visual information interact with features of the task to influence both of the proposed coordination mechanisms. Table 5-2 presents an overview of the findings from the experiments and the insight that each provides towards distinguishing between the impact of visual information on situation awareness and conversational grounding. The remainder of this section addresses the theoretical and practical implications of the findings as well as the limitations and future directions of this work.

5.6.1 Theoretical implications

As illustrated in Table 5-2, the findings support *H1-H4* and the general notion that shared visual information of the workspace supports communication and coordination. These findings replicate previous work and demonstrate that collaborative pairs perform more quickly and accurately when they share a common view of a workspace (Gergle et al., 2004b, 2006; Kraut et al., 2002b). Pairs were approximately 30-40% faster when there was immediately available shared visual information as compared to when it was absent. The value of this information, however, depended on the features of the task. Its value increased when the task objects were linguistically complex and not part of the pairs' shared lexicon. Yet, even a small delay to the transmission of the visual information severely diminished its value.

Unlike previous literature, these results show that shared visual information benefits collaboration by independently supporting both situation awareness and conversational grounding. Study 5 examined *H5-H8* in order to demonstrate the benefits that shared visual information has on conversational grounding. Together the results provide evidence that conversational grounding is a central mechanism supported by the availability of shared visual information. Rotating the Helper's view degraded the ability of the Helper and Worker to describe spatial relations and ground their piece descriptions; however, this manipulation left the ability of the pairs to track the state of the task intact. We found that pairs were over 55% faster when their views were aligned.

When the shared view was rotated, the pairs could no longer easily describe the pieces using their intrinsic spatial properties, nor could they easily describe the spatial location of the pieces using efficient unambiguous referring expressions such as "to the right of," or "above," and instead had to rely on more ambiguous locative expressions such as "by". When this was the case, it was even more critical for the pairs to have immediately available continuous visual information at their disposal. Such information helped the pairs to more easily adapt to these limitations and served as

a basis to facilitate their grounding on object names and spatial descriptions. This demonstrated that a manipulation of the spatial alignment of the shared visual information had an independent impact on the ability of the pairs to perform conversational grounding, yet their ability to track the state of the task and maintain accurate mental representations of the Worker's progress towards the solution was left intact.

Study 6 examined the benefits of shared visual information on situation awareness. The results support *H9*, the notion that the shared visual information supports both conversational grounding and situation awareness. The pairs were approximately 27% faster when shifting from no visual information to a small window of shared visual information. This gain can be primarily attributed to the fact that the pairs now have visual access to the pieces which they can use to support the linguistic grounding that provides efficient description of the pieces in the puzzle. When going from a small shared field of view to a larger field of view, it is reasoned that the main benefit would be increased access to the surrounding context. In other words, the pairs received additional visual feedback that could be used to maintain a more accurate model of the task state. In going from a small to a large field of view, the pairs were 12.6% faster, while they received an additional 7.8% boost in performance when going from a large view to a full view.

When the field of view was limited (e.g., small versus large), the pairs could no longer easily track the surrounding context of the puzzles. In this case, the pairs typically received the benefit of having the visual information for conversational grounding. For example, they could still benefit from linguistic efficiencies through, for example, the use of such deictic references as "that one." However, they had more difficulty confirming the state of the task and recognizing that actions were performed correctly. As demonstrated in the qualitative descriptions, the pairs had to rely on linguistic descriptions in place of visual evidence of the surrounding context. While the evidence pointed towards an independent effect of situation awareness, it was more ambiguous for the case of situation awareness than it was for conversational grounding (we return to this problem when discussing some drawbacks to these studies).

Table 5-2. Overview of hypotheses, quantitative results and implications for situation awareness and conversational grounding.

Short description	Study 4	Study 5	Study 6	General findings	Impact on situation awareness and conversational grounding
H1: Pairs perform quicker when they have a shared view	+		+	Pairs exhibit ≈30-40% faster performance when going from no shared visual information to having shared visual information.	Ambiguous results about whether it is situation awareness, conversational grounding, or both that play a role.
H2: Pairs perform slower when the linguistic complexity of the objects increases	+	+	+	Pairs exhibit ≈30-40% faster performance when the lexical complexity of the task objects increases.	This suggests that when referential grounding is required, the pairs are slower to complete the task.
H3: A shared view area will have additional benefits when the linguistic complexity increases	+	+	**partial**	Studies 4 and 5 demonstrate added benefit to immediately available visual information when the pieces are lexically complex plaids. However, Study 6 failed to find this between no shared view and a small shared view, and only found it for comparisons between the larger views.	Study 5 demonstrates strong evidence consistent with the notion that conversational grounding is a critical mechanism supported by shared visual information. Study 6 provides partial support for the notion that situation awareness is also a critical mechanism supported by visual information.
H4: Delay in transmission will weaken the value of a shared view	+	+		Studies 4 and 5 demonstrate strong support for the hypothesis that a delay in the immediacy of the visual information (in various forms) weakens the value of the visual information.	
H5: If visual information is primarily used for conversational grounding, pairs will perform quicker when they share a spatial perspective		+		Pairs were over 55% faster when their views were aligned than when they were rotated.	Provides unambiguous evidence that conversational grounding is a central mechanism supported by shared visual information.
H6: Alternatively, if the visual information is primarily used for situation awareness, a shared spatial perspective will have little additional benefit		-		See above.	This alternative hypothesis was not supported (*see above*).
H7: An immediately available view will have additional benefit when the shared views are rotated		+		Pairs gained additional benefit from immediate visual information when the views were misaligned.	
H8: An identical viewpoint onto the work area will have additional benefit when the linguistic		**n.s.**			

complexity of the objects increases			
H9: If the pairs are using the visual information primarily for situation awareness, larger fields of view should improve task awareness and benefit performance	**+**	Pairs are ≈27% faster when going from no shared visual information to a small amount, ≈12.6% faster in going from a small view to a large view, and ≈7.8% faster when going from a large to a full view.	This evidence suggests that both situation awareness and conversational grounding play a role. It also suggests that conversational grounding has a greater impact on performance than task awareness in our configuration.
H10: If the pairs are using the visual information primarily for situation awareness, the larger field of view should not interact with the linguistic complexity of the objects	**partial**	As expected, there was no difference between the Large and Full. However, there was an interaction between the Small and Large field of view sizes.	This evidence provides partial support for the notion that situation awareness plays an independent role in performance. However, the results remain slightly ambiguous due to the significant interaction in the range between the Small and Large field of views.
H11: If the pairs are using the visual information primarily for conversational grounding, larger fields of view should not improve their ability to identify objects nor cause them to complete the puzzles more quickly	**-**	This alternative hypothesis was not supported. As described in H9, there was an impact of field of view size on task performance.	*See above.*
H12: If the pairs are using the visual information primarily for situation awareness, we should expect Helper Manual < Automatic < Worker Manual	**n.s.**		The findings did not differentiate between Hypothesis 12 and 13.
H13: If the pairs are using the visual information primarily for conversational grounding, we should expect Automatic < Helper Manual < Worker Manual	**n.s.**		The findings did not differentiate between Hypothesis 12 and 13.

Together these results demonstrate general support for a distinction between the role that shared visual information plays in supporting conversational grounding and its role in supporting situation awareness. They provide specific support for Clark and Brennan's (1991) hypothesis that different communication features change the cost of achieving common ground and extended this work by demonstrating an application of this notion to situation awareness along with evidence that these facets also interact with particular features of the task (as proposed in Kraut et al., 2002a; Kraut et al., 2003).

This deeper understanding of the theoretical role that visual information plays in collaborative environments can be used to inform the development of collaborative systems, particularly those systems that are meant to support tightly-coupled collaborative activities that involve joint physical or virtual manipulation of objects that occur simultaneously with spoken communication. The next section examines the role that these theoretical findings may play in informing the future development of collaborative systems.

5.6.2 Practical design implications

By identifying the ways in which visual information impacts collaborative behavior, we can begin to make informed design decisions regarding when and how to support visual information in collaborative applications development. This section describes some concrete examples of real-world designs that require visual space and how they may be impacted by differences in their need for visual information in order to support situation awareness, conversational grounding, or both.

When applying these findings to the development of new collaborative systems, our data demonstrate the importance of understanding the task when determining the value of providing support in the form of shared visual information. Tasks may vary on several levels. The rate of change of the objects within the environment might be quick, as in the case of a rapidly changing world found in a massively multi-player online role-playing game. In this case, delays to the visual feedback will impact people's ability to maintain updated situational models of the current environment. Conversely, the task state, objects, and environment might change at a relatively slow pace as in the case of a system that supports collaborative 3D architectural planning. For such an application, it may be more suitable to spend effort establishing tools to support conversational grounding in discussions of the visual artifact (such as remote pointers or methods

for indicating landmarks) so that an architect can easily discuss the details of the model with a client who may lack the domain knowledge to speak in professional architectural terms.

In the architecture example mentioned above, there is a disparity in knowledge between the roles of the group members. Here, the architect may have specific domain knowledge that a client lacks. In this case, conversational grounding is likely to be a critical attribute to support interaction. However, this might not always be the case. Some tasks may rely more on successful situation awareness—as is the case with Air Traffic Control systems. Here an effective domain-specific abbreviated language exists for the controllers to communicate, yet the primary task relies on quickly and efficiently establishing shared situation awareness with other controllers and knowing when to pass off control of entities between airspaces. In this case, conversational grounding is less important, yet situation awareness is crucial for success. Ensuring that shared visual displays support the formation of situation awareness by making entities and environment states highly salient is critical to the design of a successful environment.

As with most user-centered designs of collaborative systems, a major first step in the design process is to understand the details of the task, the environment, and the roles and social structures of the group members involved. Once these are known, then an understanding of how the proposed applications will impact the availability of shared visual information can be considered in the relative light of the task requirements. Understanding a collaborative group's need for particular coordination mechanisms and then understanding how these mechanisms are impacted by particular technical limitations underlies the successful implementation of systems to support tightly-coupled collaborations.

5.6.3 Limitations and future directions

Maintaining a conceptual distinction between situation awareness and conversational grounding is useful from both theoretical and practical perspectives. Doing so provides insight into how these mechanisms impact collaboration and provides knowledge that can be applied to future designs. However, while this chapter has provided evidence of the independent existence of these mechanisms, the two are extremely difficult to distinguish in many real-world tasks as well as in the laboratory. For example, the small field of view in Study 6 provides benefits for grounding by allowing Helpers to see the piece being manipulated, but it also provides some situation awareness not available in the no-view condition. Future research needs to be developed to establish a cleaner distinction between situation awareness and grounding.

Another potential drawback of the current work is its use of the stylized puzzle task. The strength of this paradigm is that it allows precise control over the characteristics of the visual space and task along with precise measurements of performance and communication. This level of control has proven useful for providing insight into the interdependencies that exist between language functions and physical actions commonly observed in collaborative physical tasks. However, a possible limitation of this paradigm is that the puzzle task oversimplifies these interdependencies because of the limited range of instructional utterances and worker actions that are possible. However, it is important to note that many more complex real world tasks, whether it be remotely instructing the repair of a transformer, jointly building a Lego house, or simply discussing a journal article with a co-author located across the globe, are comprised of the same sorts of object identification-object positioning sequences studied here. Thus, the findings regarding the relationships among base level actions and language are likely to hold even when tasks involve a much more complex range of activities. However, future research is needed to address the scalability of these findings.

5.7 Conclusion

Visual information about a partner and the shared objects that comprise a collaborative activity provides many critical cues for successful collaboration. It impacts situation awareness by providing feedback about the state of a joint task, and facilitates conversational grounding by providing a resource that pairs can use to communicate efficiently. Technologies to support remote collaboration can selectively disrupt the ability to use visual information for situation awareness and grounding, and the extent of this disruption depends in part on task characteristics such as the lexical complexity of objects. The results clarify basic principles of communication and interaction, and provide insights for the design of future collaborative technologies.

Chapter 6

The Sequential Structure of Language Use and Visual Actions[14]

While Chapters 3 through 5 provided insight into performance differences and a high-level account of the patterns of language used (e.g., the use of acknowledgements), the work presented in this section delves deeper into the *sequential communication processes* that take place when collaboration is supported with shared visual information. It represents Stage II of the dissertation and examines the proposal that a shared view of the workspace allows a pair completing a physical task to substitute *actions* for *language* in the discourse surrounding task-oriented collaboration.

This work provides some of the first quantitative demonstrations of the ways in which actions and language interact and unfold over the duration of a communication episode, and how these sequences vary according to the presence of shared visual information. It extends previous analyses of the effects of media on interpersonal communication by providing a richer understanding of the way that physical actions and language are integrated to perform joint tasks and ground communication. At the theoretical level, it extends previous analyses of the effects of media on interpersonal communication by providing a richer understanding of how physical

[14] The work presented in this chapter was originally published in Gergle, D., Kraut, R. E., & Fussell, S. R. (2004). Action as Language in a Shared Visual Space. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW 2004)*, pp. 487-496. NY: ACM Press.

actions and language use are integrated to perform joint tasks and ground communication. At a more applied level, this knowledge is used to develop new design guidelines for technology to support distributed group work.

In order to draw these conclusions, sequential analysis techniques (Bakeman & Gottman, 1997; Bakeman & Quera, 1995; Fienberg, 1978; Goodman, 1978) are used to demonstrate how shared visual information can be used in concert with or as a replacement for speech. We briefly review prior performance findings and then detail the structural similarities and changes to communication that occur when language is complimented with visible actions. The work presented in this chapter was originally reported in Gergle et al. (2004a).

6.1 Introduction

A good portion of technology development for HCI and CSCW tacitly assumes that the primary goal is to support spoken language. For a large number of tasks, however, successful interaction does not rely solely on spoken language. Rather, communicative information can be provided in the form of linguistic utterances, visual feedback, gestures, acoustic signals, or a host of other sources, all of which play an important role in successful communication. Everyday communication requires conversants to integrate these elements in an extremely rapid, flexible, real-time and cooperative fashion. Speakers generate and monitor their own activities; however, they also monitor the *language* and *actions* of their partners, and take *both* into account as they speak.

Consider a group of architects, consultants, and lay clients working together to discuss architectural plans for the design of a new corporate headquarters. Communication in the group is not merely composed of a series of individual utterances produced sequentially and presented for others to hear. Rather, the speakers and addressees take into account their local contextual environment, what one another can see, etc. Many observational studies have demonstrated this rich interplay between speech and action that takes place in collaborative interactions (Bekker *et al.*, 1995; Goodwin, 1996; Tang, 1991).

However, detailed quantitative evidence describing the temporal patterns of interaction has not yet been collected. By identifying how visual information and speech can influence and substitute for one another, we can make informed decisions about when and how to provide this visual information in new tools to support collaboration. Telemedicine applications, remote repair

systems, and collaborative design technologies are but a few of the examples of systems and tasks that can be informed by this understanding. I now present a brief background to describe the theoretical roles played by shared visual information before examining its role in the sequential structure of interaction.

6.2 Action and language in communication

When people work together to solve a problem, they approach their task through different perspectives—different roles, spatial viewpoints, and levels of background knowledge. In order to coordinate their activities, they need a common set of goals and a shared language to discuss them. As previously described, this work relies on theoretical framing using Clark's Grounding Theory and Endsley's Situation Awareness Theory to help describe the relationship between actions and language use. The following section revisits these frameworks and discusses them with a particular focus on language and action.

Assessing comprehension. One way visual information affects communication is by acting as a source of evidence for understanding. Visible workspaces can provide situational awareness (Endsley, 1995) that provides evidence about both the current state of the task and the activity levels of the members. In order for speech to be effective, it needs to occur at the right moment. Visual information provides a mechanism for preparing subsequent statements and task descriptions by providing awareness of the task in relation to its overall end goal. It can also provide information regarding the availability and current activity levels of others.

Visual information has been described as one of the strongest sources for verifying mutual knowledge (Clark & Marshall, 1981). By witnessing the actions of a conversational partner, one can more readily recognize when the partner is behaving incorrectly, when they are confused and do not understand a directive, or when they do not understand the general task (Brennan, 2005). Hesitations, lack of action, and incorrect actions are all visible indicators of a lack of understanding. Imagine a pair in which a guide is remotely instructing a traveler on how to navigate from one part of campus to another. If the guide is given access to the proper visual information and the traveler turns left when she should have turned right, they can intervene with new instructions right away. In addition, the situational awareness provided by the visual information serves as a mechanism by which the guide can plan the timing of additional utterances. Continuing with the navigation scenario, if there is a particularly tricky sequence of turns, the guide can precisely issue directives one at a time if he can see where the traveler is.

Without visual feedback, the guide must continually query the traveler and rely on her to provide an accurate description of where she is and what she has done in order to successfully guide her across campus. Thus, visual information provides situational awareness that may change both the structure (e.g., who is speaking when) and the content (e.g., what is said when) of an interaction.

Because shared visual information facilitates awareness of whether an utterance has been understood, it allows pairs to coordinate the formulation of their shared language. For example, if the guide in the previous example tells the traveler to "go kittycorner" from where she is, and the traveler simply stands there, her inaction may be interpreted to mean that "kitty-corner" is not part of their shared language. A reformulation of "go diagonally to the left" may quickly remedy the situation. With shared visual information, such a comprehension error can easily be detected. By seeing the actions of the partner, the speaker gets immediate feedback regarding whether or not the addressee understood the instruction.

Assessing task performance. Visual information also serves a role in allowing judgments of task performance to be formed. Even if the speaker were addressing a robot, with no need for grounding, it would be important to have a feedback loop to get verification that an instruction had been heard and that it had the intended effects. This loop of action and feedback is more general than language and a basic tenet of user-centered design principles.

Synchronizing messages. Conversational partners have to time their contributions to ensure orderly turn exchanges. Features of media have been demonstrated to alter how efficiently turns are exchanged. For example, visual information allows pairs to overlap signals. When pairs must rely on speech to describe their situation, talking at the same time will likely lead to confusion and incomprehensible speech. However, when a shared visual space is available, they can overlap their signals by relying on multiple modes of communication (Gergle et al., 2004c). For example, while the speaker describes the task, the addressees can demonstrate their understanding using action—in effect, parallelizing the modes of communication. Whereas with a reliance on spoken language to achieve this, addressees often have to wait for an opportunity to interject, leading to a less efficient exchange. However, simply because this can be done does not mean it is optimal. If attentional focus is not shared, then the communicative intent of the action may be missed and yield misconceptions about the degree to which information is mutually shared.

By considering these crucial aspects of collaboration and how shared visual information may support joint tasks, we can now begin a more detailed investigation of how shared visual information plays out during an interaction sequence. To do so, data from the puzzle study is used as a basis for examining how various forms of visual information play out in the unfolding of interactions over time.

6.3 Decomposing the puzzle task

The sequential nature of the puzzle task makes it ideal for investigating interrelationships between speech and visible actions. In order to successfully add a piece to the puzzle, pairs first had to identify which is the correct piece and then guide it to the correct location. This identification-placement sequence had to be repeated four times to complete the puzzle, once for each piece. The basic task structure can be summarized as follows:

Step 1. Identify the piece
Step 2. Move the piece onto the workspace
Step 3. Position the piece spatially within the larger work area
Step 4. [Repeat steps 1 to 3 for subsequent pieces]
Step 5. Jointly agree to be finished with the trial

Each of these steps can be further decomposed into what Clark and Wilkes-Gibbs (1986) have called presentation-acceptance sequences. For example, to conversationally ground *Step 1* (piece identification), the following sequence of events is required:

- The Helper generates a referring expression for a puzzle piece
- The Worker gives off evidence of understanding (or lack thereof) of the referring expression
- If understanding is demonstrated, partners agree that the piece has been identified
- If a lack of understanding is demonstrated, the Helper repairs the referring expression

In the puzzle study, each of these component subtasks can be realized via speech, action, or a combination of the two. Helpers can identify referents using verbal descriptions such as “the red piece” or by deictic expressions like “that one”. Workers can evidence understanding by giving verbal acknowledgements (e.g., “ok”), by moving the correct piece into the workspace, or through a combination of the two. If a technology provides a shared view of the workspace,

collaborators following Clark's principle of least collaborative effort (Clark & Wilkes-Gibbs, 1986) will be more likely to use visible actions to ground each component of the task, since these actions are more efficient and less ambiguous indicators or comprehension. Table 6-1 presents the type of evidence (spoken or visual) that can be used at each step of the puzzle task.

Table 6-1. Type of information (spoken or visual) that can be used at various stages of the puzzle task.

Task Sub-goals	Component sub-tasks	Immediate Shared Visual Information	No Shared Visual Information
Object Reference	Make reference to piece	Spoken	Spoken
	Verify referent	Spoken or Visual	Spoken
Object Placement	Make reference to piece	Spoken	Spoken
	Describe spatial position	Spoken	Spoken
	Verify spatial position	Spoken or Visual	Spoken

6.4 Using sequential analysis techniques to examine grounding sequences

As addressed in the prior portions of this thesis, the visual evidence provided by a given technology appears to alter the way collaborators ground their utterances during each component of the puzzle task. However, although previous analyses suggest that communicators use visual evidence to facilitate grounding (Clark & Brennan, 1991; Clark & Krych, 2004; Fussell et al., 2000; Gergle et al., 2004b), they have not used analytical techniques that can identify the precise ways that language and action are interrelated. The current study builds upon this prior work by using sequential analysis techniques to determine if there is probable sequential structure, and if so, whether it varies by the availability of a shared view space.

Examining the patterns of communication using sequential data analysis techniques reveals a deeper understanding of both the role that visible action plays in communication and how it interacts with task structure. Consider, for example, the following examples of conversational strategies for achieving the same sub-goal of positioning a piece in the puzzle:

- Helper states piece position → Worker positions the piece → Helper identifies next piece
- Helper states piece position → Worker positions the piece → Helper states correctness

- Helper states piece position → Worker states understanding → Worker positions the piece → Helper states correctness
- Helper states piece position → Worker states understanding → Worker positions the piece → Worker states correctness → Helper restates piece position → Worker restates correctness

These are all strategies for attempting to achieve the same component subtask of telling a partner where to put a piece and ensuring that it occurs. Some may be more or less efficient; this depends on the mediated form of communication available to the pairs. For example, the sequence of "Helper states piece position → Worker positions the piece → Helper states correctness" may be extremely efficient when the Helper can see what the Worker is doing. However, in the event that the pairs do not share a visual space, this strategy may be extremely ineffective, both in the errors produced and the added time it takes to repair misunderstandings. Sequential analysis allows us to examine how these sequences differ across various conditions of shared visual information.

6.5 Hypotheses

This discussion leads to a couple of general hypotheses that can be investigated by using sequential analysis techniques. In particular,

> *H1: When a shared view of the workspace is present, Helpers will use Workers' actions as evidence of comprehension. They will be more likely to follow their own statements with another statement, without waiting for a Worker's verbal response, than when a shared view of the workspace is not present.*

> *H2: When a shared view of the workspace is present, Workers will be more likely to let their actions speak for themselves as evidence of their comprehension. They will be less likely to offer verbal acknowledgements of understanding when they know the Helper can see their actions than when they know the Helper cannot see these actions.*

6.6 Method

The basic setup and apparatus for this experiment has been described in Chapters 2 and the additional analyses that form the basis of the sequential analysis use a subset of the data collected in Study 1 in Chapter 3.

6.6.1 Measures

To investigate the relationship between the shared visual information and the dialogue structure, a theoretically-derived coding scheme was developed to capture the primary purpose of each utterance and action (for similar linguistic coding schemes see Anderson *et al.*, 1991; Carletta *et al.*, 1997). Separate media streams were transcribed that captured the *utterances* and *actions* of both Helper and Worker, permitting an investigation of the circumstances under which shared visible actions could replace spoken language. Since the Worker could speak at the same time as the Helper, it took three overlapping streams to accurately capture a pairs' interaction.

The final set of codes used in this study is represented by four major categories: Helper utterances, Worker utterances, Worker actions, and jointly occurring Worker utterances and actions. These categories are detailed in Table 6-2.

Table 6-2. Utterance and behavioral action codes.

Utterance/Action Code	Description of Code
Helper Utterances	
$H\_UTT_{REFERENT}$	Helper makes reference to a specific piece (e.g., "Take the red one")
$H\_UTT_{POSITION}$	Helper describes the position of a single piece (e.g., "Put that in the upper-left")
$H\_UTT_{ACK\_BEHAVIOR}$	Helper acknowledges a behavior (e.g., "Yes, that's perfect")
$H\_UTT_{CONTEXT}$	Helper discusses contextual information about the task or process
Worker Utterances	
$W\_UTT_{REF\_OR\_POS}$	Worker makes an utterance about a referent or a positional statement (e.g., "it's black and green?")
$W\_UTT_{ACK\_BEHAVIOR}$	Worker acknowledges a behavior (e.g., "I've done it")
$W\_UTT_{ACK\_UNDERSTAND}$	Worker acknowledges understanding (e.g., back-channels such as "mmmhmm")
$W\_UTT_{CONTEXT}$	Worker discusses contextual information about the task or process
Worker Actions	
W_ACT_{MOVE}	Worker moves a piece into the workspace
W_ACT_{REMOVE}	Worker removes a piece from the workspace
$W\_ACT_{POSITION}$	Worker positions a piece within the workspace or existing puzzle
Worker Utterances + Actions	
$W\_UTT+ACT_{ACK\_UND+MOV}$	Worker acknowledges and moves a piece close in time (e.g., "mmm-hmm" [Worker moves piece into the workspace])
$W\_UTT+ACT_{ACK\_BEH+POS}$	Worker acknowledges a behavior and positions a piece close in time (e.g., [Worker positions piece next to center square] "Done")

The original data set contained onset and offset times that captured the entire duration of the utterance or action in multi-stream event format[15]. This initial arrangement allowed a look at the data using a variety of temporal windows.

