4 How Conversation Is Shaped by Visual and Spoken Evidence

Susan E. Brennan

When two people communicate successfully, they each come to the belief that they are talking about the same things, and their individual mental representations seem to converge. How does this happen? Perhaps the simplest explanation is that as long as both speaker and addressee are rational, cooperative, and following the same linguistic conventions, understanding emerges serendipitously. As Sperber and Wilson (1986, 41) have stated, "Clearly, if people share cognitive environments, it is because they share physical environments and have similar cognitive abilities." This explanation for how speakers and addressees come to believe they are talking about the same thing emphasizes the ways their abilities, environments, and language processes are similar. Not only are two individuals in conversation likely to share some of the same biases, but the processes of production and comprehension themselves likely share the same resources. That is, what is easy for an individual to understand is often easy for that individual to produce (Brown and Dell 1986; Dell and Brown 1991).

A second sort of explanation of how people achieve shared mental representations emphasizes the interactive *coordination* of meaning, above and beyond speakers using encoding rules that match addressees' decoding rules. In other words, successful communication depends not only on conventions about the content of messages, but also on a metalinguistic process by which conversational partners interactively exchange evidence about what they intend and understand. This is *not* to say that similar abilities and biases play no role in successful communication, but that these are often not *sufficient* to achieve shared mental representations. Consider this episode of a spontaneous face-to-face conversation:

Susan: You don't have any nails, do you?
Bridget: ***(pause)***
Fingernails?

Susan: No, nails to nail into the wall.
⟨pause⟩
When 1 get bored here I'm going to go put up those pictures.

Bridget: No.

The two people in this example spoke the same native language. They lived in the same town, were members of the same university community, and in fact were both psycholinguists. They shared the same graduate advisor, interests, social milieu, and office. In this conversation they were talking about something concrete, using simple, high-frequency English words. But similarity alone did not guarantee that Bridget would immediately understand what Susan meant. It turns out that Bridget had a different meaning in mind for "nails" than Susan did, but Susan did not discover this until after Bridget tested her hypothesis with "fingernails?" What Bridget may have had in mind during the pause after Susan's "no, nails to nail into the wall" seems less clear. Was Bridget trying to recollect the state of her toolbox? Was she wondering whether Susan was just being sarcastic and really did mean fingernails? After a pause during which Bridget did not take up the attempted explanation, Susan offered a further rationale for why she needed the nails. Then Bridget's hypothesis about what Susan meant by "nails" appeared to converge with Susan's, and she provided a relevant answer. At this point, but not before, Susan could conclude that she had succeeded in asking her question.

Clearly, similarity between two conversational partners can get them part of the way toward converging mental states. And the fact that, within the same mind, the processes of speech production and comprehension share some of the same resources increases the odds that what is easy to say (e.g., high-frequency words) will be easy to understand. But similarity is not enough. Rather than simply delivering and receiving messages, speakers and addressees jointly construct and negotiate meanings in conversation. This is necessary because natural languages afford a gencrativity and a flexibility that formal languages do not. For instance, the possible mappings from word to referent change from situation to situation and from speaker to speaker (see, e.g., Brennan and Clark 1996), and speakers routinely create new words when the need arises (see, e.g., Clark 1983; Clark and Gemg 1983). Moreover, speakers and addressees process language in the face of potential noise, distractions, and limited cognitive resources.

What ensures successful communication is that speakers and addressees engage in a process of ***grounding,*** in which they continually seek and provide evidence that they understand one another (Brennan 1990; Clark and Brennan 1991; Clark and Schaefer

1987, 1989; Clark and Wilkes-Gibbs 1986; lsaacs and Clark 1987; Schober and Clark 1989). According to Clark (1994, 1996), language use is concerned with two sorts of signals: those in Track *1*, having to do with the primary, "official business" of the conversation, and those in Track 2, the secondary, often paralinguistic signals that are used to ground or coordinate the understanding of the material in Track 1 (Clark 1994, 1996). Repairs like the one between Bridget and Susan can be accounted for by considering how Track 2 signals aid in disambiguating Track 1 signals.

