HUMAN-COMPUTER INTERACTION, 2003, Volume 18, pp. 149–170

Things to Talk About When Talking About Things

Steve Whittaker
AT&T Labs–Research

ABSTRACT

This commentary reviews the existing research literature concerning support for talking about objects in mediated communication, drawing three conclusions: (a) speech alone is often sufficient for effective conversations; (b) visual information about work objects is generally more valuable than visual information about work participants; and (c) disjoint visual perspectives can undermine communication processes. I then comment on the four articles in the light of these observations, arguing that they broadly support these observations. I discuss the paradoxical failure of current technologies to support talk about objects, arguing that these need to be better integrated with existing communication applications. I conclude by outlining a research agenda for supporting talk about things, identifying outstanding theoretical, empirical, and design issues.

Steve Whittaker is a cognitive psychologist, with interests in the theory, evaluation, and design of collaborative and multimodal systems; he is a Senior Research Scientist at AT&T Labs–Research, Florham Park, NJ.

CONTENTS

1. INTRODUCTION

One strong theme of this set of articles is their attention to the role of visual information in supporting talking about things. But the role of visual information in communication turns out to be both complex and counterintuitive. So, to provide some context for my subsequent comments on the articles, I preface these by briefly reviewing what we know theoretically and empirically about the role of visual information in communication. I summarize prior research as a set of numbered observations, followed by the data that support each. I draw three main conclusions: (a) speech alone is often sufficient for effective conversations; (b) visual information about work objects is generally more valuable than visual information about work participants; and (c) disjoint visual perspectives can undermine communication processes. I then comment on the four articles in the light of these observations, arguing that

they broadly support these observations. I discuss the paradoxical failure of current technologies to support talk about objects, arguing that these need to be better integrated with existing communication applications. I conclude by outlining a research agenda for supporting talk about things, identifying outstanding theoretical, empirical, and design issues.

2. THE RESEARCH CONTEXT

1. Visual information is not always valuable, and speech is often sufficient to support communication (Sufficiency of Speech).

A great deal of previous research has viewed face-to-face communication as the touchstone for theory and design of technology to support mediated communication (Clark & Brennan, 1991; Daft & Lengel, 1984; Kraut, Fussell, & Siegel, 2003; Oviatt & Cohen, 1991; Sellen, 1995; Whittaker, 1995; Whittaker & O'Conaill, 1997). One strong intuition of previous research is the need to support visual information. Observations of face-to-face communication underscore that it is a complex multimodal process involving speech, gaze, gesture, and facial expressions (Clark, 1996; Clark & Brennan, 1991; Goodwin, 1981; Kraut et al., 2003; Whittaker, 1995; Whittaker & O'Conaill, 1997). This is especially true in complex object-centric tasks where visual behaviors play a central role. When talking about objects in face-to-face settings, people jointly orient to, gesture at, and manipulate objects, leading to observable transformations of those objects. Gaze, gesture, and facial expressions are all highly reliant on visual information, and having a shared visual frame of reference is critical for the interpretation of gaze and gesture. Observing gaze and gesture enables us to determine what objects other conversational participants are attending to and what they are likely to talk about (Cooper, 1974; Kahneman, 1973).

These observations about the role of vision in face-to-face communication give rise to theoretical hypotheses and design principles about interaction technologies. These take the view that multimodal technologies will better support complex communication than more impoverished ones providing speech-only conversation. For example, multimodal visual technologies such as video-conferencing (affording both speech and vision) should better support communication than unimodal technologies such as the telephone (speech only) or e-mail (text only). These intuitions have inspired technologies such as video-conferencing and the videophone. I now review evidence for the visual impoverishment hypothesis.

Unfortunately, the visual impoverishment hypothesis does not square with research findings. Chapanis, Ochsman, Parrish, and Weeks (1972, 1977) conducted an important set of laboratory experiments comparing the effect on

communication of various combinations of media. Somewhat surprisingly, they found for various cognitive tasks that communication using speech alone was as effective as either (a) face-to-face or (b) combined speech plus video communication. And Reid (1977) summarized 28 other similar studies, replicating the finding that adding visual information to speech-only communication does not change the outcome of cognitive tasks. These are significant results because they show that people are highly effective in using speech as a communication medium. Furthermore, the findings are not the result of implementation problems (such as low bandwidth video), because in most cases, speech is no different from face-to-face communication. Nor are they attributable to the use of outdated technologies. Recent laboratory (Sellen, 1992, 1995) and naturalistic studies (Fish, Kraut, Root, & Rice, 1992) show that video communication is not significantly different from speech communication.

Worse still, other research showed that adding visual information may impair critical aspects of spoken communication. For example, many video-conferencing systems introduce delays into speech by buffering it so that it can be synchronized with video. But several studies showed that such delays compromise important communication feedback processes that demand immediacy: for example, backchannels or interruptions (Anderson et al., 2000; Cohen, 1982; O'Conaill, Whittaker, & Wilbur, 1993; Whittaker & O'Conaill, 1997). This can affect the outcome of such conversations. So by trying to visually enrich communication channels, we may disrupt communication.

2. Visual communication environments are useful in a restricted set of circumstances.

These studies suggest a rather bleak future for visually based interaction technologies. But other studies suggest that the picture is more complex: Visual information is important in certain rather specific tasks, where nonverbal information is critical. Short, Williams, and Christie (1976) showed that visual information affected communication effectiveness for tasks that required access to emotional information. And Veinott, Olson, Olson, and Fu (1999) found that nonnative speakers benefited from video while giving directions, presumably because their lack of verbal fluency meant they were more reliant on nonverbal information.

3. Visual communication environments are difficult to design: What is shown and how it is shown are crucial.

The Short et al. (1976) and Veinott et al. (1999) studies show that visual information can be useful, but we need to refine our understanding of when and

how it helps. Two critical issues are: (a) what visual information we show; and (b) how we show that information.