Figure 6-1 visualizes a small portion of the coded behaviors from the original data when shared visual information was available. There are several points of interest in this small excerpt. In the top graph, the first two bars represent a typical presentation-acceptance pair. The Helper begins at 3:16 by issuing a positional statement that tells the Worker where to put the puzzle piece. The Helper accepts this proposal by positioning the piece in the workspace. Notice that the Worker does not comment on whether or not she understood the position, nor does she linguistically assess the quality of the move. Rather, the visual availability of her actions implies her understanding. At around 3:19, the Helper treats this move as an acceptance and continues on with the next presentation of an instruction.

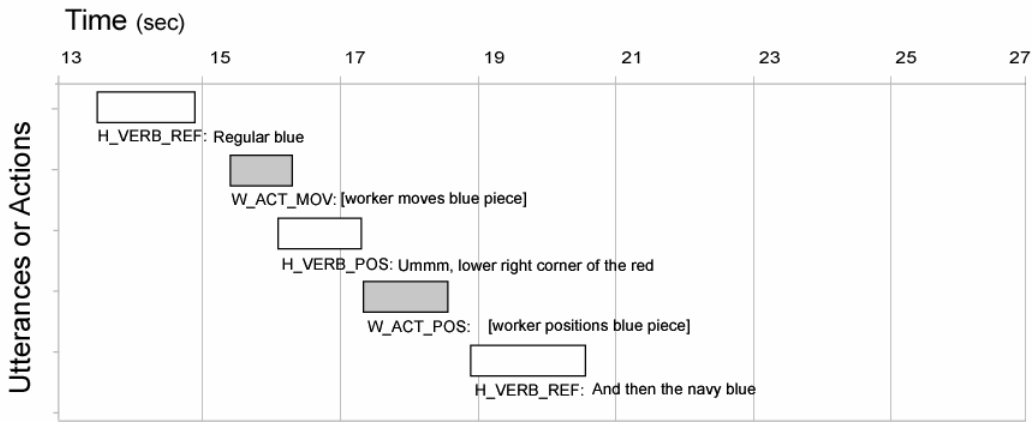


Figure 6-1. Demonstration of the coded data when shared visual information is available (white = Helper utterance; gray = Worker action; black = Worker utterance).

Contrast this with the coded behaviors presented in Figure 6-2. A quick glance reveals the dark black segments which are Worker verbal contributions, which are absent from Figure 6-1. A

[15] This data is described in Bakeman & Quera (1995) as timed event sequential data which includes overlapping data and a temporal range with a start- and end-time for each element captured. Timed event sequence data is one of the most data conserving forms, and it can later be reduced to exclusive event states if need be.

closer look at the first three exchanges reveal a similar presentation-acceptance pair as explored in the previous example. At about 3:13, the Helper issues a statement about where to place the light green piece[16]. Notice that here the Worker positions the piece (as in the previous example), and also verbally confirms her understanding and proceeding. It was this temporal richness of exchanges that the coding scheme allowed us to capture, and the differences in sequential structures were then able to be compared statistically.

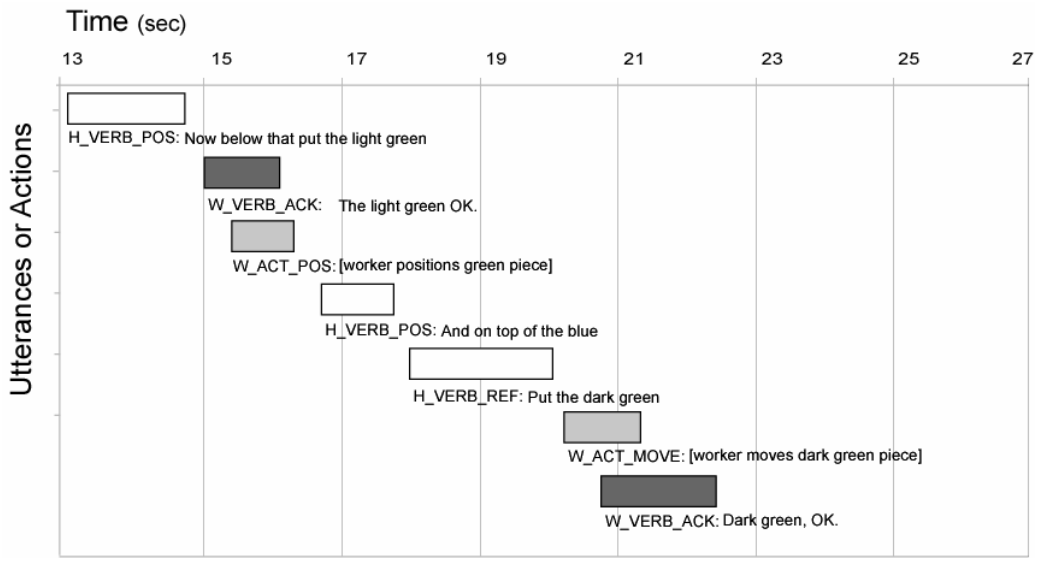


Figure 6-2. Demonstration of the coded data when shared visual information was not available (white = Helper utterance; gray = Worker action; black = Worker utterance).

Two independent coders classified a sample of utterances until they reached 90% agreement. They then each coded different transcripts, periodically coding a common transcript to ensure that the categories they used did not drift during the duration of the coding. Agreement remained high throughout.

6.6.2 Statistical analysis

The major interest in this chapter is on the impact of the availability of shared visual information on conversational structure and actions. Log-linear modeling, lag-sequential analysis, and Chi-

[16] In the previous utterance, the Helper declared that the Worker should be looking for "a light green, light minty colored piece").

square techniques were used to examine the sequential nature of the data (Bakeman & Gottman, 1997; Bakeman & Quera, 1995; Fienberg, 1978; Goodman, 1978). Using these techniques for analysis of group interactions is not a novel idea for studying group processes (e.g., see Olson *et al.*, 1994; Sanderson & Fisher, 1993; Weingart, 1997), however, it is oftentimes an under-utilized technique due to the heavy time investment required.

Log-linear modeling is a general technique for analyzing multi-way contingency tables. It is useful to assess the global nature of the sequential structure by comparing the degree to which the data are sequentially structured versus being randomly distributed. Multivariate investigations allow the exploration of how the sequential nature changes across experimental conditions. The lag-sequential method was used as a confirmatory technique to look for theoretically driven sequential patterns that occur more often than expected by chance.

After using these two techniques to determine whether or not sequential differences existed across conditions, we used theoretically-driven one degree of freedom Chi-square tests to examine particular areas of interest and determine exactly where the differences in sequence occurred.

6.7 Results

The first portion of these analyses model the sequences of data by reducing the original multi-stream timed event sequential data into individual states—or event-sequential data. Basically, each temporal encoding was reduced to a single state with the overall order determined by the onset time of the coded behavior. The original table consisted of 13 categories and 1413 cases.

Model development begins by establishing that there is sequential structure to the data. If there were no sequential order, then we would expect one category to follow another at random, dependent only on the frequency of occurrence. Cell scores would simply represent the joint probabilities of the target and given categories. This initial test can be construed as similar to an omnibus test that provides license to continue more detailed testing regarding the nature of the sequential relations.

6.7.1 References to a piece

If the process of making reference to a puzzle piece and confirming its correctness can be done through either spoken language or action (as suggested in our hypotheses), then we should expect

vast differences in the pairs' spoken communication when they had shared visual information versus when they did not. In order to explore whether this was the case, one of the most structured aspects of the task—the component subtask of identifying and making successful reference to a puzzle piece—was taken and its sequential event structure was examined.

The process of successful reference begins with the Helper issuing a statement regarding the puzzle piece to be selected. For example, "It's kinda like a mauve color" would be the starting point of such a piece reference. A 2 (**No SVS; Immediate**) × 2 (**H_UTT$_{REFERENT}$; ~H_UTT$_{REFERENT}$**) × 13 (**All Categories**) matrix was constructed to represent the sequential transitions between categories. The first dimension represents whether or not the pairs had shared visual information and is referred to as the "SVS" dimension. The second dimension is referred to as the "Given" dimension and differentiates the cases when the initial expression occurred (**H_UTT$_{REFERENT}$**) versus those when it did not (**~H_UTT$_{REFERENT}$**). The third dimension is the "Target" dimension and differentiates among the utterances and actions that the Worker or Helper could perform following the initial expression. The resulting three-dimensional matrix contains cells with the frequency of the transitions between the Target and Given events nested within the appropriate visual space condition.

An initial test of the model of independence revealed significant structure in the SVS × Given × Target matrix ($G^2_{(37)} = 564.8$, $p < 0.001$). This indicated that it was highly unlikely that the observed cell frequencies were simply the result of random transitions. In other words, there was significant dependence between the dimensions of the table. This provides statistical license to investigate the details of where this structure exists.

In order to investigate whether the sequential differences were due to whether or not the pairs had a shared visual space (i.e., whether the interaction of Given and Target categories varied across the experimental conditions), the proper model to test should include all main effects and two-way interactions. The results of such a model implied that the three-way interaction was indeed significant ($G^2_{(12)} = 33.412$, $p < 0.001$). This suggests sequential structure in the data, and that it varies across the experimental conditions.

The initial independence model (i.e., the main effects model) was investigated in order to understand specifically where the sequential differences occurred. Figure 6-3 shows the conditional probabilities and *z*-scores of the transitions between the code (**H_UTT$_{REFERENT}$**) and

several subsequent categories of interest. Note that these diagrams do not represent all of the transitions. For instructional purposes the number of nodes graphed is restricted to those that are significant and of theoretical interest.

A glance at the figure reveals where the conditional transitions vary and where large signed adjusted residuals exist (suggesting significant directional structure at greater or less than chance levels). For example, Figure 6-3 shows that following the Helper's description of a puzzle piece ($\mathbf{H\_UTT_{REFERENT}}$), the Worker moved the piece into the workspace 36% of the time when a shared visual space was available. However, when they did not have a shared visual space, this only occurred 19.6% of the time. Instead, the Helper issued an acknowledgement along with their movement 21.2% of the time. The *z*-scores in these figures serve to indicate the relative strength of the transitions while taking into account the overall frequencies of each of the categories.

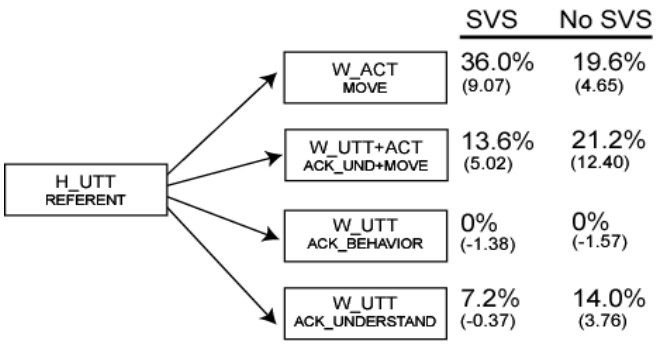


Figure 6-3. Conditional probabilities (percentages) and *z*-scores (in parenthesis) for models of piece referents.

If the pairs had been performing according to the principle of least collaborative effort, we should expect to find the transition between the Helper referent and the Worker movement more often than when there was no shared visual space. Similarly, when they had no shared space to rely on for grounding, we should expect that they would more frequently verbally acknowledge the referent or move the piece while issuing a verbal acknowledgement.

The data revealed that when the pairs had immediately available shared visual information they were much more likely to simply move the piece than to either move the piece and acknowledge that they had done so or simply acknowledge the statement (for the contrast, $\chi^2_{(1,\ N=169)} = 12.641$, $p < 0.001$). When the pairs had a shared visual space, the Worker typically responded to the

referent by simply moving the piece (as seen in the example in the left hand side of Table 6-3). However, when there was no shared visual space, the Worker typically moved the piece and provided evidence using spoken language (as seen in the right side of Table 6-3).

Table 6-3. Excerpts of pairs making object references with and without shared visual information.

Immediate shared visual information		No shared visual information	
Helper:	OK, and the orange	Helper:	Um, and then there's an orange brownish one
Worker:	[Moved correct piece]	Worker:	[Moved correct piece]
Helper:	Um, touching the right corner, right top corner of the dark blue.	Worker:	Yeah. Got it.
		Helper:	That's touching the right top of the blue one

6.7.2 Positioning a piece

As in the prior model, when the Helper gave directives on where to position the piece, the Worker could respond in several ways depending on whether or not they were taking the media into account. In order to explore whether this were the case, another commonly structured aspect of the task—the component subtask of successfully positioning a puzzle piece within the workspace—was examined for sequential structure.

This process typically begins with the Helper issuing a statement regarding where a puzzle piece should be placed. For example, "You should put it in the upper-left corner". A similar table as described above was constructed but the Given categories were replaced with the appropriate codes representing utterances about positional information (**H_UTT$_{POSITION}$**; **~H_UTT$_{POSITION}$**).

A 2 (**No SVS; Immediate**) × 2 (**H_UTT$_{POSITION}$**; **~H_UTT$_{POSITION}$**) × 13 (**All Categories**) matrix was again constructed. An initial test of the model of independence revealed significant structure in the SVS × Given × Target matrix ($G^2_{(37)} = 408.4$, $p < 0.001$). This provided statistical license to investigate the sequential structure in a more detailed fashion.

Examining whether or not this structure varied across experimental conditions again requires a test of whether the interaction of Given and Target categories varied across the experimental conditions. The results suggest that the three-way interaction was indeed significant ($G^2_{(12)} = 21.2$,

$p < 0.05$). Once again, this implies that there is sequential structure in the data, and that it varies across experimental conditions.

The main effects model was examined for significant sequential structure and differences across conditions of shared visual information, in order to understand specifically where the sequential differences occurred. Figure 6-4 shows that following the Helper's description of piece placement (**H_UTT$_{POSITION}$**), the Worker moved the piece into the workspace 36.8% of the time when an immediate shared visual space was available and only used verbal acknowledgements of any sort in 12% of the cases (combining the three other categories displayed). However, when the pairs did not have a shared visual space, they simply positioned the piece only 17.0% of the time. Instead, the Helper issued an acknowledgement along with their positioning 13.2% of the time and simply stated their understanding of where the piece should go 25.3% of the time (reserving the actual positioning of the piece an indeterminate number of turns later).

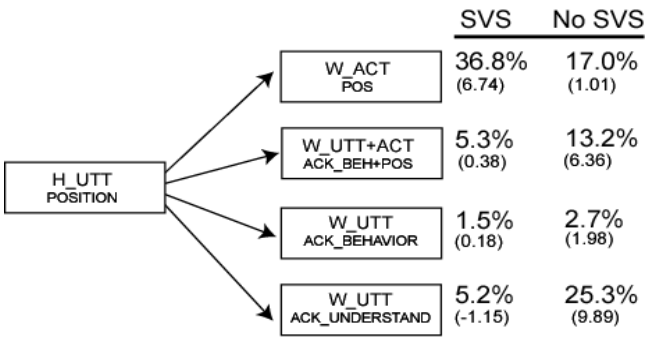


Figure 6-4. Conditional probabilities (percentages) and *z*-scores (in parenthesis) for models of piece position statements.

These differences were tested using a Chi-square analysis and it was determined that when the pairs had a shared visual space they were much more likely to simply move the piece than to either move the piece and acknowledge that they had done so or simply acknowledge the statement (for the contrast, $\chi^2_{(1, N=164)} = 34.427$, $p < 0.001$).

When the pairs had a shared visual space, the Worker typically responded to the positional information by positioning the piece (as seen in the left hand side of Table 6-4). However, when there was no shared visual space, the Worker typically positioned the piece and provided evidence of the action through spoken language (as seen in the right side of Table 6-4).

Table 6-4. Excerpts of pairs making positional references with and without shared visual information.

Immediate shared visual information		No shared visual information	
Helper:	Put it corner to corner in the lower left	Helper:	And its bottom left corner touches the top right corner of the purple one
Worker:	[Positioned piece correctly]	Worker:	[Positioned piece correctly]
Helper:	Now take a light blue	Worker:	Mmm-kay, got it
		Helper:	OK

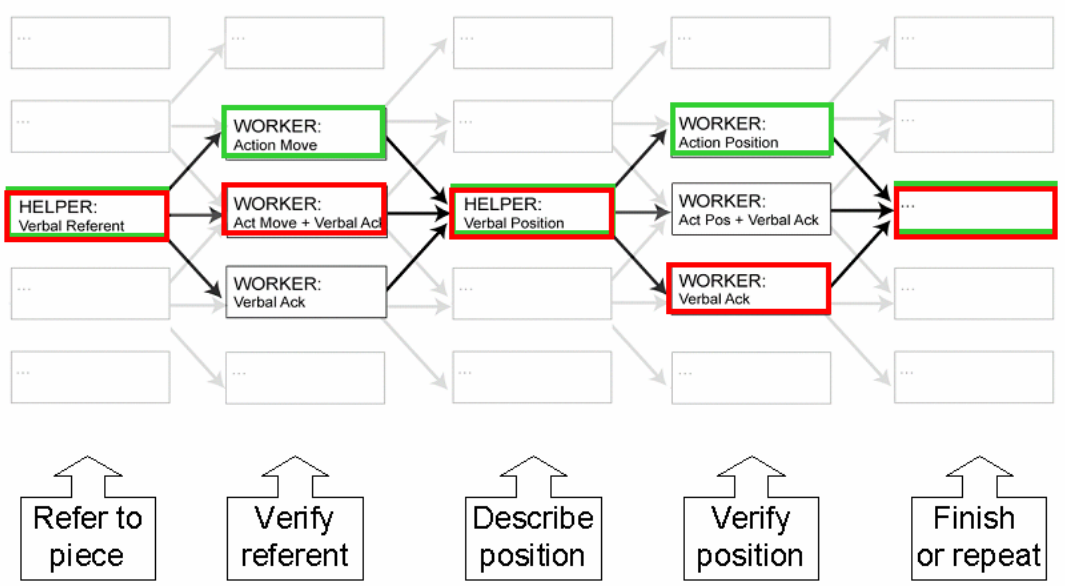


Figure 6-5. Most probable paths through the arrangement of codes starting with a piece of referent initiated by the Helper for both when the pairs had access to visual information (green) and when they did not (red).

Subsequent analyses have demonstrated that chaining these analyses together reveals interaction patterns across task components. For example, Figure 6-5 illustrates the most likely path through the coded behaviors taken by pairs when placing a piece in the workspace. Here you can see that the patterns differ between the cases where there is shared visual information (green) and those

where there is not (red). These patterns can be developed to give useful insight into the points in the task process at which different benefits are accrued. In this figure, it is during referent verification and position verification where the sequences diverge.

6.8 Discussion

These analyses provided quantitative evidence that the pairs used the visual information in two major ways. First, they used the visual information to serve as a more efficient and less ambiguous source of confirmation. For example, with a shared visual workspace the pairs were less likely to explicitly verify their actions with speech. Rather, they relied on more accurate visual information to provide the necessary communicative and coordinative cues. This was particularly evident both when the pairs made use of the visual information to support referential identity and spatial commands. Secondly, they took advantage of the visual information to recognize and remedy inconsistencies in their shared model of the state of the task. For example, pairs were able to detect errors earlier in the course of their work and remedy the situation in a timely fashion before their actions became nested and they needed to revert through several previous task states in order to fix any problems.

Overall, these results presented a much more detailed process model of the ways in which the media interacted with the pairs' behavioral patterns over the duration of the task. They clearly demonstrated that partners adapted their communication to the presence or absence of shared visual information. When a shared view of the workspace was available, the Workers were more likely to let their actions "speak" and provide evidence of their comprehension. They were less likely to present verbal acknowledgements both when attempting to select the proper puzzle piece and when positioning a relevant piece within the workspace. The sequential analyses presented in this paper demonstrated that the Workers' actions replaced a typical utterance or action + utterance sequence when they knew that the Helper could see what they were doing. Similarly, the Helpers were more likely to use the Workers' actions as evidence of understanding by simply following the actions with their next description. By using actions to help ground their utterances, pairs in the shared visual space condition were able to communicate more efficiently. This work provided a necessary step towards developing a model of interaction in the presence of shared visual information, the topic of the third and final stage of this work.

Chapter 7

Developing a Model of Referring Behavior in the Presence of Shared Visual Information[17]

The changes in patterns of language described in previous chapters provide a foundation for a model of referring behavior in the presence of shared visual information. These studies demonstrated how pairs made use of shared visual information and how it affected their ability to establish common ground and maintain situation awareness; and in doing so, it altered their use of particular types of linguistic entities. In addition, the previous chapter demonstrated how shared access to visual information impacts dialogue acts and their sequential structure. Yet, while these studies demonstrated a clear relationship between the form of visual information available and the referring expressions used, they did not explain how the relationship works. Existing theories of how visual information combines with linguistic information are underspecified, and a more detailed description of how these two forms of information combine to produce effective reference is needed. This requires deeper insight into the parameters of the linguistic mechanisms involved, including syntax and discourse level features, as well as a better understanding of the explicit features of visual information that lead to successful reference. It is the goal of the work

[17] Portions of the work presented in this chapter were originally published in Gergle, D. (2006). What's There to Talk About? A Multi-Modal Model of Referring Behavior in the Presence of Shared Visual Information. In *Proceedings of European Chapter of the Association for Computational Linguistics (EACL 2006) Conference Companion*, pp. 7-14.; and in Gergle, D. (2005). The Value of Shared Visual Space for Collaborative Physical Tasks. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2005), Extended Abstracts*, pp. 1116-1117. NY: ACM Press.

presented in the remaining chapters to produce a detailed process model that explicitly describes how visual information and linguistic information combine to account for the patterns of referring behavior observed in the puzzle studies.

As demonstrated in the previous chapters, shared visual information can have a major impact on collaborative task performance, communication efficiency, and communication processes. One mechanism continually identified as a central contributor to these benefits—both in the puzzle studies and in a host of other task-oriented studies (Barnard *et al.*, 1996; Daly-Jones et al., 1998; Fussell et al., 2003a; Gergle et al., 2004b; Karsenty, 1999; Kraut et al., 2003)—is the ability of conversational pairs to leverage shared visual information to support efficient and unambiguous object reference. This is evidenced by a pair's use of efficient referring expressions such as "this," and "that," when making reference to objects with otherwise lengthy and complex linguistic labels. As shown in Chapter 3, the pairs were more likely to replace longer noun phrase (NP) descriptions with pronouns such as "that" when shared visual information was available. This use of pronouns is demonstrated once again in Table 7-1.

Table 7-1. Use of deictic pronouns with and without shared visual information.

Immediate shared visual information		No shared visual information	
Helper:	And that over… put that on top of the red one.	Helper:	The bright blue's, the bright blue's, um, bottom left corner touches the bright red's upper right corner.

This work illustrates how a feature-based representation of shared visual information combines with linguistic cues to enable effective pronominal reference. The remaining chapters present a computational model of how people use visual information to support pronoun resolution. The current chapter presents background information, motivation and rationale for the chosen modeling approach, and an overview of the modeling framework, while the implementation details and model evaluations are presented in Chapter 8. A major goal of this work is to evaluate a language-only model, a visual-only model, and an integrated model of reference resolution when applied to a portion of the data from the PUZZLE CORPUS[18]. Results from a corpus-based

[18] The PUZZLE CORPUS refers to the data gathered using the puzzle task paradigm and contains the complete collection of data from the studies presented in Table 2-1.

analysis demonstrate that the integrated model significantly outperforms both the language-only model and the visual-only model as a model of reference resolution.

7.1 Introduction

Literature in the psycholinguistics and computational linguistics communities suggests that a number of parameters in the *linguistic context* play a central role in governing effective reference. Language-based accounts of reference typically describe the combined use of syntax, discourse metrics, and lightweight semantics as major contributors to the distributional patterns and forms of referring expressions. However, if we rely solely on language-based accounts and accept their dependence on linguistic context, a number of discrepancies appear when examining the data from the PUZZLE CORPUS. For example, the pairs often used pronouns such as "this" when the linguistic context was such that a pronoun was not licensed because the antecedent hadn't been mentioned in conversation. They also used full NPs when principles of linguistic salience suggested that a pronoun was appropriate. And they often hedged in their use of a pronoun and accompanied it with a full or partial NP (e.g., "take that [...] orange one"), even though a pronoun was linguistically licensed. While these are just brief examples, and I defer a more detailed presentation and discussion of these problems to §7.2, the key point is that by most language-based accounts of discourse these are highly atypical behaviors and existing language-based computational models of spoken discourse fail to capture many of these patterns. The following chapters argue that a major reason for this is that language-only models lack a formal way of representing the role of visual information in reference. In the following pages, I argue that language-based models of reference can be significantly improved by accounting for *visual* salience and integrating this information with existing principles of *linguistic* salience in a rule-based computational model of referring behavior.

7.1.1 Background

A number of recent studies have demonstrated that the distribution of typical linguistic patterns shifts depending on the speaker's situational context. Distributional 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 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 more syntactically simple language (Oviatt, 1997) and produce different surface realizations (Cassell & Stone, 2000) when gestures accompany their speech.

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., 2004b; Kraut et al., 2003). For example, the data in previous chapters shows that when a speaker knows their partner can see their actions but will incur a small delay before doing so, they increase their production of full NPs (Gergle et al., 2004b). Similar work by Byron and colleagues (Byron *et al.*, 2005b) demonstrates that the form of referring expression varies according to a partner's proximity to visual objects of interest.