Much research on language can be characterized as fitting either a *language-as-product* tradition (where the focus is on utterances and their processing in a generic or "default" context and where comprehension is considered apart from production) or a *language-as-action* tradition (where utterances are seen as emerging from both intra- and interpersonal processes, embedded in a physical context). Many psycholinguists work squarely in the first tradition, whereas some, along with ethnographers and conversation analysts, work in the second. A goal of this book is to bridge these traditions, in part by presenting research on spontaneous, interactive language use in carefully chosen contexts, using online measures that afford some degree of experimental control and reliability. This chapter reports previously unpublished details of an experiment that attempted to do this, before unobtrusive head-mounted eye trackers were available for measuring moment-by-moment processing (Brennan 1990). In this introduction, 1 will describe a framework from the language-as-action tradition with which to view both *intra*personal and *inter*personal processes of language use.

Seeking and Providing Evidence for Mutual Understanding

The Contribution Model

Clark and Wilkes-Gibbs (1986) originally proposed (but did not test directly) a principle of mutual *responsibility*, stating that people in conversation try to establish the mutual belief that the addressee understands a speaker's utterance to a criterion sufficient for current purposes before they move the conversation along. This proposal contrasts with theories of communication that assume that the responsibilities of conversational participants are fixed and that speakers bear all responsibility for avoiding misunderstanding (e.g., Sperber and Wilson 1986, 43). The principle of mutual responsibility and the process of grounding extend Grice's cooperative principle (Grice 1975) to the unfolding of a conversation over time.

Mutual responsibility was formalized in Clark and Schaefer's contribution model. According to this model, a contribution to a conversation is achieved jointly in two phases: a *presentation* phase and an acceptance phase. Every utterance or turn in a conversation represents a presentation, as when, in our example, Susan said to Bridget, "You don't have any nails, do you?" But a presentation cannot be presumed to be part of a speaker and addressee's common ground until its acceptance phase is complete— that is, until there is enough evidence for the speaker to conclude that the addressee has understood. An acceptance phase can be longer than a single utterance, with additional contributions nested inside it, as with Susan's first attempt at a question and the exchange that followed. Evidence of understanding can be explicit, as when a partner provides a backchannel response, a clarification question, or a demonstration, or it can be implicit, as when a partner continues with the next relevant utterance (as when Bridget finally answered, "No"). Participants set higher or lower ***grounding criteria*** for the form, strength, and amount of evidence necessary at any particular point (Clark and Wilkes-Gibbs 1986; Wilkes-Gibbs 1986, 1992). Grounding criteria vary depending on the current purposes of the conversation (Clark and Rrennan 1991; Wilkes-Gibbs 1986; Russell and Schober 1999) and also on the resources available within a communication medium (Clark and Brennan 1991; see also Whittaker, Brennan, and Clark 1991).

Clark and Schaefer supported the contribution model with examples from the London-Lund corpus of British English conversation (Svartvik and Quirk 1980). They showed how the contribution model could result in data structures (contribution trees) that emerge as the ***product*** of conversations. However, as they themselves pointed out, transcripts are only products, and the data from the Lund corpus do not capture the moment-by-moment ***processes*** by which speaker and addressee coordinate their individual knowledge states (Clark and Schaefer 1989, 273–274). In the next section, I will highlight why it is that language transcripts alone are inadequate for testing these predictions.

Conversation Online

A text transcript and the recording it originated from contain clues about what happened in a conversation. But the previous gloss of Bridget and Susan's misunderstanding is ultimately not very satisfying as a window into their processing, for at least three reasons. First, there is ample evidence that overhearers, addressees, and side participants experience a conversation differently (Kraut, Lewis, and Swezey 1982; Schober and Clark 1989; Wilkes-Gibbs and Clark 1992). Discourse analysts are, ordinarily, only overhearers.