3A. It is often better to show work objects rather than work participants when providing visual context (Objects not Participants).

Research into shared workspace applications provides unambiguous support for the importance of certain types of visual information in communication. To understand why, it is critical to distinguish the visual information provided in shared workspaces from that provided by traditional video applications.

In video-conferencing and in many videophone implementations, the visual channel shows a "head and shoulders" view of other participants, providing information about their gaze and facial expressions. This "talking heads" view contrasts with the visual information presented in shared workspaces. Instead, shared workspaces provide visual information about relevant shared objects (such as documents or a drawings) that the participants are jointly working on. Shared workspaces generally also allow all participants to directly modify those objects and to observe the effects of changes made by others.

Early studies of shared workspaces showed that adding this type of visual information improves the efficiency of speech communication (Bly, 1988; Whittaker, Brennan, & Clark, 1991; Whittaker, Geelhoed, & Robinson, 1993). For example, Whittaker et al. (1993) compared speech-only communication with speech plus a shared workspace for three different tasks: brainstorming, spatial design, and collaborative editing. Providing the workspace improved communication for spatial design and collaborative editing tasks but not for brainstorming. Analyses of linguistic behavior showed the reasons why: When the task requires reference to complex visual objects (spatial design) or complex layout (design and editing), people were able to use deictic gesture for both reference and to express complex spatial relations ("put that over here"). People were also more implicit in their communications when using the workspace, because the workspace supported situational awareness (Endsley, 1995). There is no need to explicitly communicate changes about the current state of the task if the other person can see this information directly. These effects were not found in the brainstorming task, which did not demand reference to complex objects, spatial relations, or object transformations. The results were interpreted in terms of Clark's theory of common ground (Clark, 1996; Clark & Wilkes-Gibbs, 1986). Similar effects for the efficiency of reference and situational awareness in shared workspaces were reported in Bly (1988), Karsenty (1999), McCarthy, Miles, and Monk (1991), McCarthy et al. (1993), Minneman and Bly (1991), and Tang (1991).

The shared workspace results led to a rethinking about the use of video (Whittaker, 1995; Whittaker & O'Conaill, 1997). Two research groups published articles proposing using video to show shared objects as opposed to work participants (Gaver, Sellen, Heath, & Luff, 1993; Heath, Luff, & Sellen, 1995; Nardi, Kuchinsky, Whittaker, Leichner, & Schwartz, 1996; Whittaker, 1995; Whittaker & O'Conaill, 1997). One application was telemedicine, where video was used to show distributed surgical teams' views of neurosurgeons' current actions, such as where the surgeons were cutting, the tool currently being used, and its angle of entry into the operating area (Nardi et al., 1996; Whittaker, 1995; Whittaker & O'Conaill, 1997). This situational awareness allowed nurses to anticipate the surgeons' movements and to provide surgeons with relevant tools without the need for explicit requests. It also allowed neurophysiologists working remotely to better interpret monitoring data about the state of the patient. They could then provide more timely and appropriate advice to the surgeons about the potential effects of the surgeons' current actions on the patient. Gaver et al. (1993) reported similar results. In their experiment, distributed participants could choose between different camera views when carrying out a complex spatial layout task. Views showed either other participants or the objects involved in the task. Shared object views were much more commonly selected than views of other participants, again underscoring the usefulness of information about objects as opposed to participants.

These results about the primacy of objects are also supported by research into nonverbal communication. First, looking at other people is the exception rather than the rule in conversation (Anderson, Bard, Sotillo, Doherty-Sneddon, & Newlands, 1997), and gaze at others falls to 3% to 7% of conversational time when there are interesting objects present (Argyle & Graham, 1977). Mutual gaze is even lower (Anderson et al., 1997; Kendon, 1967). These important results suggest that participants do not spend entire conversations monitoring other's facial expressions, especially when the environment contains relevant objects. This in turn would explain the greater success of video applications that depict shared objects and environments rather than showing images of other participants.

3B. Shared perspectives are critical: Disjoint perspectives may require extra work to resolve (Shared Perspectives).

An apparent exception to the findings about workspace object utility is a study by Tatar, Foster, and Bobrow (1991). Tatar et al. found that participants had extreme problems with achieving reference and consensus in a shared workspace. But one critical difference between their system and other shared workspaces lies in their implementation. Their Cognoter system was de-

signed to allow communicating participants to have *different* views on the same underlying set of objects. As a result, participants did not always share the same perspective or jointly observe transformations of those objects. This lack of a common view created problems for users in achieving a joint perspective and led to considerable difficulties when one participant made a change to an object that was not viewed by others. In contrast, most other shared workspaces present the same view of objects and are designed so that any change to an object is immediately presented to all other users (Greenberg, 1991; Minneman & Bly, 1991; Whittaker et al., 1993). This is necessary to achieve a shared perspective, which affords straightforward reference and a common view of the current state of the task.

Similar perspective-sharing problems are reported with video-conferencing systems. Video systems provide a restricted field of view, which means that participants at different ends of the video link have different perspectives. In addition, restricted camera resolution means that some objects may be hard for remote participants to see (Gaver, 1992; Kraut et al., 1996; O'Conaill et al., 1993; Whittaker, 1995; Whittaker & O'Conaill, 1997). Together these can lead to problems in object-sharing: Users report large problems in configuring systems to jointly view important objects such as documents or slides (Gaver, 1992; O'Conaill et al., 1993; Whittaker, 1995; Whittaker & O'Conaill, 1993, 1997). This is partly because it is hard to determine exactly what the remote participants can see.

3. ISSUES ARISING FROM THE ARTICLES

Having established the research context, I now turn to the findings presented by the four articles in this special issue.