Together this work suggests that the visual context shared by the interlocutors has a major impact on their patterns of referring behavior. Yet, as previously mentioned, a number of discourse-based models of reference rely solely 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, a handful of multi-modal models have emerged that integrate visual information into the resolution process. However, they were designed to describe human-computer dialogue and not human-human communication, and often rely on the restricted assumption of communication via a command language. While these models allow natural language as input, the expressions are typically part of a restricted domain and tied to particular functions (or commands) known by both the user and the system (e.g., "open" a "folder," or "delete" a "file"). Some of these models have incorporated a notion of gesture that is integrated with the command language. For example, the utterance "open that" accompanied by a recent mouse pointer position in the proximity of a folder icon, can resolve the pronoun "that" to the local folder icon. Thus, their approaches can be applied 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; for an interesting discussion of task limitations in these environments see Kehler *et al.*, 1998).

The work presented in the last portion of this thesis aims to develop a detailed process model of reference in an unconstrained and spontaneous dialogue environment that does not rely on a fixed grammar or vocabulary. As will be discussed in §7.1.2.1, such a model can support the future testing of theoretical claims regarding the state of a speaker's internal model of their

conversational partner, and can be used to account for the patterns of reference observed in experimental settings with a variety of contextual and visual conditions.

At a more practical level, this work evaluates the performance of several hypothesized models of reference resolution in contexts where speakers may or may not share a common visual workspace. In particular, it compares three alternative 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 salience provides sufficient information to describe patterns of referring expressions. While the experimental evidence in the prior chapters clearly demonstrates that this approach is incorrect, the vast majority of computational models of reference resolution operate under this assumption. A second hypothesis suggests that visual salience overrides any linguistic salience in governing the use of referring expressions. Finally, the third hypothesis posits that a balance of linguistic and visual salience is needed in order to account for patterns of referring expressions.

The remainder of this chapter begins with a discussion of the motivation for developing a computational model followed by a description of three models used to explore the aforementioned hypotheses. The subsequent chapter describes the details of the implementation and describes a study performed to assess the performance of the models. This study presents a hand-processed evaluation of the three models on a subset of the PUZZLE CORPUS data.

7.1.2 Motivation

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

7.1.2.1 Theoretical motivation

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.

Although these studies have suggested that shared visual information about the objects and workspace can significantly impact collaboration and communication in task-oriented interactions (Karsenty, 1999; Kraut et al., 2003; Monk & Watts, 2000; Nardi et al., 1993; Velichkovsky, 1995; Whittaker, 2003), an explicit theoretical description of how this is possible and the mechanisms by which it occurs are left unspecified. In fact, while Clark's Grounding Theory provides an excellent conceptualization of human communication and language use as a joint activity, it remains rather modest in the details it provides about the mechanisms and processes that underlie successful communication. A detailed computational description of these processes can put the theory on stronger footing, provide insight into why particular communication patterns occur, and expose implicit and possibly inadequate simplifying assumptions underlying our current theoretical understanding.

In an attempt to partially address this deficit, Pickering and Garrod (2004) introduced a mechanistic account of dialogue that details how automatic alignment of linguistic representations occurs and how this alignment influences the production and comprehension of language in group settings. However, a great deal of controversy surrounds this description and its treatment of the speakers as egocentric producers and consumers of language. This work highlights a major theoretical controversy surrounding the question of whether or not speakers model a listener's state of knowledge. A large body of literature suggests that speakers account for others during the generation and comprehension stages of communication (Hanna & Tanenhaus, 2004; Hanna *et al.*, 2003; Lockridge & Brennan, 2002; Nadig & Sedivy, 2002), while other researchers suggest that this is not always the case (Brown & Dell, 1987; Ferreira & Dell, 2000). Still other researchers suggest that a model of the addressee is a late-stage corrective mechanism that only comes into play after the original mental formulation of the utterance yet before the physical articulation (Horton & Keysar, 1996; Keysar *et al.*, 2000; Keysar *et al.*, 1998). This stands in contrast to the notion that pre-articulation formation of the utterance is done in a manner consistent with the addressee's current state of knowledge and attention (e.g., Clark & Marshall, 1981). Still, other researchers have suggested hybrid accounts that suggest the degree to which a partner is modeled is based on the salience and accessibility of the contextual information (Hanna & Tanenhaus, 2004). Together, these empirical studies propose a rich variety of theoretical rationale for patterns of referring behavior in situated environments. The modeling architecture used for the models is flexible enough to accommodate a number of these proposed processes and can aid in the investigation of a number of theoretically interesting phenomena.

In addition to these studies questioning the basic theoretical rationale behind reference and partner modeling, a number of studies have explored situated language use and the role that situational context plays in reference. The technique of eye-tracking has provided a measure of temporal precision in examining the role played by visual information and its influence on language formation and understanding (e.g., see Tanenhaus & Spivey-Knowlton, 1996; Tanenhaus *et al.*, 1995). These studies have shown that pairs performing referential communication tasks, similar to the puzzle studies, integrate visual information very early on in the formation of utterances (Eberhard *et al.*, 1995; Lockridge & Brennan, 2002; Tanenhaus et al., 1995). Even studies that suggest that integration does not occur until later in the production process, claim it typically takes place before the physical articulation of the speech (Keysar et al., 2000). In addition, experimental data suggest that speakers attempt to take into account what it is their partners can see during generation and comprehension and continually update this information with other lexical and discourse-based constraints. Work by Chambers and colleagues (Chambers *et al.*, 2002) describes how the domain of interpretation is updated in a speaker's internal model in a real-time fashion, and how this influences the comprehension of particular referring expressions (see also Allopenna *et al.*, 1998).

A computational modeling framework that can be used to explore these theoretical parameters will fulfill four major requirements. First, the framework needs to be able to account for continuous speech, not just single utterances or contributions. It should be able to account for mixed-initiative dialogues, as opposed to monologues, since communication is jointly constructed and a number of discourse-level factors play a role in reference. Second, the modeling environment needs to be able to handle a number of visual and situational contexts. As discussed in prior chapters, a wide variety of visual features and parameters play a role in the construction and comprehension of referring expressions. A complete model must be able to account for such variability. Third, flexibility in representing the state of various discourse participants needs to be available. For example, the current theoretical debate regarding the role of partner modeling could be put on stronger footing if a computational framework can explicitly capture the differences proposed by competing theorists and evaluate performance on the basis of the theories. To do this, the modeling framework needs to be flexible enough to capture not only what is linguistically and visually available to the speaker, but also what the speaker believes to be available to their partner. Finally, if a computational model is going to be useful for providing insight into collaboration and reference in task-oriented visual domains, the time-course with which reference resolution occurs needs to be flexible. A number of existing computational

models perform resolution at utterance segments. In other words, they wait until a full sentence is available to the model and then resolution is performed. However, recent empirical evidence demonstrates that the context in which the resolution occurs is updated in a real-time, word-by-word (if not finer) fashion. Thus, an accurate model will need to be an incremental one (Allopenna et al., 1998; Chambers et al., 2002; Eberhard et al., 1995) that allows for integration between the visual and linguistic information at a finer time course than a complete utterance.

7.1.2.2 Applied motivation

A computational model may also make valuable contributions to applied 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 may vary drastically. As demonstrated in earlier chapters, a variety of visual factors may impact communication and collaborative performance. 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, 1990, 2005; Gergle et al., 2004b, 2006, Under Review). The model described in this chapter could be used along with a detailed task analysis to inform the design and development of such technologies.

A final motivation for this work is to improve the performance of state-of-the-art models of communication currently used to support conversational interactions with intelligent agents (Allen *et al.*, 2005; Devault *et al.*, 2005; Gorniak & Roy, 2004). Many of these systems 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 computational 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.

To summarize, a theoretically viable modeling environment will (1) need to be able to handle continuous conversations from more than just a single speaker. It will also (2) need to be flexible in its ability to model a number of situated environments that may include linguistic as well as

visual entities. It should (3) contain a notion of group salience and the ability to model both ego-centric as well as partner-modeling approaches. And (4) it ought to have the ability to resolve conversation at a finer level of granularity than the simple sentence. Finally, in order to support practical application of the model, it should (5) be fully expressed computationally and able to augment or interact with existing dialogue managers and systems. While the initial modeling evaluations presented in Chapter 8 do not explore all of these parameters, the architectural framework is developed with a flexibility that allows it to handle this wide variety of constraints.

7.2 Reference in collaborative discourse

Reference is the act of using language (spoken or written) or gestures to enable a recipient (a reader, listener or viewer) to identify something (Yule, 1996). It is central to naturally occurring discourse and a major factor in determining the coherence of a given conversational excerpt. Natural language provides a number of ways for someone to refer to things. In the previous example presented in Table 7-1, the entity described as "the bright blue block" by the Helper may subsequently be referenced using a variety of forms such as: *it*, *this*, *that*, *the piece*, *that bright blue one*, *the brightest blue piece*, etc. Each of these referring expressions contains clues about the status of a given entity in a pair's current model of the discourse (Chafe, 1976; Gundel *et al.*, 1993; Prince, 1981). For example, it is unlikely that the Helper would use the pronoun "it" to refer to "the bright blue block" if they have since discussed several other pieces. Similarly, the Helper should use "the brightest blue piece," only if he knows that he shares visual access to three blocks that are different shades of blue with his partner.

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 (e.g., Hobbs, 1978), and those that integrate a combination of syntax, semantics and discourse structure (e.g., Grosz *et al.*, 1995; Strube, 1998; Tetreault, 2001). A number of these models are salience-based approaches where linguistic entities are ranked according to how salient they are to the speaker or listener based on their grammatical function, number of prior mentions, prosodic markers, etc.

7.2.1 Linguistic context in support of reference

In spoken dialogue, licensed referents are often introduced through the prior *linguistic context*. Of all available linguistic entities, there is often one that is thought of as the current topic of

discussion (also known as a focus) (Grosz *et al.*, 1983, 1995; Grosz & Sidner, 1986), and speakers can make reference to this entity in a variety of ways. Consider again, the following example drawn from the PUZZLE CORPUS 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:

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

In excerpt (7.1)[19], the first utterance uses the definite-NP "the dark red piece" to introduce a new discourse entity. This phrase refers to an actual puzzle piece that has a color attribute of dark red and resides in the shared workspace. Assuming the Worker has correctly heard the utterance, the Helper can now expect the entity to be a shared element that is the current focus as established by the linguistic context. This status provides license for the dark red piece to be subsequently referred to using a pronominal expression such as the "it" in the second utterance. This use of a pronoun to refer to a prior entity in the discourse is known as anaphoric reference. The referring expression is the "it", and the object being addressed is known as the referent.

7.2.2 Visual context in support of reference

In contrast to the examples presented in the previous section, during task-oriented collaborations with physical objects, the *visual context* often plays a critical role in determining which objects are salient parts of a conversation. In the following example it is not merely the linguistic context that determines the potential antecedents for a pronominal expression, but also the shared visual context, for example:

(7.2) Helper: All right, uh, take, um, the darkest orange block.
Worker: OK.

[19] A number of stylistic conventions are used to present the interaction excerpts. Spoken utterances contain a speaker role followed by a transcription of the utterance. The spoken utterance is presented using a Roman typeface. Referring expressions or pronouns contained within the utterance are presented using an underlined Roman typeface. Finally, visual actions are contained within brackets, "[" and "]" and a description of the action is provided using an *Italicized typeface*.

Worker: [*moved incorrect piece*]
Helper: Oh, that's not it.

In excerpt (7.2), both the linguistic and visual information provide entities that could be potential targets of a referential expression. In this excerpt, the first pronoun "that," specifies the "[*incorrect piece*]" that was physically moved into the shared visual context. While the second pronoun, "it," has as its antecedent the object co-specified by the definite-NP "the darkest orange block."

Another example of a common problem when applying models based exclusively on linguistic properties to the puzzle study data is in the prediction of the use of a pronoun. In the following example, the visual information creates ambiguity for the pair that results in a full NP being repeated, while a model based solely on linguistic context would claim it is not needed.

(7.3) Helper: The bluish block goes in the upper right corner.
Worker: [*Blue block positioned in the shared workspace*]
Worker: [*Green block re-positioned in the shared workspace*]
Helper: The bluish block should be all the way in the corner.

In excerpt (7.3), if the model only accounted for the spoken contributions and disregarded the two visible moves in the middle of the excerpt, the repeated use of "The bluish block" in the last utterance would seem incoherent. Rather, the use of a pronoun (e.g., "It should be all the way in the corner") would seem to be a more coherent statement. However, this example demonstrates that the visual information introduces ambiguity regarding the most salient entity for the pair, and hence, what entity is the most likely referent of a pronominal expression. The bluish block is one likely referent, since it has been mentioned and subsequently moved. However, the movement of the "[*Blue block*]" is immediately followed by the "[*Green block*]" being moved. In this case, the "[*Green block*]" is the most recently activated visual object in the shared visual workspace. This situation creates an ambiguity between the linguistically salient entity (i.e., "The bluish block") and the visually salient entity (i.e., "[*Green block*]"). For this reason the Helper, quite appropriately, repeats the full NP of "The bluish block" in the last utterance. This is done in order to circumvent any confusion that might arise from having a linguistic entity and a visual entity as possible referents of a pronominal expression.

Finally, there are a number of occasions where seemingly ambiguous referring patterns appear in the speech streams when shared visual information is available, for example:

(7.4) Helper: There is an orange-red block that obscures half of it, and it is to the left of it.

(7.5) Helper: Take that, no that, yeah it goes to the lower left.

A number of existing computational models of reference resolution will accurately resolve the pronoun in excerpt (7.1) but fail to do so in excerpts (7.2), (7.4) and (7.5). Similarly, the same models would have difficulty describing the use of the repeated NP in excerpt (7.3). Without becoming prematurely mired in the details of any particular model, most pronoun resolution models would fail for a number of reasons. In the simplest case, if the model does not account for the visible objects in the surrounding visual context, in excerpt (7.2) it will incorrectly resolve the "that" to "the darkest orange block." However, if the visual object were simply included as a potential referent in the model, it is still not clear what precedence ranking it should achieve relative to the linguistic entities and other surrounding visual entities. In typical language-only models, the ranking of available entities for a given referring expression is primarily the result of grammatical function. For example, subjects are more likely referents than direct objects, which in turn are more likely than indirect objects. A similar ranking would need to be established for visual objects, and a method for combining the rankings needs to be addressed.

Together these examples demonstrate a number of ways that *both the linguistic and visual context* play a central role in the ability of the conversational pairs to make use of efficient communication tactics such as pronominal reference. In order to successfully account for many of these patterns, two goals must be met. First, a method needs to be developed for capturing and ranking potential visual objects in the shared context. Second, a thorough understanding needs to be established for describing how linguistic and visual ranking combine to result in an ordering that accounts for the referring behaviors of humans in real-world situated environments. To do this requires a deeper understanding of how the visual elements and linguistic elements are combined in an integrated shared model of reference.

7.2.3 Toward an integrated model

The problems presented in the last section are often compounded in real-world and computer-mediated environments since the visual information can take many forms. For instance, pairs of interlocutors may have different viewpoints which could result in different objects being occluded for the speaker and 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 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 could fail to resolve references that the linguistic context determines are highly salient and the visual context does not. Therefore, in addition to language-only and visual-only models, these chapters develop an integrated model that uses both linguistic and visual salience to support reference resolution. In addition, I extend these models to the new task domain of the puzzle study which permits a more elaborate understanding of referential patterns in the presence of various forms of shared visual information. This corpus also allows a decomposition of the various features of shared visual information in order to better understand their independent effects on referring behaviors.

The remainder of this chapter describes an overview of the modeling framework and rationale for its development, a description of the puzzle corpus used to evaluate the models, and an overview of the major hypotheses that are tested in Chapter 8.

7.3 The general modeling framework

The modeling framework developed for this work aims to address the primary requirements defined in §7.1.2. It augments a rule-based model of spoken discourse in order to account for the reference patterns found in various visual conditions of the PUZZLE CORPUS. The approach adopts the ideas of Centering Theory originally developed by Grosz and colleagues (Grosz et al., 1983, 1995).

7.3.1 A centering approach

Centering Theory is a dynamic model that was developed to describe the mutual attentional state of discourse participants. It has been used to explore such linguistic concepts as the given/new distinction, common ground, discourse object salience, and the impact discourse structure has on interpretation and understanding (Brennan, 1995; Hudson *et al.*, 1986). It provides a real-time, dynamic method for tracking discourse focus and captures the notions of discourse entity salience and discourse coherence. In doing so, it provides a means to describe the referential complexity of a discourse as well as a method to describe the occurrence of particular forms of referring expressions. As a dynamic model of running discourse, it satisfies the first and third requirements outlined in §7.1.2. It provides a processing account of dialogue, as opposed to isolated words or sentences, and it captures a notion of group salience that can account for either an ego-centric account of dialogue or a view more compatible with Clark's view of partner modeling and grounding.

Another benefit of the centering model is that it characterizes the extent to which a given discourse segment is understandable based on the form of topic transitions between contributions, and the way in which speakers maintain old entities and introduce new ones in their evolving internal model of the discourse. In addition, it is considered by many to be one of the more psychologically plausible models for describing referential behavior in spoken interactions. However, in its original formulation, Centering Theory focused primarily on linguistic context, and in doing so it fails to account for many of the referential patterns found when visual context plays a role, as described in §7.2.2. However, Centering Theory's notion of linguistic salience provides an architecture that can be modified to account for visual entities as well. In this way, it satisfies the second major constraint that the modeling environment be flexible and adaptive in its ability to model more than just linguistic context.

As a theory of *discourse salience*, Centering Theory aims to describe which discourse entities are most likely to be at the center of conversational attention at any given time. In other words, it describes the entities that are the most salient to the conversational participants at any given time and therefore may be the most likely candidates for pronominalization. As a theory of *discourse coherence*, it attempts to characterize how understandable a given discourse segment is, based on the form of topic transitions between utterances and the way in which the speakers maintain old discourse entities and introduce new ones in their evolving internal model of the discourse. In this way, it provides a description of the relative ease with which a given discourse can be understood and offers testable predictions about preferred surface grammatical forms based on the psychological processing requirements that underlie comprehension.

Centering Theory achieves this with a system of rules and constraints that interact with existing semantic restrictions and world knowledge while making use of data structures to capture the local attentional focus. Together these elements govern the relationships between the discourse content and the surface forms of the utterances generated by the conversational participants. The original Brennan and colleagues (Brennan et al., 1987) algorithm for centering is provided in Appendix B.

7.3.2 The Left-Right Centering algorithm

One area where the original formulation of Centering Theory and its related algorithms (Brennan et al., 1987) are deficient is in the ability to describe reference in an online and real-time fashion similar to the experimental descriptions described in the psycholinguistics literature (Allopenna et al., 1998; Chambers et al., 2002). This poses a problem for extending the model to account for visual information, since the stream of visual information is continuous and not easily partitioned into discrete bins in the same way as utterances or sentences. The Left-Right Centering (LRC) algorithm (Tetreault, 2001, 2005) was developed to address this deficit and makes provisions for the incremental resolution of pronouns (for a theoretical discussion of these issues see Kehler, 1997). The LRC algorithm does this by maintaining a partially-ordered list of potential entities that are available at any given point during the construction of an utterance. This dynamic, real-time list of entities allows one to capture the attentional state of a discourse at a finer level of granularity than previous algorithms. The details of this algorithm and its extension to the original Brennan and colleagues definition (Brennan et al., 1987) is provided in Appendix C.

The LRC's ability to address incrementality has the fortunate side effect of providing a mechanism that allows the centering model to deal elegantly with continuous visual information. Therefore, in addition to getting the plausibility as a psychologically valid model, there is a major practical advantage to using the LRC approach in that it resolves incrementally, which is extremely useful for a model that aims to resolve pronouns with a continuous stream of constantly changing visual information. Thus, the LRC model serves to address the fourth major constraint which is to have a model that can resolve conversation at a finer level of granularity than the simple sentence.

7.3.3 Overview of the modeling architecture

In this section, I present an overview of the major components of the modeling framework, while a more detailed description of the implementation details is reserved for Chapter 8. The major components are presented in Figure 7-1 and consist of a Running Discourse History, a Transient Knowledge Base, a World Knowledge component, and a set of proposed ranking strategies for ordering the entities contained in the Transient Knowledge Base.

The Running Discourse History captures the spoken utterances, actions and objects in the shared visual environment and their corresponding timing information. These data streams are then parsed to extract the entities needed for inclusion in the dynamically ordered ranked-list of entities that comprise the Transient Knowledge Base. The major aim of the Transient Knowledge Base is to capture the salience of the various entities in a given discourse context at any given time. It includes both the visual and linguistic entities that may be the targets of future referential expressions. Before a referring expression can be successfully resolved, the World Knowledge applies selectional restrictions to elements in the Transient Knowledge Base. For example, the World Knowledge module is responsible for imposing verb projection restrictions such as the restriction that the object following the verb "move" must be a "moveable" object. The following presents a brief overview of these components and describes how they work together at a system level, while a more detailed description of the components follows.

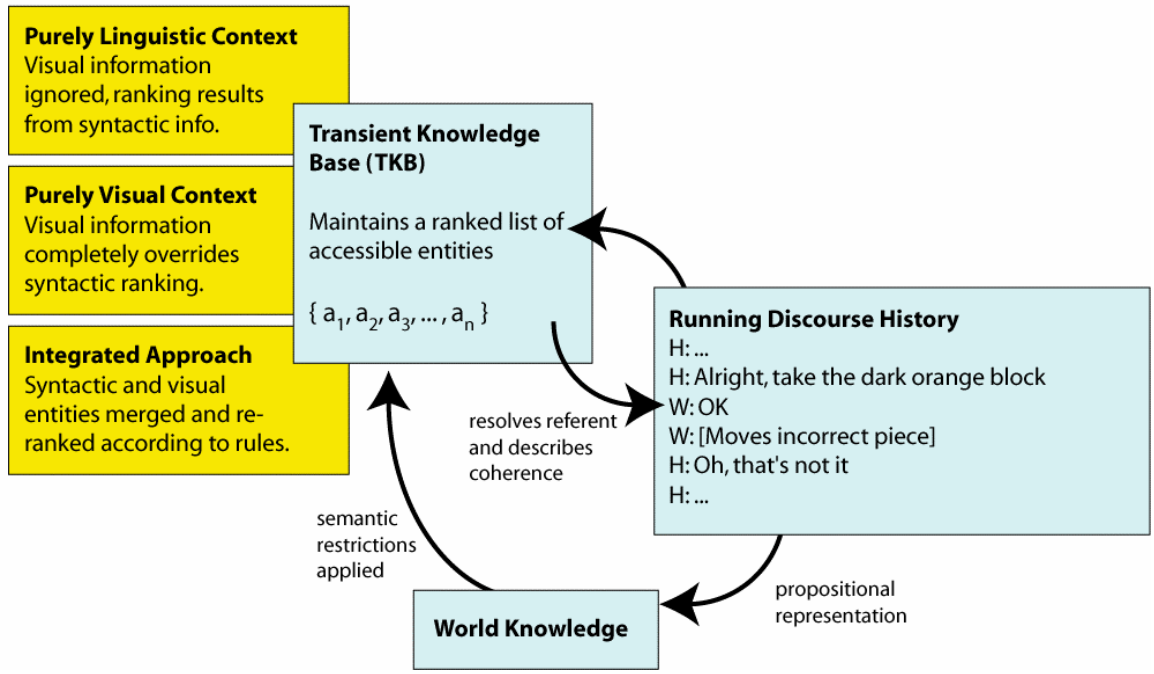


Figure 7-1. Modeling framework. Basic components (blue) and hypothesized ranking strategies (yellow).

To summarize, the basic flow of the modeling framework is as follows:

1. Parse the Running Discourse History to extract potential referential entities from both the visual and linguistic contexts.
2. Populate the Transient Knowledge Base with the linguistic and visual entities.
3. Rank the entities in the list according to a devised set of rules (e.g., grammatical function or visual salience).
4. Filter and combine the multi-modal representations of entities using a system of rules and constraints.
5. When a pronoun is encountered:
    a. Apply syntactic agreement constraints. These are constraints that ensure a match between the pronoun and the features of the referent. For example, a plural pronoun such as “they” must match in number agreement to its referent. So “the blocks” is a legal referent, while “the block” would not be. Agreement constraints for number (i.e., plurality), gender (male, female and neuter), and person (first person, second person, third person) are enforced.
    b. Apply binding constraints. These are constraints based on Binding Theory (Chomsky, 1982) that determine whether or not a pronoun needs to be bound

to a referent in its local domain. For example, a reflexive pronoun such as "himself" needs to be bound locally, as in the case of "John painted himself" versus a pronoun such as "him" that cannot be bound locally, as in the case of "John painted him."

c. Apply semantic constraints. These are restrictions such as verb argument agreement that are culled from the World Knowledge component. For example, the object following "move" needs to be a "moveable" object.
d. Select the most likely candidate from the top of the list that satisfies the syntactic agreement, binding and semantic constraints described above.

7.3.3.1 Running Discourse History

The Running Discourse History captures the utterances, actions and objects that can serve as potential referents in future utterances. From these various streams of data we can parse and extract the major units needed for inclusion in the models. The visual and linguistic information from both the Helper and Worker are captured independently and synchronized on the basis of a common timestamp. While at a first glance this approach may seem redundant, it actually allows for the capture of cases when there are asynchronies between the visual or linguistic channels of the Helper or Worker. For example, this method allows you to capture the state of the shared workspace for each individual if, for instance, the Helper is provided visual feedback that is subject to a delay as in the prior studies. Similarly, if the Helper can only see a portion of the visual space (as in the studies of limited field-of-view), independently capturing their views maintains any differences that may exist between their view and their partner's view. Details about the extraction processes are presented in §8.3.