A second problem is that a post hoc account based on a transcript does not enable predictions about why, at each juncture, Rridget and Susan would do what they did.

Although some aspects of conversation seem routine, there are in fact many options at every point. For instance, Susan could have first said "Do you have any nails?" without the tag question. Bridget could have held up the back of her hand so that Susan could see for herself. Or, after the second pause, Susan could have waited longer for an answer from Bridget, rather than providing a justification for her question. Sometimes conversants opt to provide more evidence about their own beliefs, and sometimes they opt to seek more evidence about their partner's. Are these choices systematic? How do conversants know when to stop seeking and providing evidence and conclude that they understand each other well enough? While a transcript provides some information about a conversation's product, it says little about the process from which the product emerges.

A third problem with relying on transcripts in the study of discourse is that they give no independent evidence of what people actually do understand or intend at different points in a conversation. Many referential communication studies have addressed this problem by collecting task-oriented dialogues using a variant of the card-matching task developed by Krauss and his colleagues (Clark and Wilkes-Gibbs 1986; Isaacs and Clark 1987; Krauss and Glucksberg 1969; Krauss and Weinheimer 1964, 1966, 1967; Schober and Clark 1989). The assumption is that by observing a task in which pairs of people have to move objects into some preset configuration, we can tell when they are talking about the same object and when they have misunderstood one another. However, the typical referential communication task manages to document outcomes of matching trials without uncovering much about the time course by which two people get their individual hypotheses to converge.

What is needed, then, is a way to study the grounding process *online,* as it unfolds. A classic referential communication study in the language-as-action tradition that took steps in this direction videotaped participants in a matching task (Schober and Clark 1989). Schober and Clark observed that addressees sometimes picked up a card and held it for a while before placing it in the target location described by their partners; they proposed that these addressees had reached an individual "conjecture point" prior to the "completion point" by which both partners could conclude that they understood one another and move on to the next card. That study served as the inspiration for the one reported here, which documents the online processing of evidence during conversation in greater detail. More recently, the use of head-rnounted eye trackers (e.g., Tanenhaus et al. 1995) has enabled the precise and unobtrusive tracking of eye gaze, providing a nuanced measure of what people intend and understand in conversation.

Language Use as Hypothesis Testing

Consider what individuals need to do in order to achieve shared meanings with their conversational partners. In the case of addressees, this seems straightforward; the addressee (like the reader) forms possible interpretations or ***meaning hypotheses*** (Krauss 1987) and then tests and revises them as evidence accrues (Berkovits 1981; Kendon 1970; Krauss 1987; Rumelhart 1980). Meaning hypotheses involve smaller hypotheses (some conscious, some not) that concern many dimensions of the utterance, such as who is being addressed, what word to retrieve, how best to resolve lexical or syntactic ambiguity, where the speaker's attention is, what is presupposed, what schema to evoke, what part of an utterance is new and what is given, and how the utterance is relevant to the situation. The addressee can set his criterion for rejecting a meaning hypothesis conservatively or liberally, depending on what is at stake. He has the option of pursuing more evidence if the intention behind an utterance or word is unclear.

Addressees are not the only ones doing hypotheses testing; speakers do it too. An utterance embodies a speaker's hypothesis about what might induce her addressee to recognize and take up her intention at a particular moment. The speaker monitors the addressee's responses such as eye gaze, nods, verbal acknowledgments, and other backchannels (Yngve 1970), relevant next turns and actions, and clarification questions (Bruner 1985; Clark and Schaefer 1989; Goodwin 1979; Heath 1984); these responses provide evidence of attending and understanding. The speaker evaluates the response she observes against the response she expected; she can then refashion her utterance and re-present it, or even revise her original intention so that it now converges with the one her addressee seems to have recognized.' So the speaker's hypothesis, as expressed in her utterance, plays a dual role by providing the evidence against which the addressee tests his hypothesis, while his response in turn enables the speaker to test hers.