3.1. Sufficiency of Speech

Martin and Rouncefield (2003) echo the *sufficiency of speech* observation. Their article illuminates the verbal strategies that participants use to refer to, and describe, the behaviors of objects that are invisible to their conversational partners. Martin and Rouncefield make these observations in the context of telephone banking, for relatively simple tasks like account inquiries, setting up standing orders, or paying bills. The range of objects discussed includes the bank's mainframe system and its operations, along with various physical objects such as the customer's bills, bank statements, letters, and so on. In these interactions, Martin and Rouncefield argue that the system is akin to a third party that has an important effect on the pacing, structure, and content of the interaction. Although Martin and Rouncefield do not claim that such verbal interaction is as efficient as interaction would

be if the participants found themselves face-to-face, they nevertheless document the set of skilled verbal strategies that in particular operators use to finesse system and situational constraints. These strategies involve structuring the conversation to meet the requirements of system interaction, pacing their interaction to mesh with system behaviors, and accounting for those system behaviors to their conversational partners, who are obviously unable to see the remote system and its operations. Operators also use organizationally defined scripts in an attempt to structure the conversation in ways that are convenient to the (relatively inflexible) operation of the system. Martin and Rouncefield also speculate that operators are sometimes able to exploit the fact that the system is invisible to the customer: Because the conversations are held over the phone, operators are able to hide or explain away aspects of its functioning that would otherwise be awkward or complex to account for.

The effectiveness and flexibility of speech communication are contrasted with an exploratory second study of an experimental video conferencing system, showing the customer a facial view of a remote banking expert, along with various documents such as forms or policies. The goal of the application was to allow customers access to experts who advise them about various types of financial policies. Consistent with the *sufficiency of speech* observation, Martin and Rouncefield conclude that adding video does not improve the interaction. Indeed Martin and Rouncefield offer some preliminary arguments suggesting that providing visual information may detract from the overall interaction quality (cf. O'Conaill et al., 1993; Sellen, 1995; Whittaker, 1995; Whittaker & O'Conaill, 1997). First, the system implementation is poor, so that various aspects of its functioning have to be explained to the customer. Second, Martin and Rouncefield argue that making information visible makes it hard for operators to hide irrelevant or misleading aspects of system functioning from the user, such as when the system presents inaccurate information about the customer.

Although being rather different in focus, the article by Ducheneaut and Bellotti (2003) makes similar points about the richness and flexibility of language (in their case e-mail text) in supporting object reference. They present preliminary data indicating that references to digital objects in e-mail are often extremely imprecise, and yet these are seldom misunderstood. Despite the fact that e-mail is a unimodal application supporting only text, participants are nevertheless able to refer to quite complex objects, without, Ducheneaut and Bellotti claim, becoming confused. Ducheneaut and Bellotti point out how this contradicts various influential theories (Daft & Lengel, 1984) that argue for the necessity of multimodal information for communication. As with the *sufficiency of speech*, their observations suggest that visual information is not necessary for effective object-centric interaction.

3.2. Objects Not Participants

In contrast to the mainly linguistically oriented articles of Martin and Rouncefield and Ducheneaut and Bellotti, the Kraut et al. (2003) and Luff et al. (2003) articles explore *objects not participants*. Both articles make the point that much prior research has focused on presenting "talking heads" rather than exploring the use of work objects in complex environments. Kraut et al. present two studies of a mentoring task in which a remote expert instructs a novice how to repair a bicycle. Novices wear a head-mounted camera that displays their field of view to the remote expert. The article first shows that providing this type of visual information is no more efficient than speech-only communication and is less efficient than face-to-face communication (a finding similar to the early observations about *sufficiency of speech*). But there are differences in communication processes between speech and video-mediated communication, indicating the importance of visual information. Video-mediated communication allows experts to be more proactive in offering help because situational awareness allows them to see when the novice is in difficulties. Consistent with the *objects not participants* results, when visual information is available, participants exploit this for deictic reference and spend less time discussing the state of the task or coordinating understanding. Again replicating the *objects not participants* results, communication is more implicit when visual information is provided because of situational awareness. The second study in the article also provides some fascinating data about the differences between face-to-face and video-mediated interaction, arising from participants' differing visual perspectives. For example, video-mediated interactions include more statements that attempt to coordinate perspectives ("can you see that?"), and experts are much less likely to use deixis than novices in video-mediated interaction, because novices cannot see the expert's gaze or gestures.

3.3. Shared Perspectives

These last two observations bear critically on *shared perspectives*, and they reveal two important weaknesses of the implementation (which Kraut et al., 2003, acknowledge). The first is that expert and novice do not strictly have a shared visual environment. As Kraut et al. point out, because the head-mounted camera presents a restricted field of view with limited resolution, certain objects that the novice can see may be out of the expert's field of vision or too small for the camera resolution to show clearly. This difference of perspective necessitates negotiation about exactly what the expert can see and hence knows about, and it may partially undermine the assumption that expert and novice have shared visual information. (In Clark's [1996] termi-

nology, expert and novice may not share common ground). As a result, pointing and deictic reference are compromised by the limited view offered to the remote participant. As other work on perspective taking shows, it is extremely difficult to understand another person's visual perspective when it is subtly (as opposed to radically) different from one's own (Schober, 1993). The second critical point is that experts in the mentoring task are not truly actors in the novice's environment. They cannot point to objects, and more importantly they cannot change the world they see. This leads to an important asymmetry between expert and novice both in their views of the world and in how they can refer to and act upon that world. Both have significant implications for the operation and success of the system.