7.3.3.2 Transient Knowledge Base

At the heart of the model is the dynamically updated ranked-list of entities that contains the constituent entities ordered by their relative salience. The highest-ranked entity in the Transient Knowledge Base is considered the most likely candidate for a subsequent referring expression. In this way, the Transient Knowledge Base intends to capture the current focus of the discourse, whether it is a recently mentioned object or a highly prominent visible object that has just been moved in the shared workspace. The Transient Knowledge Base allows the construction of a model that captures an egocentric model of the discourse state or one that captures that of the speaker as well as the addressee. In its original conception, the elements in the Transient Knowledge Base represent the joint state of the discourse and assume there are no asymmetries in the model between the participants (e.g., see the original formulations in Brennan et al., 1987;

Hobbs, 1978; Strube, 1998; Tetreault, 2001). However, when asymmetries exist, such as when there is an audio or visual delay and the partners have received separate information regarding the state of the task and discourse, the partner can be modeled in a couple of ways. First, a system of rules and Booleans can be used to capture the state of the partner's assumed knowledge. The second method is to replicate a Transient Knowledge Base for each partner in the discourse, and each one contains their beliefs about their own state as well as that of their partners. For the initial models evaluated in this chapter, I do not address the case of asymmetric viewpoints. However, the construction of the architecture enables future modeling work to address these issues.

A number of algorithms have been described that describe how to rank this list in spoken discourse (Brennan et al., 1987; Strube, 1998; Tetreault, 2001). However, little work has been done that explores the visual elements or an integrated model of visual and linguistic context, and how it influences the ranking of entities in a shared model of discourse.

7.3.3.2.1 Linguistic entities and their salience ranking

The linguistic entities used to populate the Transient Knowledge Base are extracted by parsing, chunking and tagging the utterances in the Running Discourse History. This makes it possible to identify NP boundaries as well as distinguish between pronouns and other types of nominal expressions that should be included in the Transient Knowledge Base. For a task-oriented dialogue, such as the one captured in the puzzle studies, these elements are directly recoverable from the transcribed speech. While higher-order referents and abstract entities such as propositions and events can be included in this list (Byron, 2002; Eckert & Strube, 2000), in the initial modeling such elements were not included. Each linguistic object has a number of features that determines its availability as a potential referent and its ranking within the list. Syntactic information such as recency of mention, grammatical function, and information status can be used to rank the objects. In addition, agreement constraints such as those based on gender or plurality (i.e., number) and binding constraints (e.g., *contra-indexing* constraints) (Chomsky, 1982; Hobbs, 1978) are used when resolving a referential expression.

7.3.3.2.2 Visual entities and their salience ranking

In addition to the linguistic entities, the Transient Knowledge Base can be populated with visual entities. In the PUZZLE CORPUS these elements may consist of the blocks and their associated properties. In richer visual environments the list of expected entities would unavoidably grow. For example, in a 3D virtual environment this may include other avatars, objects in the

environment, terrain features, etc. Similarly, in a GUI environment this may include icons, pointers, windows, text objects, or graphical objects.

In the PUZZLE CORPUS the visual entities (i.e., puzzle blocks) have several relevant features: whether or not the object is currently in view for Helper or Worker, the time since a piece has last come into view, the time since a piece was last visually available, whether or not the block is currently being moved, and the time since a piece was last moved. In addition, things like the total number of times the object has been active or its visual uniqueness in comparison to other available visual objects can play a role (Byron et al., 2005b; Chai et al., 2005; Huls et al., 1995; Kehler, 2000).

Obviously, there are a great number of visual features that can impact the visual salience of a particular entity in a particular environment (see Scholl, 2001 for a thorough review of the literature on object-based attention). However, one particular attribute that tends to be highly perceptually salient is object motion. For this reason, I use motion and the recency of object motion (i.e., activation) as the primary visual feature in the initial models. If visual information, as measured by the rather coarse attribute of perceptual salience, influences referring behaviors then a more complete investigation of visual salience is warranted in later research. It should be noted that the modeling framework described here does not preclude the use of more articulate notions of visual salience. In fact, future modeling plans include a more systematic approach to examining particular visual features and their influence on group salience.

7.3.3.2.3 Integrating the elements of the linguistic and visual salience rankings

Together, the linguistic entity list and the visual entity list are intended to capture all the entities that could potentially be referred to in the puzzle study data. The experiments presented in Chapter 8 examine the balance between visual and linguistic salience of the objects contained in the Transient Knowledge Base, and the hypothesized ranking strategies used to model the salience of the elements in a multi-modal, task-oriented environment.

7.3.3.3 World Knowledge

The World Knowledge component is used to capture any previously existing shared knowledge the pairs may have and also serves to enforce semantic restrictions on the elements available as referents in the Transient Knowledge Base. This component is used to eliminate or temporarily strike out particular elements contained in the Transient Knowledge Base if they do not match the

semantic restrictions imposed by the current utterance. One example of this is to ensure that the referent of "move it" is indeed a "moveable" object. The current models are developed to match the evaluations of earlier pronoun resolution evaluations which assume no world knowledge, and rely instead on syntactic agreement criteria and binding constraints (Strube & Hahn, 1999; Tetreault, 2001). However, this component is included in the framework in order to support future modeling that explores the use of lightweight notions of domain knowledge. For example, colored stripes in the plaid pieces may be referred to by location, but they are not "moveable" in the same way that the whole block locomotes.

7.3.4 The PUZZLE CORPUS

The corpus used for model development comes from the rich collection of referential communications captured in the puzzle task. This domain is much more difficult than a written corpus, in part, because of its increase in the number of disfluencies, speech repairs, repetitions, and interruptions. In contrast to a number of other spoken dialogue corpora (Chai et al., 2005; Gorniak & Roy, 2004; Huls et al., 1995; Kelleher & Genabith, 2004), these dialogues consist of unconstrained, spontaneous speech with no fixed grammar or vocabulary in a task-oriented collaboration between two human participants[20]. While this may make the corpus harder to approach from a modeling perspective, it has advantages in that any model produced will be more generally applicable to task-oriented spoken dialogues in other environs. The data collected using the puzzle paradigm contains over 15,000 spoken contributions with more than 5,500 referring expressions collected from over 180 unique pairs of participants.

7.4 Proposed ranking strategies

Three ranking strategies are examined, each of which corresponds to a hypothesized method for ranking possible referents in the Transient Knowledge Base. These hypotheses describe whether shared visual information is useful for supporting reference. The hypothesized ranking strategies are represented in yellow in Figure 7-1, and are described here:

[20] This distinction is important in that much of the prior work has examined constrained tasks such as performing object selections in GUI environments (Chai et al., 2005; Huls et al., 1995) or those that have used a constrained grammar or vocabulary (Alshawi, 1987; Huls et al., 1995) or were performed as monologues (Chai et al., 2005; Kehler et al., 1998; Kelleher & Genabith, 2004).

Purely linguistic context. One hypothesis is that the visual information is completely disregarded and the entities are salient purely on the basis of 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 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.

The following chapter presents the implementation details of the models and describes an evaluation experiment that aims to examine the details of these general hypotheses.

Chapter 8

Model Evaluation[21]

This chapter details the development and evaluation of a rule-based computational model of reference in the presence of shared visual information. Three models are developed to address the hypotheses presented in §7.4. They include a language-only model, a visual-only model, and an integrated model of reference resolution. Predictions from these models are made upon data from the PUZZLE CORPUS[22], and evaluation results demonstrate that the integrated model significantly outperforms the language-only and visual-only models of reference resolution.

The following sections present a detailed description of the models and their development, an empirical evaluation of their performance, and a reflection on the findings and future avenues for modeling. The hand-processed evaluation presented in this chapter uses automatic extraction methods to extract potential referential entities and is performed on a subset of the PUZZLE CORPUS.

---

[21] Portions of the work presented in this chapter were originally published in Gergle, D. (2006). What's There to Talk About? A Multi-Modal Model of Referring Behavior in the Presence of Shared Visual Information. In *Proceedings of European Chapter of the Association for Computational Linguistics (EACL 2006) Conference Companion*, pp. 7-14.; and in Gergle, D. (2005). The Value of Shared Visual Space for Collaborative Physical Tasks. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2005), Extended Abstracts*, pp. 1116-1117. NY: ACM Press.

[22] The PUZZLE CORPUS refers to the data gathered using the puzzle task paradigm and contains the complete collection of data from the studies presented in Table 2-1.

8.1 Introduction

The initial models presented in this chapter were developed and tested on two primary conditions from the PUZZLE CORPUS. The first was the "No Shared Visual Information" condition where the Helper could not see the Worker's workspace at all. In this condition, the pairs needed to successfully complete the task using only linguistic information. The second was the "Shared Visual Information" condition, where the Helper received immediate visual feedback about the state of the Worker's work area. In this case, the pairs could make use of both linguistic and visual information in order to successfully complete the task.

Final performance of each of the models was assessed across the two experimental conditions of the PUZZLE CORPUS. This approach is a rather novel validation technique in comparison to traditional corpus-based evaluations that often focus on model performance on a single corpus. Testing the model's performance across a range of controlled experimental conditions provides more detailed information regarding its performance. For example, this technique can determine if the language-only model performs well in cases where the pairs had to complete the task using only language yet is insufficient when applied to cases where the pairs had access to shared visual information. Table 8-1 presents an overview of the testing arrangement and the expected patterns of findings to the three models.

Table 8-1. Testing plan and expected findings.

	No Shared Visual Information	Shared Visual Information
Language Model	+	-
Visual Model	n/a	-
Integrated Model	+	+

Prior research suggests that pronoun resolution models typically perform in the range of 80-90% on written texts. However, difficulties can arise when applying the same models to human-to-human task-oriented spoken dialogues. Written text is often syntactically precise, fluid and well-structured. However, multiparty spoken dialogue often produces fragments, revised utterances, and non-grammatical speech. Prior research has demonstrated that the performance of reference resolution models falls off drastically when applied to this more challenging domain. For example, Tetreault's LRC algorithm with binding constraints only performs at 57.9% when applied to a task-oriented dialogue between two humans in which one person's task is to allocate and

distribute resources in a simulated game that uses a second participant as a "system" to aid in planning (Tetreault & Allen, 2004). Similarly, Walker (1989) found that performance of the BFP algorithm (Brennan et al., 1987) and Hobbs' algorithm (Hobbs, 1976a, 1976b, 1978) decreases when applied to human-to-human, keyboard-mediated dialogues that describe the construction of a plastic water pump (P. Cohen, 1984); the respective algorithms performed at 54.3% and 62%[23]. Therefore, if the referential patterns in the PUZZLE CORPUS are similar to those in other task-oriented multi-party spoken domains, performance should fall in the range of 55-65% for the baseline language-only algorithm.

A reasonable baseline score for applying a centering theory approach to the puzzle task can be established by assessing the language model's performance in the No Shared Visual Information condition. In this condition the pairs can only use spoken discourse to achieve the task, therefore, there should be no advantage to having a model that captures visual information or is based solely on visual information. The score in this condition (represented in the upper left quadrant of Table 8-1) provides a reasonable estimate for how well the centering approach applies to the puzzle task domain.

8.2 Corpus statistics

The data selected for this evaluation were a strategic selection from the PUZZLE CORPUS and included a randomly selected subset of trials from each of the experimental conditions. As Table 8-2 demonstrates, the data consisted of 14 dialogues from the No Shared Visual Information condition and 22 dialogues from the Shared Visual Information condition. Each of these dialogues was collected from a unique participant pair. This evaluation focused primarily on pronoun usage since it is one of the major linguistic efficiencies gained when pairs have access to shared visual information (Kraut et al., 2003).

Table 8-3 presents a breakdown of the referring expressions evaluated and their distributions within each of the experimental conditions. A rich variety of referential forms constitute the data. They include a number of personal pronouns (e.g., "he," "she," "it"), demonstrative pronouns (e.g., "this," "that," "they"), and a variety of singular, plural, and possessive pronouns as well as a

[23] In another piece of research that aimed specifically at resolving non-NPC pronouns (e.g., abstract entities), Byron (Byron, 2002) the performance to be as low as 37 to 43%.

range of other types. In addition, the corpus includes a rich collection of definite and indefinite referring expressions.

Table 8-2. Overview of the data included in the hand-processed evaluation.

Task Condition	Corpus Statistics			
	Dialogues	*Contributions*	*Words*	*Pronouns*
No Shared Visual Information	*14*	*336*	*1873*	*76*
Shared Visual Information	*22*	*327*	*2422*	*217*
Total	*36*	*663*	*4295*	*293*

Table 8-3. Distribution of the referring expressions evaluated.

Pronoun Form	Solid / No SVS	Solid / SVS	Plaid / No SVS	Plaid / SVS	Total
It / Them / They	19	7	42	76	144
This / That / These / Those	11	19	2	84	116
This / That / These / Those + NP	0	13	2	18	33
Total	30	39	46	178	293

8.3 Data pre-processing

There are several challenges in preparing a multi-modal corpus for use with models of reference, and a number of preparatory steps need to be taken in order to prepare the elements of the linguistic and visual context. Figure 8-1 provides an overview of the pre-processing used to prepare the data for use in this evaluation. This figure illustrates how data from both the spoken and visual channels are processed. The linguistic context is described on the top, the visual context on the bottom, and the right side of the figure illustrates the merging of the two sources.

8.3.1 Linguistic data

In order to work with spoken dialogue, it needs to be transcribed and segmented in a way that establishes appropriately-sized verbal contributions that capture the linguistic patterns of interest, while at the same time preserving the sequential aspects of the dialogue. Once the transcription and segmentation has been completed, the entities needed for the model are extracted and prepared for inclusion in the Transient Knowledge Base. The following describes this procedure.

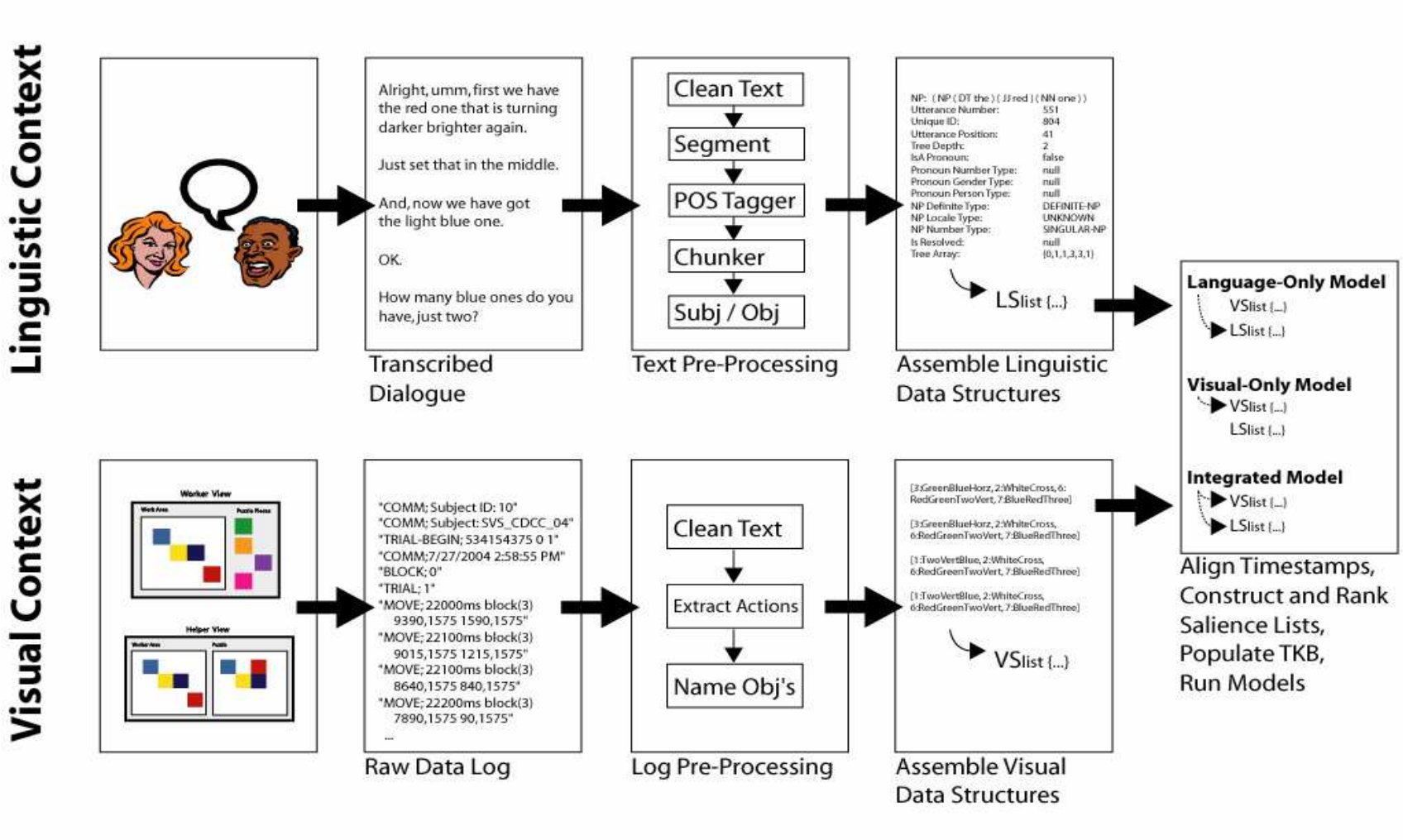


Figure 8-1. Pre-processing pipeline for linguistic information (top) and visual information (bottom).

8.3.1.1 Dialogue transcription, segmentation and alignment

A number of guidelines and heuristics were used to transcribe and segment the dialogue. They were developed based on the assumption that the pairs are participating in a spontaneous dialogue and that they are likely to follow some conventions of turn-taking. The major goal in the preparation of the dialogues was to preserve the sequential nature of the dialogue, while at the same time capturing and maintaining local phenomena so that they are not split across utterance boundaries. To do this I relied upon work by Heeman and Allen (1995) and Shriberg and colleagues (2001) that describes heuristics for segmenting unconstrained, multiparty dialogue.

Heeman and Allen's scheme is a pragmatic one developed with the intention of generating transcripts that are useful for investigating a wide variety of theories of dialogue. It aims to balance ensuring that each contribution is short enough so that it does not contain contributions from other conversational partners, with making sure that each contribution is long enough to capture local discourse phenomena. Heeman and Allen outline a number of conditions under which segmentations are appropriate. First, a segment break can be established whenever a speaker's speech stops and another speaker begins an utterance, without the original speaker trying to continue. Next, they describe that a suitable break can occur when any two of the following three criteria are met: (1) an intonational phrase boundary exists, (2) a major syntactic category boundary exists (e.g., a sentence or NP), or (3) there is a pause or breath taken. These basic guidelines were followed in preparing the transcriptions for use in the models.

Segmentation of the PUZZLE CORPUS data required three passes through the dialogue. The first pass used the notion of an utterance unit (Nakajima & Allen, 1993) or a spurt as described by Shriberg and colleagues (2001) to segment contributions based on speech pauses. This resulted in the segmentation of any speech segment separated by more than 500ms. Overlapping speech was enclosed in brackets. In this case, the utterance that was initiated first was put on the line before the interrupting or overlapping utterance, which was put on the following line.

A second pass through the transcripts refined the initial segmentation pass. During this pass, utterance breaks were added whenever there was a speaker change that resulted in a successful shift in floor control. However, backchannels (e.g., "ok") were not used to segment utterances if they were completely contained within a sentence or clause. Rather, the backchannels were added as separate utterances on the following line. In addition, unrelated clauses were hand-segmented

and put on separate lines. This made it possible for the Helper or Worker to have a number of contiguous contributions.

The third pass through the transcription was used to clean up and re-attach any spurious segments that resulted in the splitting of any major syntactic categories. This was done to maintain a notion of coherent contributions and served the secondary purpose of supporting the processing of the POS-tagger that was applied to the utterances in order to extract entities and features for the models.

8.3.1.2 POS-tagging, noun phrase extraction and subject/object tagging

To generate the appropriate features and entities, part-of-speech (POS) tagging, chunking (e.g., NP chunking), and subject/object detection was performed on the corpus. Each contribution was parsed using a memory-based shallow parser that was trained on the Penn Treebank II Wall Street Journal Corpus. The Tilburg Memory-Based Learner (TiMBL) v5.1 software package[24] (Daelemans *et al.*, 1999; Daelemans *et al.*, 2001) was used to extract the entities and tags needed for the subsequent models.

Each utterance was parsed and tagged with part-of-speech labels[25]. The chunker divided the text into syntactically related groups or clusters of words and was used to find independent (or non-overlapping) constituents. For example, the chunker clustered things like NPs and verb phrases (VPs), and made them accessible as entities for the models. Finally, the subject/object detector assigned which NP chunks it thought were the subjects or objects of particular utterances. This information was needed in the subsequent models as a method for ranking the linguistic salience of particular discourse entities. The following excerpt provides an example that will be used to demonstrate the various stages of tagging:

(7.6) *Utt*$_1$: Umm, and then there is an orange brownish one.
Utt$_2$: That is kissing the right top of the darker blue piece.
Utt$_3$: OK.

[24] The TiMBL software package is available at: http://ilk.uvt.nl/software.html .

[25] The tagging format represented throughout this chapter uses the Penn Treebank II style tags, a description of the part of speech, phrase and clause tags can be found in Appendix D. The complete annotation style manuals are available at: http://www.cis.upenn.edu/~treebank/ .

Utt$_4$: And then a yellow piece is kissing the top right of it.

Excerpt (7.6) is an example of a sequence of utterances that were segmented using the rules described in the previous section. The second utterance (*Utt*$_2$) will used to demonstrate the parsing and feature extraction. In this utterance, the Helper states "That is kissing the right top of the darker blue piece," and the following presents the tagging that is performed at each stage of the process:

Initial input:

"That is kissing the right top of the darker blue piece."

POS-tagger output:

That/DT is/VBZ kissing/VBG the/DT right/JJ top/NN
of/IN the/DT darker//JJR blue/JJ piece/NN ./.

Chunker output:

[NP That/DT NP] [VP is/VBZ kissing/VBG VP]
[NP the/DT right/JJ top/NN NP] [PNP [Prep of/IN Prep]
[NP the/DT darker//JJR blue/JJ piece/NN NP] PNP] ./.

Subject/object detector output:

[$NP_1^{Subject}$ That/DT $NP_1^{Subject}$] [VP_1 is/VBZ kissing/VBG VP_1]
[NP_1^{Object} the/DT right/JJ top/NN NP_1^{Object}]
[PNP [P of/IN P] [NP the/DT darker//JJR blue/JJ piece/NN NP] PNP] ./.

Each of these stages provides crucial syntactic information for the models. The POS-tags are used to identify pronouns of various types. The output from the chunker identifies NPs that are the essential entities required to populate the Transient Knowledge Base. These constitute both the pronouns that need to be resolved as well as the entities that make up the coreference chains and may co-specify the referents of various pronominal expressions. Finally, the subject/object detection provides additional information that is used for ranking the entities by grammatical function.

8.3.2 Visual data

In order to work with the visual information from the shared visual workspace, the actions and visible elements of the discourse were captured in a way that preserved the temporal sequence of the visual actions and allowed an abstraction of the elements in the shared visual environment. Detailed interaction logs (an example is presented in Figure 8-2)[26] were automatically generated using the puzzle study software. These logs contained information that could be pruned to develop the relevant data structures for the models described in §8.4.

Time	Log Entry Type (move / remove / update)	Piece Name (Block)	X-Pos	Y-Pos	X-Pos Distance from Solution	Y-Pos Distance from Solution
8320ms	move	red (4)	7515	3075	5250	-750
8360ms	move	red (4)	7140	3075	4875	-750
8420ms	move	red (4)	6765	3075	4500	-750
10320ms	move	dark blue (5)	9390	3075	7125	-750
12700ms	move	dark blue (5)	9015	3075	6750	-750
12740ms	move	dark blue (5)	8265	3075	6000	-750
12780ms	move	dark blue (5)	7140	3075	4875	-750
12820ms	move	dark blue (5)	6015	3075	3750	-750
12840ms	move	dark blue (5)	4890	3075	2625	-750
14480ms	move	red (4)	6015	2700	3750	-1125
14500ms	move	red (4)	5265	2325	3000	-1500
14560ms	move	red (4)	4515	1575	2250	-2250
14580ms	move	red (4)	4140	1575	1875	-2250
14600ms	move	red (4)	3765	1200	1500	-2625

Figure 8-2. Sample excerpt from puzzle study logs of the Helpers actions in the shared visual workspace.

The major elements are timing information and whether the piece was moved into, repositioned, or removed from the shared workspace. The linguistic data was then aligned with the visual data from the logs using a common timestamp. Each contribution has a start and finish time, and the visual state of the shared workspace can be resolved whenever it is needed by the model. Together, this information provided the needed material to construct the data structures used in the models.

[26] An example of the unformatted logs can be found in Appendix E.

8.4 Model overviews

As previously mentioned, the models in this evaluation are based on Centering Theory (Grosz et al., 1995; Grosz & Sidner, 1986) and the algorithms devised by Brennan and colleagues (Brennan et al., 1987) and adapted by Tetreault (2001, 2005). The language-only model is based on Tetreault's LRC model (Tetreault, 2001), the visual-only model uses a measure of visual salience to rank the objects in the visual field as possible referents, and the integrated model is a modification of the LRC model that balances the visual information along with the linguistic information to generate a ranked list of possible referents.

8.4.1 The language-only model

The LRC algorithm was chosen to serve as the base model and algorithm for the language-only model. As mentioned in the last chapter, it was chosen for a number of reasons. It is a discourse model that dynamically tracks the state of the discourse and can be extended to more than one speaker. It is flexible in its ability to account for a variety of entity types (e.g., linguistic or visual) due to its formulation as a salience-based approach (see also Poesio et al., 2004; Strube, 1998). It is extensible to the notion of group salience and partner modeling, and has mechanisms for resolving pronouns on an incremental basis, which is extremely beneficial when working with multiparty dialogues. In addition to these basic architectural advantages, the LRC algorithm performs well on task-oriented spoken dialogues and against a number of other state-of-the-art pronoun resolution models (for details see Tetreault, 2005).