Note that there is a built-in temporal asymmetry with respect to the speaker's and the addressee's hypothesis testing (Brennan 1990; Cahn and Brennan 1999). When Susan said, "You don't have any nails, do you?", Bridget recognized there might be a problem before Susan did. When Bridget replied, "fingernails?", Susan realized that Bridget had misunderstood her question before Bridget did. Because neither partner is omniscient with respect to the other's mental state, and because of this asymmetry between their distinct mental states, the computation of common ground is not determinate, but is made in the face of uncertainty. Since mutual knowledge cannot be proven, people rely on copresence heuristics in order to assume that they understand one another well enough for current purposes (Clark and Marshall 1981).

The rest of this chapter presents an experiment on the time course of how pairs of people in conversation get their meaning hypotheses to converge. Two partners' hypotheses should not converge steadily, but in phases that correspond roughly to the presentation and acceptance phases of contributions. Moreover, these phases—particularly the acceptance phase—should differ depending on the modality of the evidence available for grounding, as predicted by Clark and Brennan (1991). For this experiment, I developed a computerized spatial task that recorded two partners' mouse movements in order to provide continuous, moment-by-moment estimates of how the speaker's and addressee's meaning hypotheses were converging. This information was then synchronized with the transcript of their utterances. After describing the experiment and results I will compare this sort of behavioral measure to that of eye tracking, because each measure has something different to contribute to the study of conversation online.

Predictions

In this section 1 develop four specific predictions about moment-by-moment coordination that follow from the contribution model and indicate how these predictions were tested (additional detail can be found in Brennan 1990). These predictions are contrasted with some alternatives that arise from other assumptions or proposals about language use in conversation. These four predictions concern (1) how quickly and closely two partners' meaning hypotheses come to converge, (2) how having visual evidence about a partner's beliefs affects the time course of convergence, (3) how prior knowledge affects the time course of convergence, and (4) how the modality of evidence affects the grounding criterion, and in turn, how closely two meaning hypotheses can be made to converge. I will outline plans for testing each prediction.

Prediction 1: Addressees Form **Early** Meaning Hypotheses That They Then Ground with Their Partners The first prediction from the contribution model is that in spoken conversation, addressees should form meaning hypotheses relatively early in an exchange. Then it should take significant additional time and effort for addressees and speakers to determine and signal to each other that their hypotheses have converged, especially if they are limited to exchanging evidence via verbal utterances that have been linearized (more or less) as a sequence of speaking turns. Alternative possibilities are that an exchange could end shortly after the addressee forms a correct meaning hypothesis, or that any additional time after having formed a correct meaning hypothesis could be spent in silent deliberation. This could be the case if people in conversation did not ground or coordinate their mental states with input from their partners—that is, if addressees and speakers tested their hypotheses individually without taking an active role in each other's hypothesis testing. This is what would be expected if convergence was achieved from simply having the same biases and using the same system for encoding and decoding utterances.

I tested these predictions using a collaborative matching task in which twelve pairs of same-sex strangers who could not see each another discussed locations on identical maps displayed on networked computer screens. One person, the *director,* saw a car icon in a preprogrammed target location, and the other person, the *matcher,* used a mouse to move his own car icon to the same location. This task enabled moment-by-moment tracking of how closely the matcher's and director's hypotheses converged over time, synchronized with what they said.

Prediction 2: The Modality of Evidence Shapes the Grounding Process Many studies have found differences in conversations or tasks conducted over different media (e.g., Chapanis et al. 1972; Cohen 1984; Ochsman and Chapanis 1974; Williams 1977); the evidence available for grounding and the contribution model provide a framework for explaining these differences (Clark and Brennan 1991). The next prediction about the time course of reaching shared hypotheses concerns the impact, moment by moment, of having both visual and spoken evidence of a partner's understanding, as opposed to only spoken utterances and backchannels.

Clark and Schaefer (1989, 267) stated that it is generally up to the addressee to initiate the acceptance phase for an utterance. Typically in spoken conversation (especially when participants are not visually copresent), the addressee is in the best position to judge the goodness of his own hypothesis and to propose to the speaker that he understands what she meant. But if the distribution of responsibility is flexible, the acceptance phase should be initiated by whichever partner first amasses strong enough evidence that the addressee's hypothesis is a good one.