Luff et al. (2003) make similar observations. Again, the system they studied was object, not participant, focused. It supported object-based interaction in complex environments rather than depicting "talking heads" information. The system is different from that studied by Kraut et al. (2003), because, in addition to providing remote participants with a view into the physical environment, it is intended to give them a "presence" in the physical environment in the form of a robot avatar. This avatar has two cameras for transmitting views of the physical environment as well as the ability to point at objects in the environment, using a laser pointer controlled by the remote user's mouse. The avatar is intended to provide remote participants with a means to navigate and explore the physical environment while at the same time offering their conversational partner information about their current focus of visual attention. This was intended to help create a shared physical perspective. But Luff et al. describe preliminary observations showing that participants experienced problems in coordinating a shared point of view. In particular, participants who are present in the environment found it difficult to determine the robot's field of view. As a result, this compromised gestures and deictic reference that relied on having that shared perspective. Luff et al. report that physically present participants made large adjustments to their gestures to compensate for this, changing the time course of gestures and waiting for long periods for confirmations that gestures have been understood. They also document how pointing gestures from the remote participant (using the laser pointer) are often misunderstood. As with their earlier research on video-mediated communication (Heath & Luff, 1991), one problem seems to be that the technology presents a close, but potentially misleading, approximation to the face-to-face situation (Schober, 1993). This leads participants to rely on their normal repertoire of behaviors, which may fail because the perspective is not genuinely shared. A second important factor is the fundamental asymmetry in the situation—although the remote participant can view and move in the physical environment, she or he cannot change it.

The Kraut et al. (2003) and Luff et al. (2003) results are interesting to compare with the *shared perspectives* observation. It seems that shared workspaces (in direct contrast to Luff et al.'s findings, and in part contrast to Kraut et al's results) allow the advantages of deictic reference and situational awareness, but without the some of the disadvantages they observed. Why is this? Partly it is because most shared workspace implementations offer a genuinely shared visual perspective. Although the visual field they display is limited and the objects that can be acted on are strictly digital, the perspective they support is shared. Because of this, there is no need to negotiate what each participant can see. Furthermore, shared workspaces allow symmetric access to the environment, so both participants can act on the visible objects, and there is no need to instruct another person to make changes to that environment.

There are also key differences between the Kraut et al. (2003) and Luff et al. (2003) findings. Luff et al. report many more problems with both reference, gesture, and negotiating a shared view. One reason for this is the greater complexity of their system, especially for the remote participant. In their GestureMan system, remote participants have to make sense of separate inputs from left and right eye sensors presented on large visual displays. They also have to use a pointer for remote gesturing. In contrast, in the Kraut et al. system, it may have been more straightforward for remote experts to interpret the output of the novice's head-mounted camera that displayed their field of view. Confirming my observation about *shared perspectives*, the Luff et al. implementation of the remote video may not have allowed remote participant's to determine their coworker's perspective. It would be interesting to see whether protracted experience or a different implementation would have helped remote participants with this.

This suggests several outstanding research issues for *shared perspectives*. One obvious problem is how to present these. Although the GestureMan avatar is highly promising as a technique for signaling the focus of attention of the remote participant, it may have to be built on a larger (more human) scale to better exploit human perspective-taking abilities. For example, participants in the physical environment should not have to stoop alongside the avatar to determine its visual perspective, and a larger avatar would avoid this. It is clear, too, that better techniques are needed to allow remote participants to immerse themselves in the physical environment to promote situational awareness. The two cameras with their separate large displays in the Luff et al. setup seem to be highly confusing. Another issue concerns symmetrical access and the inability of remote participants to act on their environment. One possible reason for the success of video-based telemedicine applications reported earlier (Nardi et al., 1996; Whittaker, 1995; Whittaker & O'Conaill, 1997) is that the main experts and protagonists, the surgeons, are physically present in the displayed environment. Other participants, observing their ac-

tions virtually, are there to assist while the surgeons carry out their actions. Because of the inherent problems of asymmetric access with this class of video applications, the telemedicine configuration may be a more desirable distribution of expertise than that of Luff et al., where both participants theoretically have equally contributions to the success of the task, or Kraut et al., where the expert with the most domain knowledge cannot act in the remote world. And one condition for successful asymmetric applications may be that the most knowledgeable actors in the situation are the ones who are situated in the physical environment.

3.4. Tasks and the Paradox of Document Sharing

This in turn raises important concerns about the classes of task that we are trying to support when talking about things. As we have seen, the demands of supporting collaborative document editing (Whittaker et al., 1993) are different from complex three-dimensional design tasks (Gaver et al., 1993; Kraut et al., 2003; Luff et al., 2003) or telemedicine tasks (Nardi et al., 1996; Whittaker, 1995; Whittaker & O'Conaill, 1997). We need more work that explores the relation between technology and task. Which important user tasks require access to complex visual environments, and how frequent are these? When do people need access to complex physical (as opposed to digital) objects? What other important tasks require conversations about objects, who are the participants, and what is the distribution of their expertise?

This brings us back to a paradoxical issue about shared workspaces. Despite their demonstrated utility, these have yet to become ubiquitous. Yet we know that collaborative talk around shared paper documents is a fundamental aspect of modern office work (Luff, Heath, & Greatbatch, 1992; Sellen & Harper, 2002; Whittaker et al., 1994). For example, Whittaker, Frohlich, and Daly-Jones (1994) found that 54% of informal office interactions involve documents. And distributed work has also become much more common, suggesting an increased need to interact about documents remotely (Hinds & Kiesler, 2002). Furthermore, we have reviewed experimental studies showing that shared workspaces are a highly effective way of supporting such collaborative editing (Whittaker et al., 1993). Similar application-sharing systems have been available as products for several years now (Proshare, Timbuktu), and in some cases (e.g., NetMeeting) are available free. Why have shared work- spaces not yet become pervasive, despite their apparent utility in supporting seemingly critical work tasks? This paradox is especially striking when one compares the limited popularity of shared workspaces with the pervasiveness of other applications like e-mail or instant messaging (IM) that do not provide explicit support for document-oriented interaction but are subverted for these purposes (Ducheneaut

& Bellotti, 2003; Isaacs, Walendowski, Whittaker, Schiano, & Kamm, 2002; Nardi, Whittaker, & Bradner, 2000).