LRC uses grammatical function as a central mechanism for resolving the antecedents of anaphoric references. It resolves referents by first searching within the current utterance for possible antecedents, and makes co-specification links when it finds an antecedent that adheres to syntactic agreement and binding constraints. 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. The full LRC algorithm is reproduced in Appendix C, and readers can also be referred to Tetreault's original formulation (Tetreault, 2001)[27]. In this evaluation, the output of the subject/object

[27] It should be noted, as described in Tetreault's dissertation (CITE), that the LRC algorithm is "*loosely based on Centering Theory since it only uses one construct, the CF-list*" and that it does not strictly enforce Centering Theories Rule 1 and 2, but rather approximates them. Tetreault's justification for this is that

detector was used to generate syntactic labels that would allow a given NP to be ranked in the entity list according to grammatical function. The grammatical function ranking was determined by the following precedence ranking:

$$\text{Subject} \prec \text{Direct Object} \prec \text{Indirect Object} \prec \text{Other}$$

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.

8.4.1.1 Modifications to syntactic agreement constraints under multiparty dialogues

As previously described, a number of syntactic agreement constraints are used to support pronoun resolution. These agreement criteria are typically used in pronoun resolution algorithms to restrict the type of ties that can be made between pronouns and potential antecedents. For example, the antecedent of a plural pronoun such as "they" needs to be a plural object (e.g., "the blocks"), as opposed to a singular object (e.g., "the block").

While a number of these constraints are straightforward in written discourse or monologues, a couple of adaptations need to be made when applying them to dyadic speech. The first of these modifications is for constraints based on grammatical person role. For example, "you" does not always refer to the same entity. In a dyadic situation, the speaker role needs to be incorporated in order to successfully constrain the resolution. For example, "you" from the Helper most likely refers to the same entity as "I" from the Worker, and vice-versa[28]. The second modification is for constraints based on locality. Perception of space and locality can change depending on the speaker. For example, while the Worker may use "this" to describe a local object, the Helper may also refer to the same object as "this," or they may refer to it as a distant object and use "that". Therefore, traditional constraints based on locality cannot simply be applied wholesale; rather the perceived space in which they are constructed needs to be taken into consideration.

"*Rule 2's role in pronoun resolution is not yet known (see Kehler, 1997 for a critique of its use by BFP), and that the preliminary evaluations of the BFP algorithm showed that without perfect information, the Rules could be over-constraining and thus do more harm than good.*"

[28] This is the case in dyadic conversation, however, this constraint becomes much more complex when attempting to implement it in a dialogue with more than two participants.

8.4.2 The visual-only model

The visual-only model captured the visible actions and utilized an approach based on visual salience. This method captured the relevant visual objects in the puzzle task and ranked them according to the level of recency with which they were active.

Given the highly controlled visual environment that makes up the PUZZLE CORPUS, timing information is available about when the pieces become visible, are moved, or are removed from the shared workspace (as previously demonstrated in Figure 8-2). In the visual-only model, an ordered list of entities that comprise the shared visual space was maintained. The entities are included in the list if they were visible to both the Helper and Worker, and then they were ranked according to the recency of their activation.

8.4.3 The integrated model

The integrated model took advantage of the salience list generated from the language-only model and integrated it with that of the visual-only model. The method of integrating the list was informed by general perceptual psychology principles stating that highly active visual objects attract attentional processes (for a recent review see Scholl, 2001).

In this implementation, I defined active objects as those objects that had recently moved within the shared workspace. These objects were added to the top of the linguistic-salience list which essentially rendered them the focus of the joint activity. However, people's attention to static objects tends to fade over time. Following prior work that demonstrated the utility of a visual decay function (Byron et al., 2005b; Huls et al., 1995), a three-second threshold existed 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 which the interlocutors revert to their recent linguistic history for the context of an interaction.

The integrated model also had a more practical implementation detail that allowed it to handle cases when the visual salience list was empty yet a pronoun was encountered. In this case, the integrated model used the linguistic salience list to suggest the potential antecedent. While the number of pronouns that were successfully resolved in this case was very small, it did tend to

improve performance somewhat, particularly at the beginning of trials when the pairs were discussing strategies or higher level entities surrounding the task.

It should be noted that this modeling is a 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. Together, these three models allow the testing of the basic hypotheses outlined in §7.4.

8.5 Results

8.5.1 Measures

The basic success measure used in this experiment is the successful resolution of a pronoun. The measure used followed that provided by Mitkov (2000) and was the total number of pronouns correctly resolved over the total number of pronouns attempted[29]. However, before the model performance can be assessed, the actual antecedents of the pronouns need to be marked. This was done using two expert coders that performed coding of the antecedents for each pronoun in the corpus. Each coder went through the segmented transcripts line by line and when they identified a pronoun they marked its antecedent, whether it was a noun phrase, another pronoun, or a visual entity or action. For the evaluation set examined in this study, the coders independently rated each of the potential 292 pronouns in the corpus. Scores were counted correct if both of the coders identified the pronoun and tagged the same antecedent. However, if only one of the coders identified a pronoun, or if the antecedents were different, their coding was scored as incorrect. Overall, the coders reached a reliability of 88% overall agreement. The remaining anomalies were resolved by discussion.

8.5.2 Statistical analysis

A number of analysis techniques were used throughout this experiment to describe the performance of the models. A logistic regression was used to examine the overall performance of the models and to capture higher-order interactions of interest. The model included Model Type (Language, Visual, Integrated), Lexical Complexity (Solid or Plaid), and Pronoun Type (Personal,

[29] As the system becomes more automated, more precise and measures such as precision, recall, and the F-measure can be used to report pronoun resolution and performance.

Demonstrative, or Demonstrative + NP). Because the pronouns existed in a discourse, there was the possibility that observations within a trial were not independent of one another. Therefore, each trial was modeled as a random effect[30]. In addition, all 2-way interactions were included in the model. Three-way interactions were also investigated, but were not found to be significant, and were subsequently removed from the final analysis.

In order to directly compare the performance of the models on each pronoun encountered, a second analysis involved the creation of a confusion matrix. McNemar's test was used to test the agreement between the models and to help characterize differences in their performance. This approach examined each pronoun that had been resolved for each model, and provided an indication of whether or not a particular model faired better on the same piece of data[31] which in turn provided an aggregate statistical indication of model performance and also allowed a more detailed investigation of the patterns of failure that occurred. For example, examination of the data points in the off-diagonals of the confusion matrix could provide an indication of how one particular model outperformed another.

8.5.3 Model performance results

Table 8-4 presents the pronoun resolution rates of the three models according to whether the pairs shared visual information, and whether the puzzles included simple solid colors or more lexically complex plaid pieces.

8.5.3.1 Model performance in the No Shared Visual Information condition

As can be seen in the "Total" columns of Table 8-4, the language-only model correctly resolved 67.1% 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 language-only model was applied to the dialogues from the task conditions where shared

[30] Graphical and statistical tests of the degree of serial correlation indicated that the degree of autocorrelation was actually quite low.

[31] This is done by examining the confusion matrix between the two models (i.e., Correct / Correct; Correct / Incorrect; Incorrect / Correct; and Incorrect / Incorrect) and testing the H_0: Correct / Incorrect = Incorrect / Correct. If H_0 is rejected, this tells you that one model likely performed better than the comparison model. The off-diagonals (Correct / Incorrect; Incorrect / Correct) can then be examined to explore the qualitative differences of the models.

visual information was available, it only resolved 49.3% of the referring expressions correctly. This difference was significant, $\chi^2_{(1, N=293)} = 7.17$, $p < .01$.

Table 8-4. Success rates for resolving pronouns in the subset of the PUZZLE CORPUS **evaluated.**

	No Shared Visual Information			**Shared Visual Information**		
	Solids	*Plaids*	***Total***	*Solids*	*Plaids*	***Total***
Language Model	*70.0% (21 / 30)*	*65.2% (30 / 46)*	***67.1% (51 / 76)***	*43.6% (17 / 39)*	*50.6% (90 / 178)*	***49.3% (107 / 217)***
Visual Model	*n/a*	*n/a*	***n/a***	*66.7% (26 / 39)*	*61.2% (109 / 178)*	***62.2% (135 / 217)***
Integrated Model	*70.0% (21 / 30)*	*65.2% (30 / 46)*	***67.1% (51 / 76)***	*69.2% (27 / 39)*	*73.0% (130 / 178)*	***72.4% (157 / 217)***

The integrated model with the decay function performed at the same level as the language-only model when applied in a setting without any shared visual information. When the integrated model was evaluated on the data where only language could be used it effectively reverts back to a language-only model, thereby achieving the same 67.1% performance.

8.5.3.2 Model performance in the Shared Visual Information condition

A comparison between the three models can be made by exploring their performance on the data in the cases in which shared visual information was available. Model Type was a significant factor in the model, $G^2_{(2)} = 15.21$, $p < .001$, and contrasts between the different levels of Model Type revealed differences between the performance of each model (at $p < .05$ in all cases).

The language-only model correctly resolved 49.3% of the pronouns when applied to the trials performed in the presence of shared visual information. However, when the visual-only model was applied to the same data, it correctly resolved 62.2% of the pronominal expressions. The difference in performance between these two models was substantial, $\chi^2_{(1, N=217)} = 8.52$, $p < .01$, and indicated a major performance benefit for the visual model.

The confusion matrix presented in Figure 8-3 demonstrates that both the visual-only and language-only models correctly resolved pronouns missed by the other. An informal examination of the cases that the visual-only model correctly resolved and the language-only model failed (27.1% of the cases) revealed a few trends. A large proportion of these cases appeared to occur when an efficient referring expression was used to reference an entity that was not mentioned in

the prior linguistic stream. For example, "Oh, that is one we need, so put it to the upper left". Another case was when contrastive statements were made regarding the current visible object and the targeted referent, for example, "…a darker color than that." A small number of cases also occurred when different discourse segments made a new set of linguistic entities available, yet the proper referent was presented earlier in the discourse (i.e., in another discourse segment). For example, a new discourse segment might exist regarding the higher-level positioning of the entire puzzle, as in "OK, the whole thing should be over to the right", followed by a return to the piece initially under discussion, "OK, you can move it into position now." This is one case were particular forms of discourse segment markers could aid the performance of the language-only model. This discussion is revisited later in the chapter. There were also a small number of references that the language-only model mistook to refer to sub-features of a piece, while the visual-only model correctly suggested the whole block as an entity.

An informal examination of the cases that the language-only model correctly resolved and the visual-only model failed (14.8% of the cases) also revealed trends. First, there were a number of cases where the language-only model successfully resolved pronouns to linguistic entities where the last piece of visual information would have led to an incorrect referent. These included cases when the discourse included longer discussions regarding the details of a piece or a layout. There were also cases where the language-only model could successfully resolve references within a sentence (i.e., intrasententially). And finally, there were a small number of cases where an incorrect visual object was available and the pronoun instead referred to a previously introduced linguistic entity (e.g., "no, it is a different yellow piece").

		Language	
		Incorrect	*Correct*
Visual	*Incorrect*	50 (23.0%)	32 (14.8%)
	Correct	59 (27.1%)	75 (34.6%)

Figure 8-3. Confusion matrix between the Language Model and the Visual Model.

When the integrated model was applied to the data from the cases when the pairs had access to the shared visual information, it correctly resolved 72.4% of the referring expressions. This was significantly better than the 49.3% exhibited by the language-only model. Once again, the

performance difference between these two models was sizeable, $\chi^2_{(1, N=217)} = 26.8$, $p < .01$. Similar to the last comparison, the confusion matrix in Figure 8-4 reveals that both the integrated and language-only models correctly resolved pronouns that the other model did not. In this comparison, there appeared to be substantially more cases (33.9% of the cases) that the integrated model identified versus those that the language-only model did (10.6% of the cases). The differences between these two models were similar to those discussed above in comparing the performance of the visual-only model with the language-only model. However, in this case, the integrated model could resort to the linguistic-salience list when the shared workspace was inactive, and therefore benefit from the ranking of entities based on linguistic-salience.

		Language	
		Incorrect	*Correct*
Integrated	*Incorrect*	37 (17.0%)	23 (10.6%)
	Correct	74 (33.9%)	84 (38.5%)

Figure 8-4. Confusion matrix between the Language Model and the Integrated Model.

Finally, the integrated model's 72.4% performance was significantly better than the visual-only model's 62.2% on the same data, $\chi^2_{(1, N=217)} = 17.29$, $p < .01$; indicating a potential performance benefit to having an integrated model across both the solid and plaid conditions. Figure 8-5 presents the confusion matrix, and it is interesting to note here that the integrated model nearly completely dominates the visual-only model. There are only three instances where the visual-only model correctly resolves a referent that the integrated model did not. All three of these instances were cases where a longer visual decay parameter would have captured the proper referent. However, a longer decay would also have had the ability to hurt the performance of the integrated model by inhibiting a switch to the linguistic salience list. Without performing an analysis of decay times, it is difficult to tell how many of the 25 successfully resolved pronouns that the visual-model failed on would be retained.

		Visual	
		Incorrect	*Correct*
Integrated	*Incorrect*	57 (26.3%)	3 (1.4%)
	Correct	25 (11.5%)	132 (60.8%)

Figure 8-5. Confusion matrix between the Visual Model and the Integrated Model.

Finally, a detailed examination of the types of pronouns successfully resolved differed across the model types. In other words, there was a significant Model Type × Pronoun Type interaction in the model, depicted in Figure 8-6 (for the interaction, $G^2_{(4)} = 17.43$, $p = .001$). An examination of this interaction reveals that the language-only model appears to perform best when resolving personal pronouns and decreases in success when resolving demonstrative pronouns, while the opposite trend is seen in both the visual-only and integrated models. This revelation reveals some interesting patterns regarding the appropriateness of the various models and suggests that future lines of work might explore strategic shifts in use of the visual-salience or linguistic-salience lists triggered by the syntactic information in the utterance.

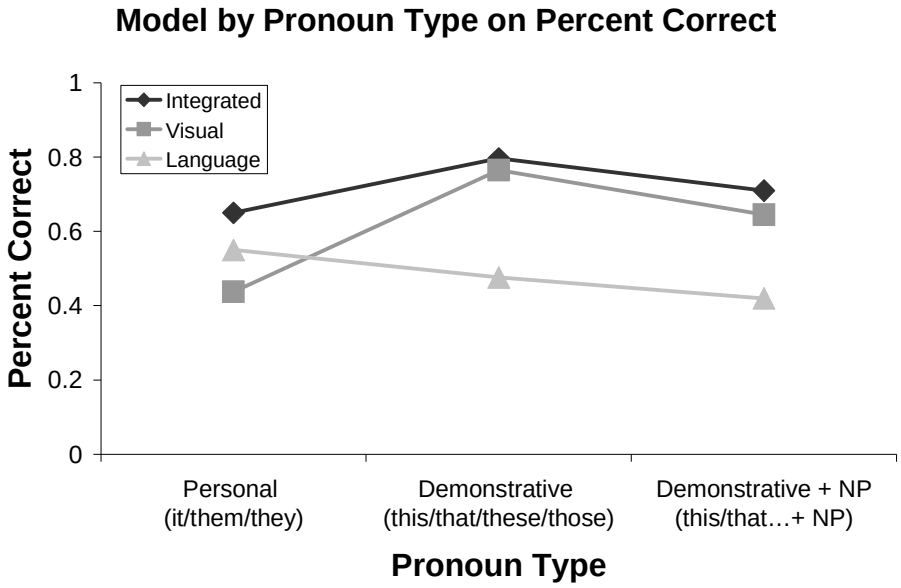


Figure 8-6. Effect of Model Type and Pronoun Type on successful pronoun resolution.

It is interesting to note that while the Plaid pieces were shown in prior studies to be linguistically complex, and while they typically required longer discourse segments and deeper reference chains, there was no indication of a significant Model Type × Lexical Complexity interaction. While initial examination of the models' performance under the plaid conditions when shared visual information was available (represented in the right-hand side of Table 8-4 under "Plaids") might suggest that the language-only model had a greater relative influence on performance when the task was linguistically complex, this interaction was not significant (for the interaction, $G^2_{(2)} = 0.31$, $p = .85$).

To summarize, 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 showed significantly increased performance over the language-only model; and the integrated model was the top performer overall. In addition, a more detailed analysis of the confusion matrices and direct comparison of the model results revealed when and where particular models worked and provided some reflection on why.

8.6 Error analysis

In order to inform further development of the model, a number of failure cases were examined in detail, particularly those in which all of the models failed. The first thing to note was that a number of the pronouns used by the pairs referred to larger visible structures in the workspace. An example of this was when the Worker would state, "like this?", and ask the Helper to comment on the overall configuration of the puzzle. In the current model, only the puzzle pieces are included as possible visual referents. One approach to alleviating this error is to integrate a richer notion of semantics with the additional visual entities in order to accurately model such situations (e.g., see Byron et al., 2005b).

Another area where the models suffered performance problems was during references to higher-order referents such as general events or the state of the world. For example, "OK, this is going to be tough" where "this" specifies the general construction of the puzzle. Similarly, non-referential "its" as in "It is easy to make something work" posed a problem for the models. These are both common problems in reference resolution and may be addressed in the future by applying recent research advances in these areas. For example, recent work by Müller (2006) provides an

automated method for filtering out non-referential "its," and this technique could be applied to refine the pronouns attempted by applying a filter earlier on in the processing pipeline. Similarly, Byron and other colleagues have demonstrated recent work aimed at addressing references to abstract entities (Byron, 2002).

In addition, there were several errors that resulted from chaining errors: When the initial referent was misidentified all subsequent chains of referents were incorrect. It should be noted that the approach used in this study to score the success of the resolved pronouns followed Walker's original description (Walker, 1989), in which she describes how chains of referents are scored as incorrect if the original binding is incorrect. This makes sense from a systems perspective where incorrect inferences could be made if the initial referent is incorrect. However, recent evaluations have used a more lenient formulation whereby a "location"-based evaluation procedure is used (Tetreault, 2001). This approach marks co-references as correct if they co-refer with an NP that has the same co-reference tag. In other words, these studies only look one step back and do not penalize performance for longer "error chains".

Finally, 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. The implementation presented here followed the notion that actions typically precede the keywords or language that is associated with a given action (Oviatt *et al.*, 1997). Future work could include a richer model of gestures and spoken language alignment in order to successfully account for such issues (e.g., Eisenstein & Christoudias, 2004).

8.7 Discussion

The results of this experiment find that the language-only model performs in the range of previous studies of pronoun resolution on spoken discourse by successfully resolving approximately 67% of the pronouns encountered. This apparent success is due, in part, to the fact that the salience-based approach works well for anaphora resolution since it captures the many well-known syntactic and psycholinguistic factors that contribute to entity salience. However, when the language-only model is applied to the portions of the corpus that include trials in which the pairs had access to shared visual information, its performance suffers. In fact, the application of the language-only model to the trials undertaken with shared visual information performs

below 50%. One reason for this is that when shared visual information is available, action and language use can become interchangeable (Gergle et al., 2004a); this is captured in the models by the fact that the visual-only model performs at 62% and it performs better in many instances.

Overall, the integrated model was the best performer in this evaluation. Its performance is equivalent to the language-only model during trials without shared visual information available to the pair. This results from the fact that there are no visual entities available for the model, and therefore it reverts to using linguistic salience as a source for resolution. However, when applied to the cases where shared visual information was available, the integrated model performed significantly better than the language-only model. This is due, in part, to the fact that it can take capture references to the physical objects that have not been mentioned using speech.

A comparison of the integrated model to the visual-only model yielded interesting results. The first area of benefit is that the integrated model successfully resolves reference when no shared visual information is available, while the visual-only model cannot[32]. When shared visual information is available the integrated model significantly outperforms the visual-only model, and this difference was there regardless of the lexical complexity of the puzzle pieces. One way in which the integrated model seemed to capture elements of the discourse that were neglected by the visual-only model occurred in cases where there was prolonged discussion about the particular features of a given entity. As a result, the decay parameter allowed the model to essentially shift its focus from active visual events to the dialogue or conversation currently taking place. In a sense, this mimics the shift in attention that occurs between participants as they fluidly move between referring to objects and actions in the environment to those discourse entities produced in the spoken dialogue stream. An interesting future avenue of research could explore whether a richer discourse model, such as one that provides discourse segment markers, could increase performance above the 72.4% currently achieved in this study. Together these findings provide strong support for the third hypothesis described in §7.4. Indeed, both linguistic entities and visual entities were central to the accurate and perspicuous accounting of the patterns of observed referring behavior.

[32] A modification to the visual-only model could be made whereby it makes use of the local visual information for each of the participants and suggests relevant entities for reference. However, such a setup would breakdown when applied to settings with asymmetries in the displays.

8.7.1 Generalizability of the models

The evaluation presented in this chapter demonstrates the strength of the integrated approach to modeling reference in dyadic, multi-modal interactions. However, a number of questions remain regarding its viability as a general approach to reference resolution in a wider variety of contexts. There are three major areas that merit further discussion, and they include questions of: (1) the applicability of this approach to a wider range of tasks and task-oriented situations, (2) the feasibility of this approach to interactions in group settings that contain more than two conversational partners, and (3) the role played by more general domain and task knowledge structure and its impact on discourse patterns.

To address the first of these concerns it is useful to return to the points made in Chapter 7. This chapter described how the architecture was designed in a flexible manner so that it could describe a number of task-oriented interactions. In particular, the framework is rich in its ability to model a number of the possible collaborative visual environments discussed throughout this dissertation. For example, asymmetries in the shared visual space can be captured by providing differential access to visual entities. This allows the model to easily adapt to things like delays in one person's access to the visual information or differences in spatial alignments that may be due to different camera orientations. Changes to the speech stream can be managed in a similar way. For example, if sentences are not heard due a large amount of acoustic noise, a particular utterance or verbal entity can be kept from inclusion in the salience list. Yet, the models developed in this chapter do need to be explored in more task environments than just the PUZZLE CORPUS. In the future it would be fruitful to extend the models to more realistic task domains such as those studied in work by Fussell and colleagues (Fussell et al., 2003a; Fussell et al., 2003b) and Kuzuoka and colleagues (Kuzuoka *et al.*, 2000). These environments are good candidates for extension since they provide task-oriented interactions yet they are unique in the visual settings and environments explored. In addition, the studies are performed in more realistic or "real world" settings. Extending the models to these environments would greatly benefit development by examining their potential to describe events in novel collaborative environments.

Another question regarding the general nature of the modeling environment is its ability to model interactions and communications in larger groups. In other words, can the same computational description be used to model groups that are larger than dyads? There is nothing in the model that prevents the modeling of larger group sizes. First, a unique Transient Knowledge Base can be generated for each actor and contain parameters that describe their current belief regarding the

information possessed by others involved in the discourse or task. A similar approach would be to modify the ranking of entities according to a global set of parameters that capture the individual states of each actor and their beliefs about the others. However, while the modeling architecture allows for this expansion, major theoretical questions remain about the theoretical and psychological plausibility of such an approach. Application of the model to larger groups may provide insight into current theoretical debates surrounding whether we maintain a complete model of all conversational partners or use a probabilistic approach when generating and comprehending speech in collaborative environments. This work could benefit, as well as be informed by, a more detailed approach to audience design and larger group interactions, and model enhancements may need to be made in order to maintain a certain degree of psychological plausibility when growing the model (see e.g., Clark & Murphy, 1982; Fay *et al.*, 2000).

Finally, a notion of domain knowledge needs to be addressed in future work. Experts bring to a task a certain level of task knowledge, and the shared knowledge between experts can make particular task relevant entities more or less salient. The current models do not take this pre-existing shared task knowledge into account. However, future work could extend the world knowledge component so that it allows the modeler to include notions of task structure that may impact speech patterns. For example, if a group of surgeons was beginning a surgery and everyone was aware of the setting and task, a surgeon may say something like, "it needs to be upped" in reference to the initial anesthetic dosage. Shared knowledge of the task structure would allow the group members to resolve the "it" to the current dosage because of its relevant task status even if it had not been previously mentioned. Such an influence of shared task knowledge about and its impact on reference is not currently accounted for in the models described in this chapter, yet this could be an area of future modeling work. This work could include general task and discourse notions such as things like the current question under discussion (Roberts, 1996) as well as more general notions surrounding general task knowledge as previously described. However, it is important to keep in mind that the integration of task and domain knowledge such as this is a bit of a double-edged sword. While it may make the model more accurate and able to predict a larger set of referring behaviors, it would do so at the cost of generalizability and parsimony of the models.

8.8 Future work

In the future, I plan to extend this work in several ways. First, a fully-automated version of the models is currently under development. This work constitutes full computational automation of

the stages represented in Figure 8-1. Part of this work will involve a large-scale computational evaluation of the entire PUZZLE CORPUS in order to examine a 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, I plan to extend the models to a wider range of task domains (as described above) in order to explore the external validity of the model predictions. Second, I plan future studies to help expand a 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. Future work will explore these parameters in detail. Third, I plan to appreciably enhance the integrated model. It appears from both the initial data analysis and a qualitative examination of the model performance that the pairs make tradeoffs between reliance on the linguistic and visual context. However, the current instantiation of the integrated model could be enhanced by taking a more theoretically informed approach to integrating the information from multiple streams.