In the current task, on half of the trials the director saw the matcher's icon superimposed on her[2] own screen (visual evidence condition), and on half she did not (verbal-only evidence condition). Having visual evidence in addition to verbal evidence may speed up the presentation phase somewhat, if the director is able to adapt her descriptions of the target location moment by moment to what she can see of the matcher's attempts to move there. However, visual evidence should have its strongest impact *late* in a contribution. When the director could see the matcher's icon, both partners should expect the director to take over the responsibility for proposing that their hypotheses had converged. This should shorten the acceptance phase considerably by providing strong evidence of convergence to the director and by saving the matcher at least one speaking turn.

A more specific expectation about the effect of evidence involves addressees' verbal backchannels. According to the contribution model, backchannels are specific, relevant signals intended for grounding (Brennan 1990; Clark and Brennan 1991; Clark and Wilkes-Gibbs 1986; Clark and Schaefer 1989). If this is the case, then matchers should take into account the changing modality of evidence available ***to their partners*** at any given moment and adapt their own responses accordingly, even when the evidence available to the matchers themselves stays the same. An alternative possibility is that backchannels are general, diffuse responses to speech, regulating the "flow of information" (Rosenfeld 1987), serving "to organize and to direct the stream of communication" (Duncan 1973, 29), or acting as reinforcers to encourage the speaker to continue talking (Duncan 1975; Wiemann and Knapp 1975). As such, the particular form that backchannels take may be simply a practiced, automatic response to the rate of information presented within a particular communication medium as experienced by an addressee; people unfamiliar with a particular medium may need to learn to produce appropriate backchannels in that medium (as Cook and Lalljee (1972) and Williams (1977) have proposed). If this is the case, then a matcher's verbal backchannel behavior should be difficult to modify; he should provide about the same number and kinds of verbal responses regardless of whether his partner has visual evidence about what he understands. This would also be expected if the main purpose of addressee responses is to reinforce a social relationship or to show general engagement in a conversation, as some analysts have proposed. I tested this set of predictions by comparing the distributions of acknowledgments in the visual and verbal conditions.

Prediction 3: Prior Shared Knowledge Has an Early Impact It is reasonable to expect that having more shared and relevant prior knowledge—which includes having more similarity in knowledge and experiences—should enable two people to understand one another more quickly, thereby shortening an exchange. The contribution model enables more specific predictions as to when and how both prior shared knowledge and type of evidence affect the grounding process. Prior shared knowledge should have its strongest effects early, during the presentation phase, where it may help the director tailor her description to the matcher's needs, enabling the matcher to form a reliable meaning hypothesis sooner. The alternative to this prediction is that prior shared knowledge should shorten all parts of an exchange roughly equally. 1 tested these possibilities by varying the knowledge that two partners could assume they shared at the outset of the task: the pairs (who were Stanford University graduate students recruited from thirteen academic departments) discussed locations half the time on maps of the Stanford campus and half the time on maps of Cape Cod (a locale with which they were unfamiliar).

Prediction 4: Grounding Is Only as Precise as It Needs to Be Some have assumed that having more evidence or copresence should lead to greater convergence in understanding than having less (see, e.g., Karsenty 1999). If this is so, we might expect that people's hypotheses (and their corresponding icon locations) would converge more closely when the director could see where the matcher's icon is located (visual evidence) than when she could not (verbal-only evidence). But according to the contribution model and the grounding framework, people in conversation do not try to get their hypotheses to converge perfectly—in fact, since neither party is omniscient, this is not even feasible. Instead, they try to reach a level of convergence that is sufficient for current purposes, satisficing in Simon's (1981) terms. Efforts at convergence are guided by the grounding criteria people set and by how effectively they can use the evidence available in a communication medium.