3.5. Using Dedicated Versus Communication Applications for Talking About Things

So far we have focused on using speech to talk about objects, but as Ducheneaut and Bellotti point out, there are other textual technologies that people use to talk about objects. Ducheneaut and Bellotti correctly draw attention to the use of e-mail attachments and URLs as a way of talking about objects, but IM and chat are also used to talk for this purpose (Isaacs et al., 2002). These are strictly communication applications providing no direct support for manipulating objects, but there is also a tradition within CSCW of building dedicated tools to explicitly support collaborative authoring (Leland, Fish, & Kraut, 1988; Mitchell, Posner, & Baecker, 1995). Again there is an apparent paradox associated with these dedicated applications. Collaborative authoring and annotation systems have been in existence for many years, but their use has not become widespread (Cadiz, Gupta, & Grudin, 2000). This lack of success is surprising because dedicated applications offer direct support for established paper-based practices, where people mark up documents and distribute these comments to others (Sellen & Harper, 2002). How can we explain their lack of success? One possibility is that current applications do not support the precise type of annotations used in paper practices, and studies by Kraut, Galegher, Fish, and Chalfonte (1992) indicate that the exact implementation of annotations is critical. Sellen and Harper (2002) point to the importance of handwritten (as opposed to textual) annotations for collaborative authoring. These handwritten comments allow people to easily distinguish the original document from markup activities and comments, contrasting with the textual annotations supported by many computer applications. But this cannot be a complete explanation for the unpopularity of such applications, because there are examples of computer-based annotation systems that support handwritten notes (e.g., Whittaker et al., 1993).

Other work suggests an alternative reason for the lack of success of collaborative annotation applications. Brush, Bargeron, Gupta, and Grudin (2002) found that annotations are more effective when they are linked to alerts. This may be because alerts inform people of changes in a timely manner. With alerts, rather than reading about a change some time after it has been proposed, collaborators on a document are able to respond to and have more rapid communications about proposed changes. Again this echoes results on shared workspaces showing that interactive discussion about changes is effective in promoting consensus and closure in collaborative tasks (Bly, 1988; Whittaker et al., 1991; Whittaker et al., 1993).

Most annotation systems also do not provide explicit support for interactive discussion about documents. In contrast, two recent systems support this type of interactive textual discussion around objects (Churchill, Trevor, Bly, Nelson, & Cubranic, 2000; Whittaker, Swanson, Kucan, & Sidner, 1997). Whittaker et al.'s (1997) TeleNotes was an IM-like system that allowed people to incorporate objects such as documents, spreadsheets, URLs, or presentations into IM interactions. Users held IM-style impromptu conversations by exchanging textual "sticky notes," and the application allowed users to embed objects (such as documents or spreadsheets) into these conversations. Users could also launch different types of conversation (speech using click to dial, or video-conferencing) from these objects, so they could use these other interaction modes as well as IMs to discuss the object. Preliminary testing of this system showed that people frequently incorporated documents into their interactive textual conversations and that IM was often used to support quick questions and answer exchanges about documents. If other participants were offline or unable to respond, then IMs (along with their attached objects) were stored in an e-mail database for later processing. More recent work on traditional IM systems shows that users frequently use IM to talk together about objects that they are both independently viewing in a separate application (Isaacs et al., 2002).

Churchill et al. (2000) built a similar system that supported anchored chat conversations about documents or other desktop objects. Unlike TeleNotes (Whittaker et al., 1997), the system allowed users to attach chats to different regions of a document. One user could propose a change to a part of the document and attach the suggested change and comment to the document in that place. Unlike traditional annotation systems, where annotations are stored and read later, this annotation was interactive, so that if other collaborators were online, the system would alert them about the change and allow an interactive chat to take place. If they were offline, then chat comments were stored in a database. Again this approach supports dynamic collaboration: Making changes to a document is a collaborative activity, and interactive discussion is important for achieving consensus about proposed changes.

Overall, this research on textual interaction about objects suggests that we can view annotation systems, IM, and chats as other ways of "talking about documents." The key to success seems to lie in designing systems that promote interactive discussion about proposed changes: either by alerting or by integration with real-time interactive applications such as IM or chat. This offers better support for collaboration because it promotes building of consensus and closure. In contrast, dedicated applications do not support this form of interaction discussion. These observations seem to support those made by made by Ducheneaut and Bellotti (2003) but generalize their observations about e-mail to other forms of textual communication.

Another issue raised by Ducheneaut and Bellotti (2003) is how textual conversations can be converted into permanent conversational resources or objects. One common observation about successful UseNet discussion groups is that their conversations become transformed into archives that are useful conversational resources for people who have not participated in prior interaction. One common method for capturing results of prior conversations is the list of frequently asked questions (FAQ). Here a discussion moderator identifies repeated themes and their responses and presents these as question–answer pairs (Whittaker, 1996; Whittaker, Terveen, Hill, & Cherny, 1998). The intention is that new discussion participants will read the FAQ and avoid bringing up repeatedly discussed topics. Others have proposed organizing conversations around a fixed set of topics administered by a moderator. In a study of Lotus Notes, however, Whittaker (1996) found that moderation and the creation of predefined categories stifled discussion and the use of archival functions, suggesting the value of allowing freeform as opposed to structured interaction. The whole issue of converting conversations into archival resources is, as Ducheneaut and Bellotti note, an important one. Although progress has been made in developing visualizations of long-term conversations (Donath et al., 1999; Erickson & Kellogg, 2000; Smith & Fiore, 2001), we currently lack good tools for capturing and distilling conversations for reuse.