Chapter 9

Conclusion

The work in this dissertation detailed a collection of empirical studies, described a number of statistical and analytical methodologies, and developed and evaluated an explicit rule-based computational model. Together this work presents a rich, multidisciplinary investigation of the role that shared visual information plays during collaborative task-oriented interactions, and its contributions fall into three major categories: theoretical, methodological, and applied. The theoretical contributions advance our understanding of the ways in which pairs use visual evidence for collaborative purposes and illuminate the basic principles of conversation and interaction in a variety of communication settings. The methodological contributions include the application of statistical and analytical techniques to provide novel insights into collaborative interactions in multimodal settings. The applied contributions result from the knowledge uncovered regarding our understanding of how communication media influence collaboration, and they provide a level of understanding that can be applied to the future design and development of collaborative technologies. The following sections provide a review of many of the central findings within each of these three major categories.

9.1 Theoretical contributions

The goal of the first stage of this work was to provide a detailed description of how features of the visual environment, commonly traded off in technologies designed to support distance collaboration, interact with particular task features to impact the ability of pairs to coordinate and communicate in an efficient manner. The results of these studies provide several theoretical contributions.

Broad empirical support for the cooperative model of communication

The data in this dissertation provide broad support for the cooperative model of communication; for example, Workers adapted their communication and behavior to compensate for what the Helper could or could not see. It is important to note that in the puzzle task, the Worker's view of the workspace was always the same regardless of whether the Helper could see it. Therefore, if Workers were using a purely egocentric approach to communication, they would not change their communication behavior in response to variations in the shared visual information because their view of the space never changed. Instead, they changed their communicative behavior in response to what they knew their partner could see. When the Helper could not see the work area, Workers used more words to complete the task, were more likely to describe the work area after they made moves, and were more likely to indicate explicitly whether they understood an instruction.

Detailed empirical support and an extension to Clark & Brennan's (1991) hypothesis that different media features change the cost of achieving common ground

The results in this dissertation are also consistent with Clark and Brennan's (1991) framework for analyzing the costs and benefits of different communication technologies. When media provide visual information about what the Worker is doing, the Workers' ability to ground their utterances via actions reduces their need to provide verbal indicators of comprehension. Instead, they let their actions demonstrate their understanding of the Helpers' instructions. The sequential analysis techniques showed that the Helper's instructions were more likely to be followed by the Worker's movement of a puzzle piece when the shared visual information was available versus when it was not. In contrast, a Helper's instructions were more likely to be followed by a Worker's acknowledgement of understanding when there was no shared visual information available. These findings, and others presented throughout the studies, demonstrate that the availability of various sub-features of the shared visual information can influence the resources available for grounding.

An improved theoretical understanding of how features of the task interact with features of the media to impact communication and coordination

This work extends the work of Clark and Brennan (1991) by illustrating how features of the task interact with features of the communication setting to influence the grounding process. In several of the studies, the value of the shared visual information depended upon the task being performed. For example, the shared visual information served performance and conversational efficiency

more when the tasks were dynamic or the objects in the environment were rapidly changing. However, when the objects were easy to describe or the environment and objects were static and unchanging, the benefits from the shared visual information were greatly diminished. These interactions between the features of the shared visual information and the features of the task demonstrate the importance of understanding task characteristics when determining the value of a shared visual workspace. These findings help to rectify the disparity between early and more recent research on the value of visual information in distributed communication.

An improved understanding of the roles played by conversational grounding and situation awareness in collaborative task performance

The work presented in these chapters provides broad empirical support for the coordination mechanisms of conversational grounding and situation awareness. Convergent findings from the collection of studies also provide some of the first empirical data to differentiate between the theoretical importance of conversational grounding and that of situation awareness as coordination mechanisms for group communication and performance.

An understanding of where visual information is used by conversational partners as they ground their utterances during a collaborative task

This work also provided an understanding of *where* the visual information was particularly useful during task-oriented collaboration. The application of sequential analysis techniques uncovered how visible actions support understanding in the discourse and allowed detailed statistical examination of the patterns of language use and actions that led to successful collaborative performance.

In addition, these techniques provided insight into the process level details of the locations in the overall task in which the shared visual information was particularly useful. This results in a significant advance for theories of interpersonal communication by providing a richer description of the importance of visual evidence for communication and conversational efficiency along with a detailed analysis of precisely where this visual information is used by participants during a collaborative task.

Developing a novel methodology for integrating contextual entities with linguistic entities in a real-time computational model of discourse salience and coherence

Finally, the development of a computational model of reference in the presence of visual information contributes to our theoretical knowledge by describing exactly *how* the visual information can play a role in supporting communication and collaboration. The results of the work presented in Chapters 1 through 6, as well as prior literature, suggested that a primary area of impact that shared visual information had was on the ability of pairs to effectively make use of shared visual information to resolve ambiguity and generate efficient referring expressions. The development of a computational model detailing how visual information is combined with linguistic cues to enable effective reference-making during tightly-coupled task-oriented collaborations furthers our theoretical understanding of how visual information influences language use and expresses this understanding in a computational form.

This work established a novel methodology for integrating contextual entities with linguistic entities in a real-time computational model of discourse salience and coherence. It also helped to uncover the relevant linguistic and visual structures at play during task-oriented collaborative interactions and served to describe their interactions. Finally, it has begun to describe the primary linguistic and visual features that are required for a successful model of reference in the presence of shared visual information.

9.2 Methodological contributions

Overall, this work provides a unique demonstration of a multidisciplinary research approach that applies techniques from behavioral research, discourse analysis and computational linguistics in a closely integrated fashion to produce complementary findings and demonstrate a fruitful and efficient exploration of a design space. In addition to this demonstration of the merits of such a multidisciplinary approach, a number of concrete methodological contributions were made. Among these are the development of a rigorous experimental paradigm for decomposing the elements of shared visual space and studying their impact on collaborative performance. This work demonstrated a technique for experimentally manipulating features of a shared visual environment and provides a useful empirical tool for observing the influence of these manipulations on task performance and communication processes. It also demonstrated a systematic method for examining small group interactions as they unfold over time by applying sequential modeling techniques to multi-actor, multi-stream data commonly found in collaborative applications. Finally, it provided a number of contributions along the lines of understanding and evaluating multimodal data. In addition, two major statistical adaptations were

used in this work that can provide a demonstration of the application of a wide-range of methodologies to enhance our theoretical understanding.

Application of sequential analysis techniques

The use of sequential analysis techniques and the detailed coding of speech and actions demonstrate in much greater depth how the availability of a shared view of the workspace affects communicative interaction. The use of log-linear modeling and multi-way contingency table analysis yields deeper insight into the communication processes as they unfold over time. This application of sequential analysis techniques to multimodal interactions in order to understand their sequential structure is a unique methodological contribution of this work.

Application of MARS methodology

The application and demonstration of the use of a statistical method that allows us to examine collaborative task performance over a continuous range of visual delays is another unique methodological contribution of this work. This methodology provides detailed insight into the range of delays within which collaborative task performance is not affected, as well as uncovers the points at which performance begins to break down. In addition, examination of the corresponding slope coefficients provides an indication of the relative impact of additional delays on performance. Application of this methodology to an outstanding theoretical quandary provided a unified description of what was previously a collection of disparate findings from earlier work that examined discrete levels of delay but could not pinpoint the precise time at which collaborative performance breaks down in the presence of delayed visual information.

9.3 Applied contributions

The practical contributions of this work address a wide range of applications and can inform the development of future collaborative systems. Knowledge of the mechanisms by which visual information can augment and change communication is crucial for the design of systems that support remote collaboration, particularly in instances where support for collaborative physical tasks is the goal. By identifying the ways in which visual information and speech interoperate, we can begin to make informed design decisions regarding ways to support visual information in collaborative applications. The results presented in this dissertation highlight the importance of making it clear that people know precisely what remote collaborators can see in a shared workspace. It is not enough to simply allow others to see what is going on, but rather, mutual understanding of what is available to one another is needed. When confusion exists regarding

what the Helpers can see, the pairs spend time trying to identify the mutually shared visual field. This reduces their overall efficiency since significant time is needed to determine what visual information is and is not shared (Kraut et al., 2003).

In the puzzle task there are two levels at which the visual information seems particularly useful. At a higher level, the pairs find it useful for task planning. For example, when planning subsequent directives the Helper often looks at surrounding contextual information. Previous work has suggested that the Helper often looks to the instructions while formulating her description of the next step (Fussell et al., 2003b). In this case, providing a wide-angle view of the workspace (i.e., a context-oriented view) is useful. However, when pairs are performing lower level coordination of their language it is useful to have a focus-view of the workspace centered on the actions. Thus, for high-level task planning it may be useful to have a wider view of the work area, while for grounding communications it may be more useful to have focused views. A potential design avenue for simultaneously supporting these two levels might be through the creation of task specific focus + context designs. Initial design avenues in this area have been explored by Schafer and Bowman in exploring collaborative spatial navigation (Schafer & Bowman, 2003), and by Greenberg, Gutwin and Cockburn as general techniques in groupware applications (Greenberg *et al.*, 1996). Coupling these design explorations with detailed knowledge of how visual information serves the task may lead to a fruitful line of collaborative applications development for joint physical tasks. This dissertation also demonstrated that when collaborators are aware of their partners' fields of view, asymmetric interfaces in which different parties have different modes of accessing the environment appear to be surprisingly functional. Developing ways of providing awareness of others' views can enable efficient grounding and be crucial to the development of successful applications for remote collaboration on physical tasks.

Throughout this dissertation, the findings suggest that actions provide a more efficient mechanism for establishing mutual understanding. Rather than relying on imprecise conversation to determine if something had been done correctly, having it in view to verify mutual understanding was extremely useful, particularly in a tightly coordinated activity or one in which the expertise is distributed. This may suggest that the use of schematic representations in lieu of direct video feeds in low bandwidth conditions may be more useful to participants if they represent the group actions rather than the others' faces or bodies. For example, sensors might provide schematic feedback about what objects have been selected or moved. The value of schematic representations has been shown in similar settings by such tools as Gutwin and

Penner's telepointer traces (Gutwin & Penner, 2002), which provide feedback about a partner's trajectory of cursor movements within a shared workspace.

In addition to the benefits of the empirical work, the modeling performed in the latter portion of this thesis provides a number of practical benefits. For example, its explicit computational description can be used to develop more natural conversational interactions with human actors in a variety of human-to-human, human-to-computer, human-to-agent, and human-to-robot interactions. It can be used to augment and improve the performance of state-of-the-art models of communication currently used in natural language generation systems, and can be used to develop systems that emulate increasingly naturalistic and realistic human conversational behavior.

The models can also be used to provide insight into when, how and why certain pieces of visual information need to be presented to remote collaborators. In particular, a computational description of ambiguous and incoherent states of a discourse can be used to augment systems to provide an indication of when and where particular pieces of verbal of visual information might be needed.

9.4 Closing remarks

Throughout this dissertation, I have argued that shared visual information is essential for complex task-oriented collaborations because it facilitates the ability of the pairs to maintain awareness of the task state, helps them to reduce errors and ambiguities when the environment is visually complex, and facilitates grounding and communication by allowing the use of efficient language as a method for monitoring comprehension. The effects of new communication technology are not superficial, and their developers should not be guided by surface characteristics. By considering the ways that technologies, and the tasks we attempt with their aid, interact with, modify, and rely on language, greater strides can be made in understanding and design. Moreover, these developments illuminate basic principles of conversation and group behavior in profound ways, bringing into focus not only technological but traditional communication processes. While further work remains in order to completely understand the impact of shared visual information on task-oriented collaboration, this thesis provides a major step forward in providing a theoretically grounded understanding of the ways in which shared visual information influences collaborative performance, as well as a direction for the future development of technologies to better enable distance collaboration.

Bibliography

Akaike, H. (1978). A Bayesian Analysis of the Minimum AIC Procedure. *Annals of the Institute of Statistical Mathematics, 30*(1), 9-14.

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 (Recent progress on TRIPS). In Proceedings of *The Association for Computational Lingustics (ACL '05), Companion Volume*, pp. 85-88.

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

Alshawi, H. (1987). *Memory and context for language interpretation*. Cambridge: Cambridge University Press.

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E. H., Doherty, G. M., Garrod, S. C., et al. (1991). The HCRC Map Task Corpus. *Language & Speech, 34*(4), 351-366.

Argyle, M., & Cook, M. (1976). *Gaze and mutual gaze*. Cambridge: Cambridge University Press.

Bakeman, R., & Gottman, J. M. (1997). *Observing interaction: An introduction to sequential analysis*. Cambridge: Cambridge University Press.

Bakeman, R., & Quera, V. (1995). *Analyzing interaction: Sequential analysis with SDIS and GSEQ*. Cambridge: Cambridge University Press.

Barnard, P., May, J., & Salber, D. (1996). Deixis and points of view in media spaces: An empirical gesture. *Behaviour and Information Technology, 15*(1), 37-50.

Bekker, M. M., Olson, J. S., & Olson, G. M. (1995). Analysis of gestures in face-to-face design teams provides guidance for how to use groupware in design. In Proceedings of *ACM Conference on Designing Interactive Systems (DIS '95)*, pp. 157-166. ACM Press.

Bolstad, C. A., & Endsley, M. R. (1999). Shared mental models and shared displays: An empirical evaluation of team performance. In Proceedings of *43rd Meeting of the Human Factors & Ergonomics Society*, pp. 213-217.

Boyle, E. A., Anderson, A. H., & Newlands, A. (1994). The effects of visibility on dialogue and performance in a cooperative problem solving task. *Language & Speech, 37*(1), 1-20.

Brennan, S. E. (1990). *Seeking and providing evidence for mutual understanding.* Unpublished doctoral thesis, Stanford University.

Brennan, S. E. (1995). Centering attention in discourse. *Language & Cognitive Processes, 10*(2), 137-167.

Brennan, S. E. (2005). How conversation is shaped by visual and spoken evidence. In J. Trueswell & M. Tanenhaus (Eds.), *Approaches to studying world situated language use: Bridging the language-as-product and language-as-action traditions* (pp. 95-130). 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.

Brennan, S. E., & Lockridge, C. B. (In preparation). Monitoring an addressee's visual attention: Effects of visual co-presence on referring in conversation.

Brown, P., & Dell, G. S. (1987). Adapting production to comprehension: The explicit mention of instruments. *Cognitive Psychology, 19*, 441-472.

Byron, D. K. (2002). Resolving pronominal reference to abstract entities. In Proceedings of *40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pp. 80-87.

Byron, D. K., Dalwani, A., Gerritsen, R., Keck, M., Mampilly, T., & Sharma, V. (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. 83-96.

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

Carletta, J., Isard, S. D., Doherty-Sneddon, G. M., Isard, A., Kowtko, J. C., & Anderson, A. H. (1997). The reliability of a dialogue structure coding scheme. *Computational Linguistics, 23*(1), 13-31.

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.

Chafe, W. L. (1976). Givenness, Contrastiveness, Definiteness, Subjects, Topics, and Point of View. In C. N. Li (Ed.), *Subject and Topic* (pp. 25-55). New York, NY: Academic Press.

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 (IUI '05)*, pp. 43-50. NY: ACM Press.

Chambers, C. G., Tanenhaus, M., Eberhard, K. E., Filip, H., & Carlson, G. N. (2002). Circumscribing referential domains in real-time sentence comprehension. *Journal of Memory & Language, 47*, 30-49.

Chapanis, A., Ochsman, R., Parrish, R., & Weeks, G. (1972). Studies in interactive communication: I. The effects of four communication modes on the behavior of teams during cooperative problem-solving. *Human Factors, 14*(6), 487-509.

Chomsky, N. (1982). *Some concepts and consequences of the theory of government and binding*. Cambridge, MA: MIT Press.

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

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. Resnick, J. Levine & S. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127-149). Washington DC: American Psychological Association.

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

Clark, H. H., & Marshall, C. R. (1981). Definite reference and mutual knowledge. In A. K. Joshi, B. L. Webber & I. A. Sag (Eds.), *Elements of discourse understanding*. Cambridge: Cambridge University Press.

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

Clark, H. H., & Schaefer, E. F. (1989). Contributing to discourse. *Cognitive Science, 13*(2), 259-294.

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

Cohen, K. (1982). Speaker interaction: Video teleconferences versus face-to-face meetings. In Proceedings of *Teleconferencing and Electronic Communications*, pp. 189-199. University of Wisconsin Press.

Cohen, P. (1984). The pragmatics of referring and the modality of communication. *Computational Linguistics, 10*, 97-146.

Craven, P., & Wahba, G. (1979). Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik, 31*, 317-403.

Daelemans, W., Buchholz, S., & Veenstra, J. (1999). Memory-based shallow parsing. In Proceedings of *Third Computational Natural Language Learning Workshop (CoNLL)*, pp. 53-60.

Daelemans, W., Zavrel, J., van der Sloot, K., & van den Bosch, A. (2001). *TiMBL: Tilburg memory based learner, version 4.0, reference guide.* (No. ILK Technical Report 00-01): Tilburg University.

Daft, R., & Lengel, R. (1986). Organizational information requirements, media richness and structural design. *Management Science, 32*(2), 554-571.

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 (ACL 2005), Companion Volume*, pp. 1-4.

Doherty-Sneddon, G., Anderson, A., O'Malley, C., Langton, S., Garrod, S., & Bruce, V. (1997). Face-to-face and video mediated communication: A comparison of dialog structure and task performance. *Journal of Experimental Psychology: Applied, 3*, 105-125.

Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., & Tanenhaus, M. K. (1995). Eye Movements as a Window into Real-Time Spoken Language Comprehension in Natural Contexts. *Journal of Psycholinguistic Research, 24*(6), 409-436.

Eckert, M., & Strube, M. (2000). Dialogue acts, synchronising units and anaphora resolution. *Journal of Semantics, 17*(1), 51-89.

Eisenstein, J., & Christoudias, C. M. (2004). A salience-based approach to gesture-speech alignment. In Proceedings of *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pp. 25-32.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors Special Issue: Situation Awareness, 37*(1), 32-64.

Endsley, M. R., & Garland, D. J. (2000). *Situation awareness analysis and measurement*. Mahwah, NJ: LEA.

Fay, N., Garrod, S. C., & Carletta, J. (2000). Group discussion as interactive dialogue or as serial monologue: The influence of group size. *Psychological Science, 11*(6), 481-486.

Ferreira, V. S., & Dell, G. S. (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology, 40*, 296-340.

Fienberg, S. E. (1978). *The analysis of cross-classified categorical data*. Cambridge, MA: MIT Press.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *Annals of Statistics, 19*, 1-141.

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

Fussell, S. R., Kraut, R. E., & Siegel, J. (2000). Coordination of communication: Effects of shared visual context on collaborative work. In Proceedings of *ACM Conference on Computer Supported Cooperative Work (CSCW 2000)*, pp. 21-30.

Fussell, S. R., Setlock, L. D., & Kraut, R. E. (2003a). Effects of head-mounted and scene-oriented video systems on remote collaboration on physical tasks. In Proceedings of *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 513-520. ACM Press.

Fussell, S. R., Setlock, L. D., & Parker, E. M. (2003b). Where do helpers look? Gaze targets during collaborative physical tasks. In Proceedings of *ACM Conference on Human Factors in Computing Systems (CHI '2003) (Extended Abstracts)*, pp. 768-769.

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. (2004a). Action as language in a shared visual space. In Proceedings of *ACM conference on Computer supported cooperative work*, pp. 487-496. ACM Press.

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

Gergle, D., Kraut, R. E., & Fussell, S. R. (2006). The Impact of Delayed Visual Feedback on Collaborative Performance. In Proceedings of *ACM Conference on Human Factors in Computing Systems (CHI 2006)*, pp. 1303-1312. NY: ACM Press.

Gergle, D., Kraut, R. E., & Fussell, S. R. (Under Review). Using Visual Information for Grounding and Awareness in Collaborative Tasks.

Gergle, D., Millen, D. E., Kraut, R. E., & Fussell, S. R. (2004c). Persistence matters: Making the most of chat in tightly-coupled work. In Proceedings of *ACM Conference on Human Factors in Computing Systems (CHI 2004)*, pp. ACM Press.

Goodman, L. A. (1978). *Analyzing qualitative/categorical data: Log-linear models and latent structure analysis*. Cambridge, MA: Abt Books.

Goodwin, C. (1996). Professional vision. *American Anthropologist, 96*, 606-633.

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

Greenberg, S., Gutwin, C., & Cockburn, A. (1996). Awareness through fisheye views in relaxed-WYSIWIS groupware. In Proceedings of *Graphics Interface*, pp. 28-38.

Grosz, B. J., Joshi, A. K., & Weinstein, S. (1983). Providing a unified account of definite noun phrases in discourse. In Proceedings of *Proceedings of Association of Computational Linguistics (ACL-83)*, pp. 44-50.

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.

Gujarati, D. N. (1995). *Basic Econometrics* (3rd ed.): McGraw-Hill, Inc.

Gundel, J. K., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language, 69*(2), 274-307.

Gutwin, C. (2001a). The effects of network delays on group work in real-time groupware. In Proceedings of *European Conference on Computer-Supported Cooperative Work*, pp. 299-318.

Gutwin, C. (2001b). *Groupware latency and jitter on the Internet* (Technical Report): Department of Computer Science, University of Saskatchewan.

Gutwin, C., Benford, S., Dyck, J., Fraser, M., Vaghi, I., & Greenhalgh, C. (2004). Revealing delay in collaborative environments. In Proceedings of *ACM Conference on Computer Human Interaction (CHI 2004)*, pp. 503-510. NY: ACM Press.

Gutwin, C., & Penner, R. (2002). Improving interpretation of remote gestures with telepointer traces. In Proceedings of *Computer Supported Cooperative Work (CSCW 2002)*, pp. 49-57.

Hanna, J. E., & Tanenhaus, M. (2004). Pragmatic effects on reference resolution in a collaborative task: Evidence from eye movements. *Cognitive Science, 28*, 105-115.

Hanna, J. E., Tanenhaus, M., & Trueswell, J. (2003). The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory & Language, 49*(1), 43-61.

Heeman, P. A., & Allen, J. (1995). *Dialogue transcription tools*: Trains TN 94-1, Computer Science Dept., University of Rochester.

Hobbs, J. R. (1976a). *A computational approach to discourse analysis*. New York: Department of Computer Science, City College, City University of New York.

Hobbs, J. R. (1976b). *Pronoun resolution*. New York: Department of Computer Science, City College, City University of New York.

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.

Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition, 59*, 97-117.

Hudson, S. B., Tanenhaus, M. K., & Dell, G. S. (1986). The effect of the discourse center on the local coherence of a discourse. In Proceedings of *Proceedings of the 8th Annual Conference of the Cognitive Science Society*, pp. 96-101. Lawrence Erlbaum Associates.

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

Hupet, M., Seron, X., & Chantraine, Y. (1991). The effects of the codability and discriminability of the referents on the collaborative referring procedure. *British Journal of Psychology, 82*(4), 449-462.

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

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. (1997). Current theories of centering for pronoun interpretation: A critical evaluation. *Computational Linguistics, 23*(3), 467-475.

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.

Kehler, A., Martin, J., Cheyer, A., Julia, L., Hobbs, J. R., & Bear, J. (1998). On representing salience and reference in multimodal human-computer interaction. In Proceedings of *American Association for Artificial Intelligence (AAAI '98)*, pp. 33-39.

Kelleher, J., & Genabith, J. v. (2004). Visual salience and reference resolution in simulated 3D environments. *Artificial Intelligence Review, 21*(3), 253-267.

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science, 11*(1), 32-38.

Keysar, B., Barr, D. J., Balin, J. A., & Paek, T. S. (1998). Definite reference and mutual knowledge: Process models of common ground in comprehension. *Journal of Memory & Language, 39*(1), 1-20.

Krauss, R. M., & Bricker, P. D. (1967). Effects of Transmission Delay and Access Delay on the Efficiency of Verbal Communication. *Journal of the Acoustical Society of America, 41*(2), 286-292.

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

Krauss, R. M., & Weinheimer, S. (1966). Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personality & Social Psychology, 4*(3), 343-346.

Kraut, R. E., Fussell, S. R., Brennan, S. E., & Siegel, J. (2002a). Understanding Effects of Proximity on Collaboration: Implications for Technologies to Support Remote Collaborative Work. In P. Hinds & S. Kiesler (Eds.), *Distributed Work* (pp. 137-164). Cambridge, MA: MIT Press.

Kraut, R. E., Fussell, S. R., & Siegel, J. (2003). Visual information as a conversational resource in collaborative physical tasks. *Human-Computer Interaction, 18*(1), 13-49.

Kraut, R. E., Gergle, D., & Fussell, S. R. (2002b). The use of visual information in shared visual spaces: Informing the development of virtual co-presence. In Proceedings of *ACM*

Conference on Computer Supported Cooperative Work (CSCW 2002), pp. 31-40. ACM Press.

Kraut, R. E., Miller, M. D., & Siegel, J. (1996). Collaboration in Performance of Physical Tasks: Effects on Outcomes and Communication. In Proceedings of *Proceedings, Computer Supported Cooperative Work Conference, CSCW'96.*, pp. 57-66. ACM Press.

Kuzuoka, H., Oyama, S., Yamazaki, K., Suzuki, K., & Mitsuishi, M. (2000). GestureMan: A mobile robot that embodies a remote instructor's actions. In Proceedings of *ACM Conference on Computer Supported Cooperative Work (CSCW 2000)*, pp. 155-162.

Levelt, W. J. M. (1982). Cognitive styles in the use of spatial direction terms. In R. J. Jarvella & W. Klein (Eds.), *Speech, place and action*. Chichester: Wiley.

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

Lockridge, C. B., & Brennan, S. E. (2002). Addressees' needs influence speakers' early syntactic choices. *Psychonomic Bulletin & Review, 9*(3), 550-557.

Miller, G. A., & Johnson-Laird, P. N. (1976). *Language and perception*. Cambridge, MA: Harvard University Press.

Mitkov, R. (2000). Towards a more consistent and comprehensive evaluation of anaphora resolution algorithms and systems. In Proceedings of *Discourse anaphora and reference resolution conference (DAARC2000)*, pp. 96-107.