So when visual evidence is readily available, people should be only as detailed and persistent in their evidence providing and evidence seeking as they need to be to satisfy current purposes. When evidence of a partner's understanding is weaker or less direct, as in the verbal condition, they will have to set their grounding criteria higher to ensure the same level of performance, causing them to be more accurate on average (that is, more convergent) than they need to be and less efficient overall. This yields the specific and somewhat counterintuitive prediction that two people's meaning hypotheses about a spatial location may come to converge less closely when they have visual evidence than when they do not. 1 tested this prediction by giving pairs a criterion for the task: they were to get their car icons parked in the same spot so that if their two screens (each measuring 1024 x 768 pixels) were superimposed, at least part of their icons (each measuring 16 x 12 pixels) would overlap.

Language-Action Transcripts: Design and Analysis

Each of the twelve pairs described a total of eighty different preprogrammed target locations. Within a pair, one person acted as director for four blocks of ten trials (where a tial involved describing one location on a map) and then switched roles with the partner in order to act as matcher for the subsequent four blocks. After each block, either the evidence condition or the map changed. The evidence condition alternated between visual and verbal-only evidence; in the verbal-only condition, the director could not see where the matcher's icon was, and they had to do the task entirely by conversing aloud, whereas in the visual condition, the director could, in addition, monitor the position and movement of the matcher's icon on her own screen, with the matcher's icon appearing in a different color than the her own. The maps, both in black and white and equally legible and detailed, alternated every two blocks between one of the Stanford campus and one of Cape Cod. These two locales varied in their familiarity to participants, and the maps were therefore intended to vary in the amount of prior knowledge the participants were likely to share. Map, evidence condition, director-matcher roles, and presentation order were completely counterbalanced for a repeated-measures design both within pairs and within locations. Directors and matchers were explicitly informed at the outset of each block as to whether the director would be able to see the matcher's icon.

The director began a trial by clicking on her icon, which then moved automatically to a preprogrammed target location and stopped. Whenever the matcher was ready, he selected his icon by clicking on it and then freely moved it over the map. Once the matcher believed his icon to be in the target location, he then parked it by clicking again. This concluded the trial; once the icon was parked, it could not be moved again until the director initiated the next trial. Conversation during each trial was audiotaped in stereo, and a time-stamped log file of both partners' mouse clicks and the *x*- and y-coordinates of their icons was generated automatically. This log began when the director initiated the trial, continued through when the matcher picked his icon up, and ended when he finally parked it.

Action Transcripts

The log files were reduced by a filtering program that automatically identified and timed all trial durations, pauses, and intervals during which the matcher stopped moving within a specified radius of the target location. There was some "jitter" associated with pausing the mouse, so a pause in the action was considered to be any period of time when the matcher moved his icon by two pixels or fewer. After the data were reduced, the distance between the director's and matcher's icons was calculated for each time increment; these data were the basis for the action transcript, which was represented as the plot of the distance between their icons over time. Figure 4.1 displays a prototypical time-distance plot generated by one matcher during one trial.

Analysis of the action transcripts was based on the assumption that the matcher's icon movements provided an estimate of what he understood at that point. Of course, icon location did not always correspond precisely to the matcher's hypothesis about

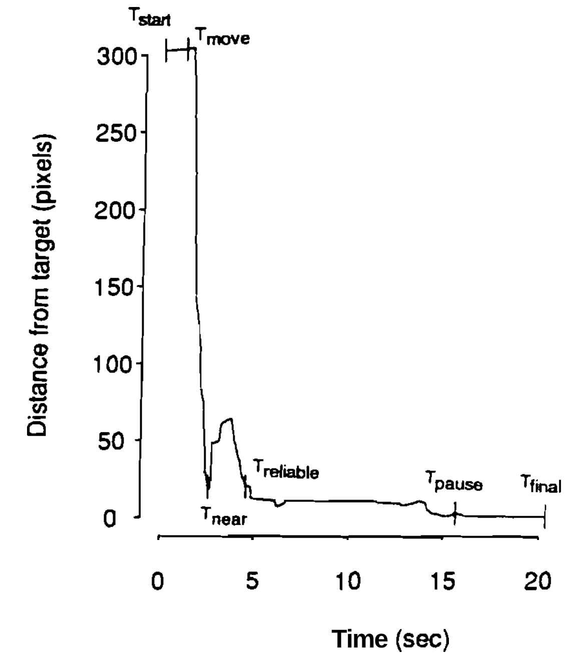


Figure 4.1
Time-distanceplot and points of interest for one typical trial.