4. FUTURE RESEARCH FOR TALKING ABOUT THINGS

4.1. Theory

Developing Common Ground Theory

The best developed theory in this area is Clark's common ground (Clark, 1996; Clark and Brennan, 1991; Clark & Wilkes-Gibbs, 1986; Whittaker et al., 1991; Whittaker et al., 1998; Whittaker & O'Conaill, 1997). However, the theory can currently only provide adequate explanations for synchronous interaction, whereas much talk about objects is clearly asynchronous (Ducheneaut & Bellotti, 2003; Whittaker, 1996; Whittaker et al., 1997; Whittaker et al., 1998). Notions of common ground clearly begin to break down when participants work independently, which is the case in asynchronous communication (Whittaker et al., 1991; Whittaker et al., 1998). Furthermore, how can the theory account for situations like those observed in the Luff et al. (2003) and Kraut et al. (2003) studies, where participants believe that they shared common ground but differences of perspective mean that they do not?

Distributed Cognition

Another theory that has been used to explain talk about things is distributed cognition (Ackerman & Halverson, 1998; Hutchins, 1995). Distributed cognition describes various aspects of how artifacts are used in work settings, as shared representations that coordinate activities between coworkers, as methods to offload memory into the environment, and as devices to restructure complex tasks. Although distributed cognition provides good descriptions of these phenomena, it needs to be made more precise if it is to make predictions or lead to generative principles for the design of shared artifacts.

Explaining the Success of Speech and Textual Communication

As an example of the weakness of our current theories, none seems to be able to explain the utility of speech-only communication. The continued success of the phone and the lack of penetration of video-conferencing and even shared workspaces require an explanation. Similarly, text-based messaging such as IM or e-mail would intuitively seem to be a limited way to conduct interaction. Yet both grounding and distributed cognition theories argue for the utility of visual support for conversation. The theories obviously need to be refined to explain why linguistic technologies still dominate.

Taxonomies of Visual Information

We also need to develop better taxonomies of the role of different types of visual information in communication. Early attempts to do this are presented by Whittaker and O'Conaill (1997) and Kraut et al. (2003). Both taxonomies analyze different types of visual information (e.g., gaze, gesture, and environmental information) and suggest how this information contributes to fundamental interaction processes (e.g., deictic reference). But these taxonomies need to be further refined and evaluated.

4.2. Empirical Work

We also need more empirical studies of many aspects of object-centered interaction:

Task Taxonomies

Kraut et al. (2003) and Luff et al. (2003) have developed systems that support complex visual coordination, but are such tasks prevalent? And how frequent are tasks involving the synchronous sharing of digital objects (e.g., documents,

spreadsheets, or slides)? These important practical questions bear on which technologies and applications are the most important to address in the near term.

Why Aren't Shared Workspaces Used More?

Despite the ubiquity of document-centric interaction in face-to-face settings and the demonstrated success of shared workspaces in laboratory settings (Whittaker et al., 1993), these have yet to become pervasive. It is important to understand the reasons for this. Is their lack of popularity to do with established work practices, such as the fact that remote participants do not want to work together on shared objects in real time? Alternatively, real-time application sharing systems may have been misimplemented. Certainly, informal observations of technologies such as ProShare or NetMeeting suggest that they are hard to set up and use, but more research is needed into this question (Mark, Grudin, & Poltrock, 1999).

Why Aren't Annotation Systems Used More?

Another related question concerns annotation systems. Why have these yet to reach general use, when again there is evidence that collaborative talk and markup of documents are ubiquitous office activities (Sellen & Harper, 2002)? Is it better to integrate objects into existing communication systems such as e-mail or IM (Churchill et al., 2000; Whittaker et al., 1997) than to build stand-alone systems (Leland et al., 1988; Mitchell et al., 1995)?

Converting Conversations Into Archives

We also need more studies of long-term conversations and how these become used as conversational repositories (Whittaker, 1996; Whittaker et al., 1998). The emergence of FAQ is an intriguing phenomenon, but we need to understand more about how and when these are created and how they are used in subsequent group interactions.

4.3. Design Work

Finally we need more work into the design of systems to support talking about things.

Representing Discrepant Perspectives

When representing complex visual environments, we need better methods to show others' perspectives when these are not shared, possibly highlighting differences in perspective.

Asymmetric Access

We need to address that fact that remote participants often have reduced ability to affect remote visual environments. One possibility might be to develop effectors in the environment (like prosthetic hands and arms). Failing this, it might be possible to identify tasks where symmetric access is less critical, where remote participants can successfully contribute without needing to act on the environment (Nardi et al., 1996; Whittaker, 1995; Whittaker & O'Conaill, 1997).

Object-Enabling Existing Communication Systems

We need to develop technologies that support extremely lightweight object sharing that is integrated with preexisting communication systems—whether the communication is e-mail, IM, or speech. With some exceptions (e.g., Churchill et al., 2000; Whittaker et al., 1997), current object-sharing technologies are not well integrated with communication systems and they are extremely hard to set up and use.

Tools for Converting Conversations Into Archives

Although FAQs (Whittaker, 1996; Whittaker et al., 1998) and novel visualizations (Donath, Karahalios, & Viegas, 1999; Erickson & Kellogg, 2000; Smith & Fiore, 2001) provide useful tools for interrogating prior interactions, we need more research into reliable methods for extracting and depicting long-term conversational structures.

NOTES

Author's Present Address. Steve Whittaker, AT&T Labs–Research, 180 Park Ave., P.O. Box 971, Florham Park, NJ 07932–0971. E-mail: stevew@research.att.com.

REFERENCES

Ackerman, M., & Halverson, C. (1998). Considering an organization's memory. *Proceedings of the CSCW 88 Conference on Computer-Supported Cooperative Work*. New York: ACM.