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

Müller, C. (2006). Automatic detection of nonreferential it in spoken mult-party dialog. In Proceedings of *11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, pp. 49-56.

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

Nakajima, S., & Allen, J. (1993). A study on prosody and discourse structure in cooperative dialogues. *Phonetica, 50*, 197-210.

Nardi, B., Schwarz, 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 *ACM Conference on Human Factors in Computing Systems (INTERCHI 1993)*, pp. 327-334.

Nardi, B., & Whittaker, S. (2002). The Place of Face to Face Communication in Distributed Work. In P. Hinds & S. Kiesler (Eds.), *Distributed Work*. Cambridge, MA: MIT Press.

O'Conaill, B., & Whittaker, S. (1997). Characterizing, predicting, and measuring video-mediated communication: A conversational approach. In K. E. Finn, A. J. Sellen & S. B. Wilbur (Eds.), *Video-Mediated Communication* (pp. 107-131): LEA.

Olson, G. M., Herbsleb, J. D., & Rueter, H. H. (1994). Characterizing the sequential structure of interactive behaviors through statistical and grammatical techniques. *Human-Computer Interaction, 9*(3-4), 427-472.

Olson, G. M., & Olson, J. S. (2000). Distance Matters. *Human-Computer Interaction, 15*(2-3), 139-178.

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

Oviatt, S. L., DeAngeli, A., & Kuhn, K. (1997). Integration and Synchronization of Input Modes during Multimodal Human-Computer Interaction. In *Proceedings of ACM CHI 97 Conference on Human Factors in Computing Systems* (Vol. 1, pp. 415-422).

Pickering, M. J., & Garrod, S. C. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences, 27*, 169-226.

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

Prince, E. F. (1981). Toward a taxonomy of given-new information. In P. Cole (Ed.), *Radical Pragmatics* (pp. 223-255). New York, NY: Academic Press.

Roberts, C. (1996). Information structure in discourse: Towards an integrated formal theory of pragmatics. In J. H. Yvon & A. Kathol (Eds.), *OSU Working Papers in Linguistics* (Vol. 49).

Sanderson, P. M., & Fisher, C. (1993). Exploratory sequential data analysis in practice. In Proceedings of *ACM INTERCHI'93 - Adjunct Proceedings*, pp. 221.

Santorini, B. (1995). *Part-of-speech tagging guidfelines for the Penn Treebank Project* (3rd ed.).

Schafer, W., & Bowman, D. (2003). A comparison of traditional and fisheye radar view techniques for spatial collaboration. In Proceedings of *Graphics Interface (GI 2003)*, pp. 23-46.

Schober, M. F. (1993). Spatial perspective-taking in conversation. *Cognition, 47*, 1-24.

Schober, M. F. (1995). Speakers, addressees, and frames of reference: Whose effort is minimized in conversations about locations. *Discourse Processes, 20*, 219-247.

Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology, 21*(2), 211-232.

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

Short, J., Williams, E., & Christie, B. (1976). Visual communication and social interaction. In R. Baecker (Ed.), *Readings in groupware and computer-supported cooperative work* (pp. 153-164). San Francisco, CA: Morgan Kaufmann Publishers.

Shriberg, E., Stolcke, A., & Baron, D. (2001). Observations on overlap: Findings and implications for automatic processing of multi-party conversation. In Proceedings of *EUROSPEECH, '01*, pp. 1359-1362.

Snook, S. A. (2000). *Friendly fire: The accidental shootdown of U. S. Blackhawks over northern Iraq* (Vol. 280). Princeton, NJ: Princeton University Press.

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

Strube, M., & Hahn, U. (1999). Functional centering: Grounding referential coherence in information structure. *Computational Linguistics, 25*(3), 309-344.

Tanenhaus, M. K., & Spivey-Knowlton, M. J. (1996). Eye-tracking. *Language & Cognitive Processes, 11*(6), 583-588.

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

Tang, J. C. (1991). Findings from observational studies of collaborative work. *International Journal of Man-Machine Studies, 34*, 143-160.

Tang, J. C., & Isaacs, E. A. (1993). Why do users like video? Studies of multimedia-supported collaboration. *CSCW: An International Journal, 1*, 163-196.

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.

Tetreault, J. R., & Allen, J. (2004, July 19-21). Semantics, dialogue, and refrence resolution. In Proceedings of *The 8th Workshop on the Semantics and Pragmatics of Dialogue (CATALOG '04)*, pp. 131-137.

Vaghi, I., Greenhalgh, C., & Benford, S. (1999). Coping with inconsistency due to network delays in collaborative virtual environments. In Proceedings of *VRST '99*, pp. 42-49.

Veinott, E. S., Olson, J. S., Olson, G. M., & Fu, X. (1999). Video helps remote work: Speakers who need to negotiate common ground benefit from seeing each other. In Proceedings of *ACM Conference on Human Factors in Computing Systems (CHI '99)*, pp. 302-309. ACM Press.

Velichkovsky, B. (1995). Communicating attention: Gaze position transfer in cooperative problem solving. *Pragmatics & Cognition, 3*(2), 199-224.

Walker, M. A. (1989). Evaluating discourse processing algorithms. In Proceedings of *27th Annual Meeting of the Association for Computational Linguistics*, pp. 251-261.

Walker, M. A., Joshi, A. K., & Prince, E. F. (1998). *Centering theory in discourse*. Oxford: Clarendon Press.

Wang, H., & Chu, P. (1997). Voice source localization for automatic camera pointing system in videoconferencing. In Proceedings of *International conference on acoustics, speech and signal processing (ICASSP '97)*, pp. 187.

Weingart, L. (1997). How did they do that? The ways and means of studying group processes. *Research in Organizational Behavior, 19*, 40-89.

Whittaker, S. (2003). Things to talk about when talking about things. *Human-Computer Interaction, 18*(1-2), 149-170.

Whittaker, S., Geelhoed, E., & Robinson, E. (1993). Shared workspaces: How do they work and when are they useful? *International Journal of Man-Machine Studies, 39*, 813-842.

Whittaker, S., & O'Conaill, B. (1997). The role of vision in face-to-face and mediated communication. In K. E. Finn, A. J. Sellen & S. B. Wilbur (Eds.), *Video-Mediated Communication*: LEA.

Williams, E. (1977). Experimental comparisons of face-to-face and mediated communication: A review. *Psychological Bulletin, 84*(5), 963-976.

Yule, G. (1996). *The study of language* (2nd ed.). Cambridge: Cambridge University Press.

Appendix A: Puzzle Study Coding Manual

This appendix presents the original coding manual used for the communication content analyses performed in the first puzzle study and augmented in subsequent studies. It is a multi-level coding scheme that aims to capture elements of the task, the utterances and language used, and the actions and movements performed in the shared visual environment.

Coding notes

Pieces are numbered from 1 to 8 by their original location on the staging area, starting at the top left and working across each row.

Task status

This level of the coding captures basic task information. For example, when a trial begins and ends, piece movements within a trial, and the beginning and ending of referents to pieces in the shared workspace.

Code	Description and Example
BEGIN TRIAL	Mark the beginning of each trial.
END TRIAL	Mark the end of each trial.
BEGIN PIECE #	When the participants begin work on each new piece, where # is the number of the piece. The beginning of a new piece is defined as the first referential statement to that piece.
END PIECE #	When participants end work on each piece, where # is the number of the piece. Work ends on a piece when the participants stop discussing or moving the piece actively.
BEGIN REFERENT	Add the tag before the first referential statement for each piece. (i.e., < x, R, x >)
END REFERENT	Add the tag after the piece is moved, or if that does not occur, when the discussion moves on to the next piece (usually after a move behavior).
BEGIN POSITION	Add the tag before the first position statement of each piece. (i.e., < x, P, x >)
END POSITION	Add the tag after the piece has been positioned, or if that does not occur, when

	the discussion moves on to the next piece (usually after a position behavior).
BEGIN CONTEXT	Add the tag at the beginning of each context-gathering exchange of greater than two utterance.
END CONTEXT	Add the tag at the end of each context-gathering exchange of greater than two utterances.
TIMEOUT	Add the tag whenever a timeout or abnormal disruption occurs.

Utterances

The following present the different forms of utterances that could be coded. Each utterance is coded as a tuple, where each has a <*Form, Utterance Type*, and *Deixis*> component. Each of these are described in turn.

Form

Code	Description and Example
Q (Question)	Used for questions, including rhetorical questions, For example: "The red one?"
A (Answer)	Used for answers or responses to questions. For example: Helper: How many reds do you have? Worker: Uh, three. < A, CR, N >
K (acKnowledgement)	Used for acknowledgements. Utterances comprised mainly of phrases such as "mmm hmm," "okay," "yeah," and/or repetition of the previous speaker's words. For example: Helper: Put it in the right hand corner. Worker: Yeah, the right hand corner, OK. < K, AU, S >
S (Statement)	Used for statements, typically all other utterances. For example: "Take the red one."
?? (Unknown)	Used for utterances that are completely or partially inaudible or incomprehensible and that thus can not be coded. For example: (code (??,??,??) if completely unusable, otherwise only those categories that make sense.

If an utterance can be interpreted as multiple forms, they use the order of precedence shown above (Question, Answer, acKnowledgement, Statement).

Utterance types

Code	Description and Example
R (References)	Code R for references to and attempts to describe a specific piece. Note that descriptions of several pieces are coded as CR (task context—referent). For example: "Take the red one."
P (Positions)	Code P for attempts to describe the position of a specific piece, even if its position is described relative to one or more already placed pieces. Note that the position of several pieces together is coded CP (task context—position). For example: "Put that one touching the upper right corner of the blue one."
AU (Acknowledgements of Understanding))	Code AU for acknowledgements of understanding. Thus, "mmm hmm," "okay," and similar statements are only coded AU when they directly follow a statement

	or acKnowledgement. Note that those phrases when used after behaviors should instead be coded as acknowledgments of behavior (code AB). For example: Helper: Take the red one. Worker: Mmm hmm. < K, AU, N >
AB (Acknowledgements of Behavior)	Code AB for acknowledgements of behavior. Thus, "mmm hmm," "okay," and similar statements are only coded AB when they directly follow a behavior. Note that those phrases when used after a statement or acKnowledgement should instead be coded as acknowledgments of understanding (code AU). For example: Helper: Take the red one. Worker moves red piece. Worker: Mmm hmm. < K, AB, N >
CR (Task Context – Referential))	Code CR (task context—referent) for referential contextual information about the task not already covered by other categories. This includes utterances regarding what colors are available, etc. For example: "What colors do you have?" "I have three reds."
CP (Task Context – Positional)	Code CP (task context—position) for position-related contextual information about the task not already covered by other categories. This includes descriptions of the position of several pieces together (patterns), etc. For example: "The last three blocks should form, like, a diagonal line."
C? (Task Context – Other)	Code C? (task context—other) for contextual information about the task not already covered by CR, CP, or any other categories. This includes the degree of shared visual space, etc. For example: "There's a delay."
IP (Internal Perceptions)	Code IP for utterances relating to participants' internal perceptions. For example: "I can't see anything." "Did you hear that?"
IT (Internal Thoughts)	Code IT for utterances relating to participants' non-perceptual internal thoughts and beliefs regarding the task, including most statements using the verb "to think," etc. For example: "I thought you could see my moves."
ES (Explicit Strategies)	Code ES for utterances relating to explicit strategies. Do not code ES for anything other than explicit strategies for completing the task. For example: "Let's do this to establish what colors are available…"
F (Fragment)	Code F for speech fragments, when the speaker either is cut off or does not finish their utterance, making it impossible to classify it otherwise. For example: "What about—" "It's probably too—"
*** (Other)*	Code ** (other) for utterances that are not related to the experiment or simply can not be put into any of the other categories. For example: "This is so boring." "I'm sorry."
?? (Unknown)	Code ?? (uncodable) for utterances that are completely or partially inaudible or

	incomprehensible and that thus can not be coded. For example: code < ??, ??, ?? > if completely unusable, otherwise only those categories that make sense.

Deixis

If deixis is present, include the appropriate letters, as described below, in the order **THS**. For example, an utterance with this/that and spatial deixis would be coded **HS**, one with temporal and spatial deixis would be **TS**, and one with all three deixis would be **THS**. Otherwise, code **N** (no deixis).

Code	Description and Example
T (Temporal)	Code T for utterances that use temporal deixis (now, then, changing, etc.) For example: "Now it's pink." "It's changing from purple to red to green."
H (Deictic pronouns)	Code H for utterances that use the deictic terms "this," "that," "there," or other related terms. For example: "Move that one up half a square." "This one is wrong." "Put the red one over there."
S (Spatial Deictics / Locatives)	Code S for utterances that use spatial deixis other than "this," "that," etc, such as "above," "below," "in front of," "on top of," "next to," "behind," "right," "left," "up," "down," "touching," etc. For example: "Put it in the upper right-hand corner." "It's on top of the upper half of the red one." "Place it next to the blue one."
?? (Unknown)	Code ?? (uncodable) for utterances that are completely or partially inaudible or incomprehensible and that thus can not be coded. For example: code < ??, ??, ?? > if completely unusable, otherwise only those categories that make sense.

Behaviors

The final level of the coding scheme captures information about the piece movements in the workspace and their quality.

Actions

Code	Description and Example
M (Moves)	Code M for moves. A move occurs when a piece is moved from the staging area to be eventually placed on the work area. A move begins at the frame before the movement is visible. It ends at the first frame where a piece is both fully in the work area AND stopped or moving slowly.
R (Removes)	Code R for removes. A remove occurs when a piece is moved from the work area back onto the staging area. A remove begins at the frame before the movement is visible. It ends at the first frame where a piece is both fully in the staging area AND stopped or moving slowly.
P (Positioning)	Code P for positioning. Positioning occurs when a piece is moved from an arbitrary position or one where it had previously been positioned to a new position where it is explicitly released and left for at least a short period of time. Positioning begins when a piece moves from stop or slow movement (often directly after the move). It ends when the piece is in its final position.
S (Showing)	Code S for showing. Showing occurs when the worker temporarily moves pieces from the staging area into the work area and quickly removes them without the intent of integrating them into their solution. It begins at the frame before the movement is visible, and ends when the piece has been replaced on the staging area.

Accuracy

Code	Description and Example
C (Correct)	Code C (correct) in these cases: o For moves and removes, code C when the piece being moved/removed is the one they have most recently been directed to use by the helper AND matches the helper's solution. o For positioning, code C when the piece being moved (regardless of its correctness) is put in the correct position relative to the pieces previously placed as per the helper's most recent instructions AND the helper's solution. Thus, if the previous positioning was incorrect, the current positioning's correctness is based solely on its position relative to the other pieces (including the incorrect one), even if the overall pattern is now incorrect. o For showing, code C if the piece would have been correct to move.
I (Incorrect)	Code I (incorrect) in these cases: o For moves and removes, code I when the piece being moved/removed is not the one they have most recently been directed to use by the helper OR does not match the helper's solution. o For positioning, code I when the piece being moved (regardless of its correctness) is put in the incorrect position relative to the pieces previously placed as per the helper's most recent instructions AND the helper's solution. o For showing, code I if the piece would have been incorrect to move.

N (Not applicable)	Code N (not applicable) when no information has been given to the worker, and they thus cannot make any judgments about accuracy. Also code N for positioning when a set of pieces are repositioned simply to make room for an additional piece, without any change to their relative positions to one another.

Other notes

Note that not all piece movement is coded. If a piece is moved around the staging area but not moved into the work area, it is not coded. If a piece is moved only slightly from one arbitrary location to another on the work area, it is not coded either.

Divide an utterance into multiple utterances if it consists of more than one sentence or idea, has relatively long breaks of silence, or contains more than one utterance type.

When it is apparent that a subject is using a word or phrase which would normally be divided from the rest of an utterance and coded separately (i.e. OK) as a habit of speech, group that utterance with the larger utterance and code it as such (i.e. OK, now take the reddish blue one < *S*, *R*, *N* >).

If a portion (or all) of an utterance is inaudible or incomprehensible, use two question marks ("??") as a substitute for that portion in the Utterance field.

Appendix B: The Basic Centering Algorithm

The following presents an overview of Centering Theory (Grosz et al., 1983) and the original centering algorithm as proposed by Brennan and colleagues (Brennan et al., 1987).

Centering theory

Centering Theory is a framework composed of a system of rules and constraints that interact with semantic restrictions and world knowledge and make use of data structures to capture the local attentional focus of a discourse. These elements come together to govern the relationships between the discourse content and the surface forms of the utterances generated by the conversational participants.

The basic centering model

The Centering model describes discourse, or in the case of the PUZZLE CORPUS, the shared visible actions and spoken language that constitute a collaborative activity. Such a discourse can be broken out into component segments that serve as the base units for the discourse model. For the purpose of this document, we will describe each discourse segment (Grosz & Sidner, 1986) as consisting of a sequence of utterances, $U_1,\ldots,U_m$, even though in practice these units will be used to capture both physical actions and spoken elements. The following summarizes the basic centering model as described in (Grosz et al., 1983) and as refined in (Brennan et al., 1987; Grosz et al., 1995; Walker *et al.*, 1998).

The notion of centers

A "center" is a semantic entity that captures the notion of the current state of focus (i.e., the "topic") of a given utterance, taking into account existing context, prior spoken discourse, etc. In order to capture the notion of shared attentional topic, each utterance, U_n is associated with a

ranked list of potential *forward-looking centers* (*CF*), a *preferred center* (*CP*), and a *backward-looking center* (*CB*). Together, these elements (described in detail below) provide mechanisms for predicting the preferred interpretation of the current topic of the discourse (with the *CP*) and for looking back to the previous discourse (with the *CB*) to determine the fluidity with which the discourse is proceeding.

The list of forward-looking centers for a given utterance, $CF(U_n)$, is a ranked-list of partially-ordered discourse entities that are realized by the linguistic expressions in the utterance. The ranking of a given entity in this list is based on grammatical function and roughly corresponds to the likelihood that the entity will be chosen as the focal center of the following utterance, U_{n+1}. This ranked list provides an indication of the relative salience of the local discourse entities. The most highly ranked entity in this list is referred to as the preferred center, $CP(U_n)$, and it is the most likely candidate to be the focal center of the next utterance. The backward-looking center, $CB(U_n)$, captures the actual discourse entity that the current utterance is about. In many formulations, this entity must be realized in the immediately preceding utterance, U_{n-1}, although this is one of the parametric instantiations that often challenged (for details see Poesio et al., 2004).

Constraints and rules

In addition to these basic structures for describing discourse centers, the centering model includes a set of rules and constraints based on psycholinguistic accounts of language generation and comprehension (these constraints and rules are drawn from Brennan et al., 1987):

Constraints,

- There is precisely one backward looking center, *CB*.
- Every element of $CF(U_n)$ must be realized in U_n.
- $CB(U_n)$ is the highest-ranked element of $CF(U_{n-1})$ that is realized in U_n.

Rules (adapted from Brennan et al., 1987; Grosz et al., 1995),

- **Rule 1**: If any element of $CF(U_n)$ is realized by a pronoun in U_{n+1} then the $CB(U_{n+1})$ must be realized by a pronoun also.
- **Rule 2**: Transition states (as defined below and shown in Table B.1 and Table B.2) are ordered such that sequences of continuations are preferred over sequences of retaining;

sequences of retaining are preferred to sequences of smooth-shifts; and sequences of smooth-shifts are preferred to sequences of rough-shifts.

The coherence of transitions

The coherence of a discourse segment is captured by the transition relation between the prior utterance's preferred center, $CP(U_{n-1})$, and the backward-looking center of the current utterance, $CB(U_n)$, as well as the relation between the current preferred center, $CP(U_n)$, and the current backward-looking center, $CB(U_n)$. This notion of discourse coherence is captured through a typology of transitions defined in Table B.1.

When a speaker has been talking about a particular entity and intends to continue talking about that same entity, a CONTINUE transition should occur. This is captured by the fact that the $CB(U_n)$ is the same as it was in the prior utterance and that the entity is also the highest ranked entity in the current set of forward-looking centers (i.e., it is the $CP(U_n)$ as well).

Table B.1. The four main transition definitions used to capture discourse coherence.

	$CB(U_n) = CB(U_{n-1})$ or $CB(U_{n-1}) = NIL$	$CB(U_n) \neq CB(U_{n-1})$
$CB(U_n) = CP(U_n)$	CONTINUE	SMOOTH-SHIFT
$CB(U_n) \neq CP(U_n)$	RETAIN	ROUGH-SHIFT

If however, the $CB(U_n)$ is the same as it was in the prior utterance but the entity is not the highest ranked entity in the current set of forward-looking centers (i.e., it is not the $CP(U_n)$ as well), then a transition type of RETAIN is said to occur. Grosz and Sidner (1986) propose that this transition occurs in the situation when a speaker intends to shift the conversational topic to new entity and they signal this by demoting the current center in the ranked list of discourse entities.

Finally, if the relation between the $CB(U_n)$ and the $CB(U_{n-1})$ no longer holds, we enter one of the shift states indicated in the right half of Table B.1. These transitions exist when the backward-looking center is not retained in any way between subsequent utterances. Brennan and colleagues (Brennan et al., 1987) identified a finer distinction in the case of shifts and proposed two types. In the case where the $CB(U_n)$ is still the highest ranked entity, $CP(U_n)$, we have what is referred to as a SMOOTH-SHIFT. However, when the $CB(U_n)$ is not the highest ranked entity, we have a ROUGH-SHIFT transition.

Together, these transition types describe the relative smoothness with which the discourse proceeds. As exemplified in Table B.2, the CONTINUE transition is the smoothest (or most coherent) transition, while the ROUGH-SHIFT is the least coherent transition.

Table B.2. Transition rankings.

CONTINUE $\prec$ RETAIN $\prec$ SMOOTH-SHIFT $\prec$ ROUGH-SHIFT

Consider a speaker that has a number of things to talk about. The most structured and coherent way for her to present the information would be to provide all the information needed about a given entity before introducing and shifting to new topics. For example, in our PUZZLE CORPUS it would be considered much more conversationally coherent for the Helper to first describe a piece, then describe its location in the workspace, and use this `<piece, position>` cycle for the remaining pieces, than it would be for the Helper to first describe a piece, then describe the second piece and eventually switch to a placement strategy in a `<piece, piece,…,piece>` `<placement, placement,…,placement>` strategy where the entity of focus changes back and forth more frequently.

The centering algorithm

Given the previously described rules and constraints, Brennan and colleagues proposed the following pronoun binding algorithm. It is presented here in its original form as originally described by Brennan and colleagues (1987) to serve as a basis for the following discussions.

There are four major stages to the Centering algorithm: **(1) Construct**, **(2) Filter**, **(3) Classify** and **(4) Select**. During the Construct stage all of the possible elements for anaphoric reference are identified. Then, the pronouns are mapped to the discourse entities, maintaining any agreement features. During the second stage, Filter, the possible mappings are discarded based on the aforementioned constraints and rules. The Classify stage classifies each of the possible transitions according to one of the four transition types presented in Table B.1. Finally, the Select stage chooses the best possibility among the classified types using the preference rankings presented in Table B.1.

(1) Construct:

This stage primarily deals with the construction of the potential referring expressions and anaphoric candidates.

1. Create set of referring expressions (REs).
2. Order REs by grammatical function.
3. Create a set of possible *CF*-lists. Expand each element of (2) according to the whether it is a pronoun, description, or proper noun.
    a. *Pronouns* expand into a set with entries for each RE in the preceding *CF* list (i.e., $CF(U_{n-1})$) that matches the following:
        i. Its agreement features.
        ii. The selectional constraints projected by the verb.
        iii. The contraindexing constraints of other elements in the current *CF* list being expanded.
    b. *Descriptions* are not expanded; rather they are represented by their intention and an index.
    c. *Proper nouns* expand into a set with an entry for each discourse entity it could realize.
4. Create a list of potential backward looking centers (i.e., the *CB*). This is the list of all the entities in $CF(U_{n-1})$ plus the additional entry of NIL.
5. Generate the proposed referential anchors using the cross-product of steps (3) and (4).

(2) Filter:

Possibilities are discarded unless all of the following criteria are met (the following are directly taken from the description in Brennan et al., 1987):

1. Filter by what are referred to as contra-indices. These are the cases when the same antecedent exists for two pronouns or there is an antecedent proposed for a prior existing pronoun with which it is contra-indexed. These selections are removed from consideration.
2. The $CF(U_{n-1})$ list is traversed and the objects kept are those that exist in the *CF* list of the anchor. If the proposed *CB* is not the first element of this list then the given anchor is eliminated. This provides a guarantee that the *CB* will be the highest ranked element of the $CF(U_{n-1})$ in the current utterance.
3. If the proposed *CB* does not match any of the entities realized in the proposed *CF* list then this anchor is eliminated. This provides a guarantee that if any element is realized as a pronoun then the *CB* is realized as a pronoun.

(3) Classify:

Classify each potential RE anchor in the list using the transitions previously described. Use U_{n-1} as the previous utterance and U_n as the current utterance and examine the potential transitions between them.

(4) Select:

Each of the possible transitions from is ranked according to the transition rankings in Table B.2. Then $CB(U_n)$ is set to the proposed *CB* and $CF(U_n)$ is set to the proposed *CF* of the highest ranked anchor.

A worked example using centering

This section presents a walkthrough of the algorithm on the following utterance from an excerpt in the PUZZLE CORPUS:

(B.1)	Helper:	There's like a red one.
	Helper:	That touches the bottom left corner of the blue one.
	Helper:	OK and there's like a brown one.