the target, since at times he may have had only a vague location (or set of locations) in mind, or he may have been temporarily "stuck." But at any particular moment, the icon location constituted an observer's best estimate of the matcher's hypothesis about what the director had presented so far. When the matcher's icon was motionless and when the distance between the matcher's and the director's icons was equal to zero, the matcher was assumed to have a perfectly convergent hypothesis of where the target location was.

The time-distance plots show to what extent the matcher's overall progress toward the target was monotonic. A steep negative slope indicates movement directly toward the target. An abrupt change in slope indicates one of three things. First, an acute angle where the slope changes its sign from negative to positive indicates that the matcher moved on a continuous path that went right by the target. His close approach might have been due not to a precise, correct hypothesis, but to chance. Second, when the slope changes from positive to negative, the matcher has gone from moving away from the target to moving toward it. Finally, a change from a sloped line to a horizontal one indicates that the matcher has stopped moving altogether or else has moved at a consistent distance from the target (essentially circling it). These possibilities are distinguished by the star-shaped points on the plot, which correspond to moments when the experiment software detected a movement from the mouse; a horizontal region with few points on it indicates that the matcher's icon was stationary or nearly so. Points located close together on a slope indicate slower movement than points located farther apart.

Several moments of interest during the time course of a trial were identified automatically and are labeled on figure 4.1: T_{start}, T_{move}, T_{near}, $T_{reliable}$, T_{pause}, and T_{final}. T_{start} was the moment the matcher first clicked to pick up his icon. T_{move} was the moment he started to move it. Presumably, at that point he must have had some idea where to move it (even if only in which general direction). Sometimes there was initial movement away from the target at the very beginning of a trial. This initial movement was probably not a reliable indicator of where a matcher thought the target was; some individuals seemed to make a large movement right after they clicked to select their icons, perhaps just to see if the click had taken effect. Apart from this difference in mouse technique at the very beginning of some trials, icon movement was assumed to provide an index into the matcher's understanding of the director's verbal presentation of the target location.

T_{final} was the moment when the matcher parked his icon, ending the trial. This point corresponds to the "completion point" in the referential communication study of Schober and Clark (1989). Our window into a matcher's understanding, then, consists of the period of time from the matcher's first icon motion, or T_{move}, to when he finally decided to park. If the matcher paused along the way, the locations where he paused were taken to approximate his intermediate meaning hypotheses. T_{near} was between T_{move} and T_{final}, when the matcher first arrived within close radius of the target location. This was the earliest location where the matcher would have been correct, or very nearly so, had he parked the icon there. Of course, chance movements could also have led the matcher closer to the target, so to properly identify the moment when the matcher had a ***reliable*** hypothesis, not only did his icon need to be overlapping the director's icon, but it should not move out of range of the target again before T_{final}. So $T_{reliable}$ was determined as the moment when the center of the matcher's icon arrived within a 20-pixel range of the center of the target icon, thereafter staying within this range. It was by definition equal to or later than T_{near}. It corresponded to the best estimate of when the matcher reached a meaning hypothesis that turned out to be correct.[3] T_{pause} was the moment when the matcher finally stopped moving, just before parking his icon (T_{final}). This is where, presumably, he had the opportunity to perform any final processing or checking before concluding that he and the director had the same target location in mind.