Anderson, A. H., Bard, E., Sotillo, C., Doherty-Sneddon, G., & Newlands, A. (1997). The effects of face-to-face communication on the intelligibility of speech. *Perception and Psychophysics, 59,* 580–592.

Anderson, A. H., Smallwood, L., MacDonald, R., Mullin, J., Fleming, A., & O'Malley, C. (2000) Video data and video links in mediated communication: What do users value? *International Journal of Human–Computer Studies, 52,* 165–187.

Argyle, M., & Graham, J. (1977). The Central Europe Experiment—Looking at persons and looking at things. *Journal of Environmental Psychology and Nonverbal Behaviour, 1,* 6–16.

Bly, S. (1988). A use of drawing surfaces in collaborative settings. *Proceedings of the CSCW 88 Conference on Computer Supported Cooperative Work* (pp. 250–256). New York: ACM.

Brush, A. J., Bargeron, D., Gupta, A., & Grudin J. (2002). Notification for shared annotation of digital documents. *Proceedings of CHI 2002 Conference on Human Factors in Computing Systems*. New York: ACM.

Cadiz, J., Gupta, A., & Grudin, J. (2000). Using Web annotations for asynchronous collaboration around documents. *Proceedings of the CSCW 2000 Conference on Computer-Supported Cooperative Work*. New York: ACM.

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,* 487–509.

Chapanis, A., Ochsman, R., Parrish, R., & Weeks, G. (1977). Studies in interactive communication: II. The effects of four communication modes on the linguistic performance of teams during cooperative problem solving, *Human Factors, 19,* 487–509.

Churchill, E., Trevor, J., Bly, S., Nelson, L., & Cubranic, D. (2000). Anchored conversations: Chatting in the context of a document. *Proceedings of the CHI 2000 Conference on Human Factors in Computing Systems.* New York: ACM.

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

Clark, H., & Brennan, S. (1991). Grounding in communication. In L. B. Resnick, J. Levine, & S. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127–149). Washington, DC: APA Press.

Clark, 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. *Proceedings of Teleconferencing and Electronic Communications* (pp. 189–199). Madison: University of Wisconsin Press.

Cooper, R. (1974). The control of eye fixation by the meaning of spoken language. *Cognitive Psychology, 6,* 84–107.

Daft, R., & Lengel, R. (1984). Information richness: A new approach to managerial behavior and organizational design. In B. Straw & L. Cummings (Eds.), *Research in organizational behaviour* (pp. 191–233). Greenwich, CT: JAI.

Donath, J., Karahalios, K., & Viegas, F. (1999). Visualizing conversations. *Proceedings of HICSS-32.* Maui, HI: IEEE.

Ducheneaut, N., & Bellotti, V. (2003). Ceci n'est pas un objet? Talking about objects in e-mail. *Human–Computer Interaction, 18,* 85–110.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors, 37,* 32–64.

Erickson, T., & Kellogg, W. (2000). Social translucence: An approach to designing systems that mesh with social processes. *Transactions on Computer–Human Interaction, 7,* 59–83.

Fish, R., Kraut, R., Root, R., & Rice, R. (1992). Evaluating video as a technology for informal communication. *Proceedings of the CHI 92 Conference on Human Factors in Computing Systems* (pp. 37–48). New York: ACM.

Gaver, W. (1992). The affordances of media spaces for collaboration. *Proceedings of CSCW 92 Human Factors in Computing Systems* (pp. 17–24). New York: ACM.

Gaver, W., Sellen, A., Heath, C., & Luff, P. (1993). One is not enough: Multiple views in a media space. *Proceedings of the CHI 93 Conference on Human Factors in Computing Systems* (pp. 335–341). New York: ACM.

Goodwin, C. (1981). *Conversational organization: Interaction between speakers and hearers.* New York: Academic.

Greenberg, S. (Ed.). (1991). *Computer supported cooperative work and groupware.* London: Academic.

Heath, C., & Luff, P. (1991). Disembodied conduct: Communication through video in a multi-media environment. *Proceedings of the CHI 91 Conference on Human Factors in Computing Systems* (pp. 99–103). New York: ACM.

Heath, C., Luff, P., & Sellen, A. (1995). Reconsidering the virtual workplace: Flexible support for collaborative activity. *Proceedings of the ECSCW 95 European Conference on Computer Supported Cooperative Work.* Amsterdam: Kluwer.

Hinds, P., & Kiesler, S. (Eds.). (2002). *Distributed work.* Cambridge, MA: MIT Press.

Hutchins, E. (1995). *Cognition in the wild.* Cambridge, MA: MIT Press.

Isaacs, E., Walendowski, A., Whittaker, S., Schiano, D., & Kamm, C. (2002). The character, functions, and styles of instant messaging in the workplace. *Proceedings of CSCW 2002 Conference on Computer Supported Cooperative Work.* New York: ACM.

Kahneman, D. (1973). *Attention and effort.* Englewood Cliffs, NJ: Prentice Hall.

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,* 283–315.

Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica, 26,* 1–47.

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

Kraut, R., Galegher, J., Fish, R., & Chalfonte, B. (1992). Task requirements and media choice. *Human–Computer Interaction, 7,* 375–407.

Kraut, R. E., Miller, M. D., & Siegel, J. (1996). Collaboration in performance of physical tasks: Effects on outcomes and communication. *Proceedings of the CSCW 96 Computer Supported Cooperative Work Conference* (pp. 57–66). New York: ACM.

Leland, M., Fish, R., & Kraut, R. (1988). Collaborative document production using Quilt. *Proceedings of the CSCW 88 Conference on Computer Supported Cooperative Work.* New York: ACM.