Example 5.1: That touches the bottom left corner of the blue one [excerpt (B.1), Utt_2]

```
Step.Construct.1    ([that][bottom left corner][blue one])
Step.Construct.2    ([that][bottom left corner][blue one])
Step.Construct.3    ([RED_ONE:that][bottom left corner][blue one])
Step.Construct.4    ([RED_ONE], [NIL])
Step.Construct.5    <[RED_ONE], ([RED_ONE:that][bottom left corner][blue one])>
                    <NIL, ([RED_ONE:that][bottom left corner][blue one])>

Step.Filter.1       <[RED_ONE], ([RED_ONE:that][bottom left corner][blue one])>
                    <NIL, ([RED_ONE:that][bottom left corner][blue one])>
Step.Filter.2       <[RED_ONE], ([RED_ONE:that][bottom left corner][blue one])>
Step.Filter.3       <[RED_ONE], ([RED_ONE:that][bottom left corner][blue one])>

Step.Classify.1     <[RED_ONE], ([RED_ONE:that][bottom left corner][blue one])>
                    Transition Type: CONTINUE

Step.Select.1       CB(Utt2) = RED_ONE
                    CF(Utt2) = ([RED_ONE:that][bottom left corner][blue one])
                    (trivial)
```

Appendix C: The Left-Right Centering Algorithm

The following is a replication of the LRC algorithm originally described in (Tetreault, 2001). This description piggybacks on some of the formal notation and structures detailed in the previous appendix. Tetreault's LRC algorithm was developed in response to the BFP algorithms lack of incremental processing. It is an incremental resolution algorithm that follows the centering constraints. A major difference between the LRC algorithm and the BFP is the LRC's ability to first search intrasententially, and if a referent is not found for the referring expression, it can then search intersententially. In other words, the algorithm begins by first looking for a possible antecedent in the current utterance. If one is not found, the algorithm then begins to search the previous the utterance's *CF*-list in a left-to-right fashion for an antecedent.

The following is reproduced from the original description of the LRC algorithm presented in (Tetreault, 2001):

1. **Preprocessing:** From the previous utterance, $CB(U_{n-1})$ and $CF(U_{n-1})$ are available.
2. **Process the utterance:** Parse and extract incrementally from U_n all references to discourse entities. For each pronoun do:
    a. Search for an antecedent intrasententially in $CF\text{-}partial(U_n)$ that meets feature and binding constraints. If one is found, proceed to the next pronoun within the utterance. Else go to (b).
    b. Search for an antecedent intersententially in $CF(U_{n-1})$ that meets feature and binding constraints.
3. **Create CF:** Create the *CF*-list of U_n by ranking discourse entities of U_n according to grammatical function. Tetreault's original implementation used a left-to-right breadth-first walk of the parse tree to approximate sorting by grammatical function.

Appendix D: Penn Treebank II POS Tags

The following describes the part of speech and bracketing conventions used with Penn Treebank II style tags. This information is originally published in Santorini's style guide (Santorini, 1995) and the bracketing style manual. Both are available at: http://www.cis.upenn.edu/~treebank.

Part of speech tags

CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential *there*
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PRP$	Possessive pronoun
RB	Adverb

RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	*to*
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun
WRB	Wh-adverb

Phrase level tags

ADJP	Adjective Phrase.
ADVP	Adverb Phrase.
CONJP	Conjunction Phrase.
FRAG	Fragment.
INTJ	Interjection. Corresponds approximately to the part-of-speech tag UH.
LST	List marker. Includes surrounding punctuation.
NAC	Not a Constituent; used to show the scope of certain prenominal modifiers within an NP.
NP	Noun Phrase.
NX	Used within certain complex NPs to mark the head of the NP. Corresponds very roughly to N-bar level but used quite differently.
PP	Prepositional Phrase.
PRN	Parenthetical.
PRT	Particle. Category for words that should be tagged RP.
QP	Quantifier Phrase (i.e. complex measure/amount phrase); used within NP.
RRC	Reduced Relative Clause.
UCP	Unlike Coordinated Phrase.
VP	Verb Phrase.
WHADJP	Wh-adjective Phrase. Adjectival phrase containing a wh-adverb, as in how hot.
WHAVP	Wh-adverb Phrase. Introduces a clause with an NP gap. May be null (containing the 0 complementizer) or lexical, containing a wh-adverb such as how or why.
WHNP	Wh-noun Phrase. Introduces a clause with an NP gap. May be null (containing the 0 complementizer) or lexical, containing some wh-word (e.g., who, which book, whose daughter, none of which, or how many leopards).
WHPP	Wh-prepositional Phrase. Prepositional phrase containing a wh-noun phrase (such as of which or by whose authority) that either introduces a PP gap or is contained by a WHNP.
X	Unknown, uncertain, or unbracketable. X is often used for bracketing typos and in bracketing the...the-constructions.

Clause level tags

S	Simple declarative clause, i.e. one that is not introduced by a (possible empty) subordinating conjunction or a wh-word and that does not exhibit subject-verb inversion.
SBAR	Clause introduced by a (possibly empty) subordinating conjunction.
SBARQ	Direct question introduced by a wh-word or a wh-phrase. Indirect questions and relative clauses should be bracketed as SBAR, not SBARQ.
SINV	Inverted declarative sentence, i.e. one in which the subject follows the tensed verb or modal.
SQ	Inverted yes/no question, or main clause of a wh-question, following the wh-phrase in SBARQ.

Appendix E: Raw Data Log of Visual Information

File Header Information *Contains basic information about trial and block number, start time, set of stimulus blocks and positions, and settings of the shared visual space*	"COMM;Subject ID: 2" "COMM;Subject SVS_CDCC_04-07-21-1500_2_W" "SEND;P941140000011000" "BLOCK; 0" "TRIAL; 1" "BLOCK-COLOR; 4" "BLOCK-DELAY; 0ms" "BLOCK-PUZZLESET; 1" "BLOCK-SHAREDSPACE; True" "BLOCK-COLORCYCLE; 0" "VIEWPORT; full" "KEY-BLOCK; 0"
Key Block Information *Contains X,Y coordinates and movements of the key block (0). These coordinates are then used to calculate the relative positioning of the other blocks in the puzzle to facilitate scoring.*	"MOVE; 6020ms block(0) 5265,4200" "MOVE; 6040ms block(0) 4890,4200" "MOVE; 6080ms block(0) 4515,4200" "MOVE; 6100ms block(0) 4140,4200" "MOVE; 6120ms block(0) 3390,4200" "MOVE; 6140ms block(0) 3015,4200" "MOVE; 6160ms block(0) 2640,4200" "MOVE; 6180ms block(0) 2265,4200"
Block Information *Contains absolute X,Y coordinates and relative X,Y coordinates to the key block.*	"MOVE; 8320ms block(4) 7515,3075 5250,-750" "MOVE; 8360ms block(4) 7140,3075 4875,-750" "MOVE; 8420ms block(4) 6765,3075 4500,-750" "MOVE; 10320ms block(5) 9390,3075 7125,-750" "MOVE; 12700ms block(5) 9015,3075 6750,-750" "MOVE; 12740ms block(5) 8265,3075 6000,-750" "MOVE; 12780ms block(5) 7140,3075 4875,-750" "MOVE; 12820ms block(5) 6015,3075 3750,-750" "MOVE; 12840ms block(5) 4890,3075 2625,-750" "MOVE; 14480ms block(4) 6015,2700 3750,-1125" "MOVE; 14500ms block(4) 5265,2325 3000,-1500" "MOVE; 14560ms block(4) 4515,1575 2250,-2250" "MOVE; 14580ms block(4) 4140,1575 1875,-2250" . . .
Footer Information *Closeout information for the trial.*	"TRIAL-TIME;73300ms" "TRIAL-END;16735093" "COMM;7/21/2004 3:15:13 PM"

Appendix F: Additional Statistical Details

Included in this appendix are the additional statistical details for data presented in the various chapters. Each major section contains additional statistics for a given chapter.

Chapter 3 appended statistical details

Task performance model, analysis of variance table.

			Stable			Drift		
			Immediate	**Delayed**	**None**	**Immediate**	**Delayed**	**None**
Row	**Dependent Variable**	**N**	*Mean (SE)*	*Mean (SE)*	*Mean (SE)*	*Mean (SE)*	*Mean (SE)*	*Mean (SE)*
A	Completion Time (seconds)	288	48.5 [a] (5.87)	58.2 [a] (6.20)	56.4 [a] (6.11)	56.1 [a] (5.89)	81.0 [b] (6.15)	97.0 [c] (6.11)

Speaker communication models, analysis of variance tables

			Helper			Worker		
			Immediate	**Delayed**	**None**	**Immediate**	**Delayed**	**None**
Row	**Dependent Variable**	**N**	*Mean (SE)*	*Mean (SE)*	*Mean (SE)*	*Mean (SE)*	*Mean (SE)*	*Mean (SE)*
B	Word Production Rate (nLog)	144	4.35 [a] (.176)	4.47 [a] (.182)	4.51 [a] (.186)	1.58 [b] (.176)	2.33 [c] (.182)	3.10 [d] (.186)
C	Number of Acknowledgements of Behavior	140	1.79 [a] (.363)	1.59 [a] (.359)	-0.75 [b] (.393)	1.58 [a] (.362)	2.3 [a,c] (.392)	3.05 [c] (.338)
D	Number of Acknowledgements of Understanding	140	0.669 [a] (.461)	0.454 [a] (.458)	1.1 [a,b] (.503)	1.92 [a,b] (.462)	2.57 [b] (.500)	5.09 [c] (.430)
E	Number of Deictic Pronouns	140	2.07 [a] (.351)	1.37 [a,b] (.348)	0.95 [b,c] (.382)	0.92 [b] (.351)	0.65 [b,c] (.380)	0.08 [c] (.327)
F	Number of Spatial Deictics	140	4.48 [a] (.459)	5.91 [b] (.452)	5.56 [b] (.491)	1.17 [c] (.455)	1.37 [c] (.494)	1.26 [c] (.430)

Row A: Task performance effect tests

A. Performance Model

Terms	DF_{Num}	DF_{Den}	F-Ratio	p-value	sig.
Block	1	258	0.78	0.379	
Puzzle Difficulty	1	258	5.75	0.017	**
Color Drift	1	10	10.19	0.010	**
Shared Visual Information	2	258	24.15	<.001	***
Block × Shared Visual Information	2	258	3.13	0.045	**
Block × Puzzle Difficulty	1	258	0.09	0.765	
Block × Color Drift	1	258	0.46	0.496	
Puzzle Difficulty × Color Drift	1	258	3.15	0.077	*
Shared Visual Information × Color Drift	2	258	11.41	<.001	***
Shared Visual Information × Puzzle Difficulty	2	258	1.01	0.367	
Block × Puzzle Difficulty × Color Drift	1	258	0.00	0.998	
Block × Puzzle Difficulty × Shared Visual Information	2	258	1.29	0.278	
Block × Color Drift × Shared Visual Information	2	258	0.91	0.406	

Speaker communication models effect tests

B. Word Rate Model

Terms	DF_{Num}	DF_{Den}	F-Ratio	p-value	sig.
Block	1	110	2.70	0.103	
Time (log minutes)	1	110	42.27	<.001	***
Speaker Role	1	110	292.95	<.001	***
Puzzle Difficulty	1	110	1.24	0.268	
Color Drift	1	12	0.54	0.475	
Shared Visual Information	2	110	11.45	<.001	***
Block × Puzzle Difficulty	1	110	0.47	0.495	
Block × Color Drift	1	110	5.12	0.026	**
Puzzle Difficulty × Color Drift	1	110	0.03	0.855	
Block × Speaker Role	1	110	0.92	0.340	
Puzzle Difficulty × Speaker Role	1	110	3.23	0.075	*
Color Drift × Speaker Role	1	110	1.08	0.301	
Shared Visual Information × Speaker Role	2	110	10.81	<.001	***
Shared Visual Information × Color Drift	2	110	3.80	0.025	**
Shared Visual Information × Puzzle Difficulty	2	110	0.18	0.832	
Block × Shared Visual Information	2	110	0.84	0.436	
Block × Shared Visual Information × Speaker Role	2	110	10.66	<.001	***

C. Acknowledgements of Behavior

Terms	DF_{Num}	DF_{Den}	F-Ratio	p-value	sig.
Block	1	105	1.10	0.297	
Time (log minutes)	1	105	0.04	0.834	
Speaker Role	1	105	8.68	0.004	***
Puzzle Difficulty	1	105	0.00	0.976	
Color Drift	1	10	0.86	0.377	
Shared Visual Information	2	105	3.60	0.031	**
Block × Puzzle Difficulty	1	105	0.09	0.764	
Block × Color Drift	1	105	4.60	0.034	**
Puzzle Difficulty × Color Drift	1	105	0.27	0.604	
Block × Speaker Role	1	105	0.00	0.945	
Puzzle Difficulty × Speaker Role	1	105	1.83	0.179	
Color Drift × Speaker Role	1	105	0.02	0.898	
Shared Visual Information × Speaker Role	2	105	33.56	<.001	***
Shared Visual Information × Color Drift	2	105	3.41	0.037	**
Shared Visual Information × Puzzle Difficulty	2	105	0.28	0.759	
Block × Shared Visual Information	2	105	2.16	0.120	
Block × Shared Visual Information × Speaker Role	2	105	1.05	0.355	
Words	1	105	7.79	0.006	***

D. Acknowledgements of Understanding

Terms	DF_{Num}	DF_{Den}	F-Ratio	p-value	sig.
Block	1	105	1.02	0.315	
Time (log minutes)	1	105	17.41	<.001	***
Speaker Role	1	105	15.48	<.001	***
Puzzle Difficulty	1	105	0.34	0.559	
Color Drift	1	10	0.58	0.464	
Shared Visual Information	2	105	12.43	<.001	***
Block × Puzzle Difficulty	1	105	1.87	0.174	
Block × Color Drift	1	105	0.40	0.527	
Puzzle Difficulty × Color Drift	1	105	0.17	0.683	
Block × Speaker Role	1	105	0.07	0.790	
Puzzle Difficulty × Speaker Role	1	105	0.41	0.524	
Color Drift × Speaker Role	1	105	6.21	0.014	**
Shared Visual Information × Speaker Role	2	105	8.66	<.001	***
Shared Visual Information × Color Drift	2	105	5.30	0.006	***
Shared Visual Information × Puzzle Difficulty	2	105	0.83	0.437	
Block × Shared Visual Information	2	105	0.55	0.579	
Block × Shared Visual Information × Speaker Role	2	105	4.53	0.013	**
Words	1	105	0.17	0.679	

E. Deictic Pronouns

Terms	DF_{Num}	DF_{Den}	F-Ratio	p-value	sig.
Block	1	105	1.10	0.297	
Time (log minutes)	1	105	2.63	0.108	
Speaker Role	1	105	3.75	0.055	*
Puzzle Difficulty	1	105	0.27	0.602	
Color Drift	1	10	0.07	0.797	
Shared Visual Information	2	105	5.47	0.006	***
Block × Puzzle Difficulty	1	105	2.86	0.094	*
Block × Color Drift	1	105	4.17	0.044	**
Puzzle Difficulty × Color Drift	1	105	0.11	0.745	
Block × Speaker Role	1	105	0.00	0.985	
Puzzle Difficulty × Speaker Role	1	105	0.01	0.917	
Color Drift × Speaker Role	1	105	0.36	0.549	
Shared Visual Information × Speaker Role	2	105	0.37	0.692	
Shared Visual Information × Color Drift	2	105	0.56	0.570	
Shared Visual Information × Puzzle Difficulty	2	105	0.91	0.407	
Block × Shared Visual Information	2	105	7.20	0.001	***
Block × Shared Visual Information × Speaker Role	2	105	0.26	0.769	
Words	1	105	0.45	0.502	

F. Spatial Deixis

Terms	DF_{Num}	DF_{Den}	F-Ratio	p-value	sig.
Block	1	105	0.13	0.723	
Time (log minutes)	1	105	0.01	0.938	
Speaker Role	1	105	45.85	<.001	***
Puzzle Difficulty	1	105	2.15	0.146	
Color Drift	1	10	0.24	0.634	
Shared Visual Information	2	105	2.66	0.074	*
Block × Puzzle Difficulty	1	105	0.00	0.966	
Block × Color Drift	1	105	0.13	0.721	
Puzzle Difficulty × Color Drift	1	105	0.38	0.538	
Block × Speaker Role	1	105	1.27	0.263	
Puzzle Difficulty × Speaker Role	1	105	5.11	0.026	**
Color Drift × Speaker Role	1	105	2.92	0.091	*
Shared Visual Information × Speaker Role	2	105	2.15	0.122	
Shared Visual Information × Color Drift	2	105	3.21	0.045	**
Shared Visual Information × Puzzle Difficulty	2	105	3.65	0.029	**
Block × Shared Visual Information	2	105	0.67	0.513	
Block × Shared Visual Information × Speaker Role	2	105	0.18	0.837	
Words	1	105	8.91	0.004	***

Chapter 4 appended statistical details

The following tables present the results from the piecewise linear regression models for Study 2 and Study 3.

Study 2: Full Model Results

Terms	Estimate	Std. Error	DF_{Num}	DF_{Den}	F-Ratio	Lower CI (95%)	Upper CI (95%)	p-value	Sig.
Visual Delay	0.48	2.87	1	610	0.03	-5.16	6.12	0.867	
Visual Delay - 939	22.83	6.21	1	610	13.53	10.64	35.01	<.001	***
Visual Delay - 1798	-28.45	7.30	1	610	15.21	-42.78	-14.12	<.001	***
Linguistic Complexity	-11880.69	722.67	1	610	270.27	-13299.88	-10461.50	<.001	***
Block	-965.17	261.32	1	610	13.64	-1478.34	-452.00	<.001	***
Trial	-3548.27	583.17	1	610	37.02	-4693.54	-2403.01	<.001	***
Visual Delay × Linguistic Complexity	2.42	2.86	1	610	0.71	-3.21	8.04	0.399	
Visual Delay - 939 × Linguistic Complexity	8.23	6.25	1	610	1.73	-4.05	20.51	0.188	
Visual Delay - 1798 × Linguistic Complexity	-23.43	7.26	1	610	10.41	-37.69	-9.17	0.001	***
Block × Linguistic Complexity	85.78	270.05	1	610	0.10	-444.53	616.09	0.751	
Trial × Linguistic Complexity	2276.05	583.17	1	610	15.23	1130.79	3421.31	<.001	***

Study 3: Models for Various Cycle Rates

Terms	Estimate	Std. Error	DF_{Num}	DF_{Den}	F-Ratio	Lower CI (95%)	Upper CI (95%)	p-value	Sig.
Moderate Change Rate									
Delay Rate	-2.16	13.91	1	265	0.02	-29.55	25.23	0.877	
Delay Rate – 431*ms*	147.49	51.32	1	265	8.26	46.49	248.50	0.004	***
Delay Rate – 558*ms*	-141.14	41.10	1	265	11.79	-222.03	-60.26	0.001	***
Block	-320.80	448.29	1	265	0.51	-1203.20	561.60	0.475	
Trial	-2404.84	834.37	1	265	8.31	-4047.69	-761.99	0.004	***
Fast Change Rate									
Delay Rate	-68.41	60.55	1	278	1.28	-187.56	50.74	0.259	
Delay Rate – 191*ms*	91.98	61.75	1	278	2.22	-29.54	213.50	0.137	
Delay Rate – 1738*ms*	-28.31	7.62	1	278	13.80	-43.31	-13.32	<.001	***
Block	-828.95	495.80	1	278	2.80	-1804.64	146.74	0.096	*
Trial	-2191.27	1005.16	1	278	4.75	-4169.94	-212.59	0.030	**
Very Fast Change Rate									
Delay Rate	-341.19	215.87	1	254	2.50	-766.15	83.76	0.115	
Delay Rate – 154ms	409.47	228.94	1	254	3.20	-41.23	860.17	0.075	*
Delay Rate – 450ms	-65.29	23.94	1	254	7.44	-112.43	-18.15	0.007	***
Block	-979.23	722.63	1	254	1.84	-2401.91	443.44	0.177	
Trial	483.97	1421.14	1	254	0.12	-2314.71	3282.65	0.734	

Chapter 5 appended statistical details

The following tables present the results from the piecewise linear regression models for Study 4, 5 and 6.

Study 4

			Solid			Plaid		
			Immediate	**Delayed**	**None**	**Immediate**	**Delayed**	**None**
Row	**Dependent Variable**	**N**	*Mean (SE)*	*Mean (SE)*	*Mean (SE)*	*Mean (SE)*	*Mean (SE)*	*Mean (SE)*
G	Completion Time (seconds)	288	48.0 [a] (5.78)	58.11 [b] (5.78)	55.7 [a,b] (5.70)	54.50 [a,b] (5.78)	79.96 [c] (5.78)	93.54 [d] (5.70)

G. Performance Model

Terms	DF_{Num}	DF_{Den}	F-Ratio	p-value	sig.
Shared Visual Information	2	266	25.32	<.0001	***
Puzzle Difficulty	1	266	7.42	0.0069	***
Lexical Complexity	1	10	9.62	0.0114	**
Shared Visual Information × Puzzle Difficulty	2	266	0.38	0.6848	
Shared Visual Information × Lexical Complexity	2	266	11.22	<.0001	***
Puzzle Difficulty × Lexical Complexity	1	266	0.96	0.3278	
Block	1	266	2.13	0.1453	
Trial	1	266	17.49	<.0001	***

Study 5

			Solid		Plaid	
			Immediate	Snapshot	Immediate	Snapshot
Row	Dependent Variable	N	*Mean (SE)*	*Mean (SE)*	*Mean (SE)*	*Mean (SE)*
H	Completion Time (seconds)	721	39.92 [a] (4.03)	53.43 [b] (4.03)	60.99 [b] (4.03)	91.63 [c] (4.03)

			Aligned		Rotated	
			Immediate	Snapshot	Immediate	Snapshot
Row	Dependent Variable	N	*Mean (SE)*	*Mean (SE)*	*Mean (SE)*	*Mean (SE)*
I	Completion Time (seconds)	721	32.13 [a] (3.19)	42.01 [b] (3.19)	68.77 [c] (3.19)	103.05 [d] (3.19)

H & I. Performance Model

Terms	DF_{Num}	DF_{Den}	F-Ratio	p-value	sig.
View Alignment	1	721	581.44	<.0001	***
Lexical Complexity	1	30	31.01	<.0001	***
Immediacy of Visual Information	1	721	118.80	<.0001	***
View Alignment × Lexical Complexity	1	721	2.32	0.1285	
View Alignment × Immediacy	1	721	36.30	<.0001	***
Lexical Complexity × Immediacy of Visual Info	1	721	17.89	<.0001	***
Block	1	721	20.17	<.0001	***
Trial	1	721	212.04	<.0001	***
Block × Immediacy	1	721	0.02	0.8923	
Block × Lexical Complexity	1	721	1.40	0.2371	
Block × View Alignment	1	721	5.46	0.0205	**
Trial × Immediacy	1	721	1.21	0.2723	
Trial × Lexical Complexity	1	721	10.99	0.001	***
Trial × View Alignment	1	721	84.73	<.0001	***
Block × Trial	1	721	0.28	0.6001	
View Alignment × Lexical Complexity × Immediacy	1	721	2.05	0.1524	

Study 6

Results from performance model examining the influence of the field of view

			Solid				Plaid			
			Full	Large	Small	None	Full	Large	Small	None
Row	Dependent Variable	N	*Mean (SE)*	*Mean (SE)*	*Mean (SE)*	*Mean (SE)*	*Mean (SE)*	*Mean (SE)*	*Mean (SE)*	*Mean (SE)*
J	Completion Time (seconds)	748	39.88 [a] (4.68)	44.57 [a,b] (4.68)	48.30 [b] (4.66)	70.91 [c] (4.68)	69.15 [c] (4.68)	73.77 [c] (4.66)	86.88 [d] (4.66)	113.90 [e] (4.68)

J. Performance Model

Terms	DF_{Num}	DF_{Den}	F-Ratio	p-value	sig.
Field of View Size	3	707	136.10	<.0001	***
Lexical Complexity	1	707	593.12	<.0001	***
Block	1	707	96.83	<.0001	***
Trial	1	707	146.36	<.0001	***
Field of View Size × Block	3	707	3.14	0.0248	**
Lexical Complexity × Block	1	707	0.47	0.4927	
Field of View Size × Lexical Complexity	3	707	5.73	0.0007	***
Trial × Lexical Complexity	1	707	9.99	0.0016	***
Trial × Field of View Size	3	707	5.05	0.0018	***

			Solid			Plaid		
			Automatic	Manual (Helper)	Manual (Worker)	Automatic	Manual (Helper)	Manual (Worker)
Row	Dependent Variable	N	*Mean (SE)*	*Mean (SE)*	*Mean (SE)*	*Mean (SE)*	*Mean (SE)*	*Mean (SE)*
K	Completion Time (seconds)	380	45.71 [a] (7.74)	46.77 [a] (7.77)	46.66 [a] (7.74)	71.99 [b] (7.74)	82.52 [b] (7.74)	85.97 [b] (7.74)

K. Performance Model

Terms	DF_{Num}	DF_{Den}	F-Ratio	p-value	sig.
Camera Control	2	21	0.265	0.7697	
Field of View Size	1	343	25.16	<.0001	***
Lexical Complexity	1	343	415.51	<.0001	***
Block	1	343	66.13	<.0001	***
Trial	1	343	61.85	<.0001	***
Camera × Block	2	343	4.97	0.0074	***
Field of View Size × Block	1	343	0.70	0.405	
Lexical Complexity × Block	1	343	7.35	0.007	***
Camera × Field of View Size	2	343	0.73	0.4844	
Camera × Lexical Complexity	2	343	5.58	0.0041	***
Field of View Size × Lexical Complexity	1	343	8.26	0.0043	***