Luff, P., Heath, C., & Greatbatch, D. (1992). Tasks-in-interaction: Paper and screen based documentation in collaborative activity. *Proceedings of the CSCW 92 Conference on Computer Supported Cooperative Work* (pp. 163–170). New York: ACM.

Luff, P., Heath, C., Kuzuoka, H., Hindmarsh, J., Yamazaki, K., & Oyama, S. (2003). Fractured ecologies: Creating environments for collaboration. *Human–Computer Interaction, 18,* 51–84.

Mark, G., Grudin, J., & Poltrock, S. (1999). Meeting at the desktop: An empirical study of virtually collocated teams. *Proceedings of the ECSCW 99 European Conference on Computer Supported Cooperative Work*. Amsterdam: Elsevier.

Martin, D., & Rouncefield, M. (2003). Making the organization come alive: Talking through and about the technology in remote banking. *Human–Computer Interaction, 18,* 111–148.

McCarthy, J. C., Miles, V. C., & Monk, A. F. (1991). An experimental study of common ground in text-based communication. *Proceedings of the CHI 91 Conference on Human Factors in Computing Systems* (pp. 209–215). New York: ACM.

McCarthy, J. C., Miles, V. C., Monk, A. F., Harrison, M. D., Dix, A. J., & Wright, P. C. (1993). Text-based on-line conferencing: A conceptual and empirical analysis using a minimal prototype. *Human–Computer Interaction, 8,* 147–184.

Minneman, S., & Bly, S. (1991). Managing a trois: A study of a multi-user drawing tool in distributed design work. *In Proceedings of the CHI 91 Conference on Human Factors in Computing Systems* (pp. 217–224). New York: ACM.

Mitchell, A., Posner, I. R., & Baecker, R. M. (1995). Learning to write together using groupware. *Proceedings of the CHI 95 Conference on Human Factors in Computing Systems* (pp. 288–295). New York: ACM.

Nardi, B., Kuchinsky, A., Whittaker, S., Leichner, R., & Schwarz, H. (1996). Video as data: Technical and social aspects of a collaborative multimedia application. *Computer Supported Cooperative Work, 4,* 73–100.

Nardi, B., Whittaker, S., & Bradner, E. (2000). Interaction and outeraction: Instant messaging in action. *Proceedings of the CSCW 2000 Conference on Computer Supported Cooperative Work* (pp. 79–88). New York: ACM.

O'Conaill, B., Whittaker, S., & Wilbur, S. (1993). Conversations over videoconferences: An evaluation of the spoken aspects of video mediated interaction. *Human–Computer Interaction, 8,* 389–428.

Oviatt, S., & Cohen, P. (1991). Discourse structure and performance efficiency in interactive and non-interactive spoken modalities. *Computer Speech and Language, 5,* 297–326.

Reid, A. (1977). Comparing the telephone with face-to-face interaction. In I. Pool (Ed.), *The social impact of the telephone* (pp. 386–414). Cambridge, MA: MIT Press.

Schober, M. F. (1993). Spatial perspective-taking in conversation. *Cognition, 47,* 1–24.

Sellen, A. (1992). Speech patterns in video-mediated communication. *Proceedings of the CHI 92 Conference on Human Factors in Computing Systems* (pp. 49–59). New York: ACM.

Sellen, A. (1995). Remote conversations: The effects of mediating talk with technology. *Human–Computer Interaction, 10,* 401–441.

Sellen, A., & Harper, R. (2002). *The myth of the paperless office*. Cambridge, MA: MIT Press.

Short, J., Williams, E., & Christie, B. (1976). *The social psychology of telecommunications*. London: Wiley.

Smith, M., & Fiore, A. (2001). Visualization components for persistent conversations. *Proceedings of the CHI 2001 Conference on Human Factors in Computing Systems* (pp. 136–143). New York: ACM.

Tang, J. (1991). Findings from observational studies of collaborative work. *International Journal of Man–Machine Studies, 34,* 143–160.

Tatar, D., Foster, G., & Bobrow, D. (1991). Design for conversation: Lessons from Cognoter. *International Journal of Man–Machine Studies, 34,* 185–210.

Veinott, E.S., Olson, J.S., Olson, G. M., & Fu, X. (1999). Video helps remote work. *Proceedings of the CHI 99 Conference on Human Factors in Computing Systems* (pp. 302–309). New York: ACM.

Whittaker, S. (1995). Rethinking video as a technology for interpersonal communications: Theory and design implications. *International Journal of Man–Machine Studies, 42,* 501–529.

Whittaker, S. (1996). Talking to strangers: An evaluation of the factors affecting electronic collaboration. *Proceedings of the CSCW 96 Conference on Computer Supported Cooperative Work* (pp. 409–418). New York: ACM.

Whittaker, S., Brennan, S., & Clark, H. (1991). Coordinating activity: An analysis of computer supported co-operative work. *Proceedings of CHI 91 Conference on Human Factors in Computing Systems* (pp. 361–367). New York: ACM.

Whittaker, S., Frohlich, D., & Daly-Jones, O. (1994). Informal workplace communication: What is it like and how might we support it? *Proceedings of CHI 94 Human Factors in Computing Systems* (pp. 130–137). New York: ACM.

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. (1993). Evaluating videoconferencing. *Proceedings of the CHI 93 Human Factors in Computing Systems* (pp. 135–136). New York: ACM.

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. Wilbur (Eds.), *Video-mediated communication: Computers, cognition, and work* (pp. 23–49). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Whittaker, S., Swanson, G., Kucan, J., & Sidner, C., (1997). Telenotes: Managing lightweight interactions in the desktop. *Transactions on Computer–Human Interaction, 4,* 137–168.

Whittaker, S., Terveen, L., Hill, W., & Cherny, L. (1998). The dynamics of mass interaction. *Proceedings of the CSCW 98 Conference on Computer Supported Cooperative Work* (pp. 257–264). New York: ACM.