In figure 4.1, the matcher's icon began about *300* pixels away from the target (T_{start}). After about a second and a half (T_{move}), he moved very slightly away and then rapidly toward the target. He passed within 20 pixels of the target (T_{near}) and then moved right by it. About 2 seconds later, he came back within 20 pixels of the target ($T_{reliable}$). At this point there is an abrupt "elbow"—that is, a change in slope from steep to nearly flat. After this point, progress toward the target is more gradual, and the plot shows a long, flat "tail." T_{pause} represents our closest estimate for the point in which he reached his final meaning hypothesis, and the time between T_{pause} and T_{final} indicates how long it took to conclude hypothesis testing and complete the trial.

Language Transcripts

Six of the twelve pairs of subjects were chosen at random and their conversation was transcribed.[4] This yielded a corpus of 480 conversational interchanges about the eighty map locations. These language transcripts were coded for level of description of the target location (general, specific, and detailed), acknowledgments, deictic cues, questions, and instances where speakers truncated their own utterances or interrupted their partner's.

Language-Action Transcripts

During each trial, there was a .5-second audible beep when the matcher clicked to make his icon movable and a .25-second audible beep when he clicked to park it. The beeps were marked in the language transcripts with #'s; the first # indicated where in the speech the beep began, and the second #, where it ended. The beeps were used to synchronize[5] the language transcripts with the action transcripts for a subset of forty-eight trials performed by the six pairs whose speech was transcribed. These forty-eight trials included eight different map locations, half on each map and half in each evidence condition. Within these constraints, the forty-eight trials were chosen randomly. For these trials, language-action transcripts were generated, where superscripts on the language transcripts correspond to numbered labels on the time-distance plots.

Findings

Next 1 will discuss the findings[6] with respect to each of the four predictions.

Early vs. Late Meaning Hypotheses

The contribution model, with its presentation and acceptance phases, led to the expectation that the distance between the matcher's and the director's icons would not decrease at a steady rate over time, but relatively rapidly at first (as they established a description from which the matcher could derive a meaning hypothesis), followed by at least one elbow in the time-distance plot (an estimate of the point during which the matcher had formed a promising meaning hypothesis) and then a relatively horizontal phase (during which the description continued to be grounded—that is, the matcher's meaning hypothesis was tested and accepted). As expected, the time-distance plots in the verbal-only evidence condition at the left of figures 4.2 and 4.3 show this pattern. Both T_{near}, the point at which the matcher's icon managed to approach the target closely enough to touch it, and $T_{reliable}$, the point at which the matcher appeared to have a reliable hypothesis (one that not only turned out to be correct, but after which any further approach to the target was monotonic), occurred relatively early, especially in the verbal-only evidence condition. This pattern contrasted with the possibility of a more gradual progression to the target that would have been expected if grounding did not occur—that is, if the matcher's icon were to have reached the correct location just before the end of the exchange, without an apparent acceptance phase before continuing to the next trial.

Impact of Visual vs. Verbal Evidence on Grounding

Elsewhere we have predicted (without systematically testing) that the evidence available for grounding shapes both the conversations and task performance (Clark and Brennan 1991). Consistent with that prediction, this task was most efficient with visual evidence. A trial, measured from the matcher's first icon movement (T_{move}) to when he parked (T_{final}), was more than twice as long with verbal-only evidence than with visual evidence, 20.68 seconds to 9.03 seconds, $min\ F'(1,35) = 125.86$, $p < .001$.[7] Consistent with this finding, fewer than half as many words were spoken in the visual condition as in the verbal-only condition (for the 480 transcribed trials, $min\ F'(1,13) = 45.51$, $p < ,001$). It is not surprising that a spatial task proceeds faster when there is visual evidence available; what is of particular interest, however, is that this evidence did not facilitate convergence evenly over the whole time course of a trial.

There was a striking qualitative difference in the shape of the time-distance plots in the two evidence conditions; compare the left-hand versus right-hand plots in figures 4.2 and 4.3. In the verbal-only condition, after the matcher reached the target location the plots showed a relatively long, nearly horizontal tail before he parked his icon, whereas in the visual condition, trials ended shortly after he reached the target. This

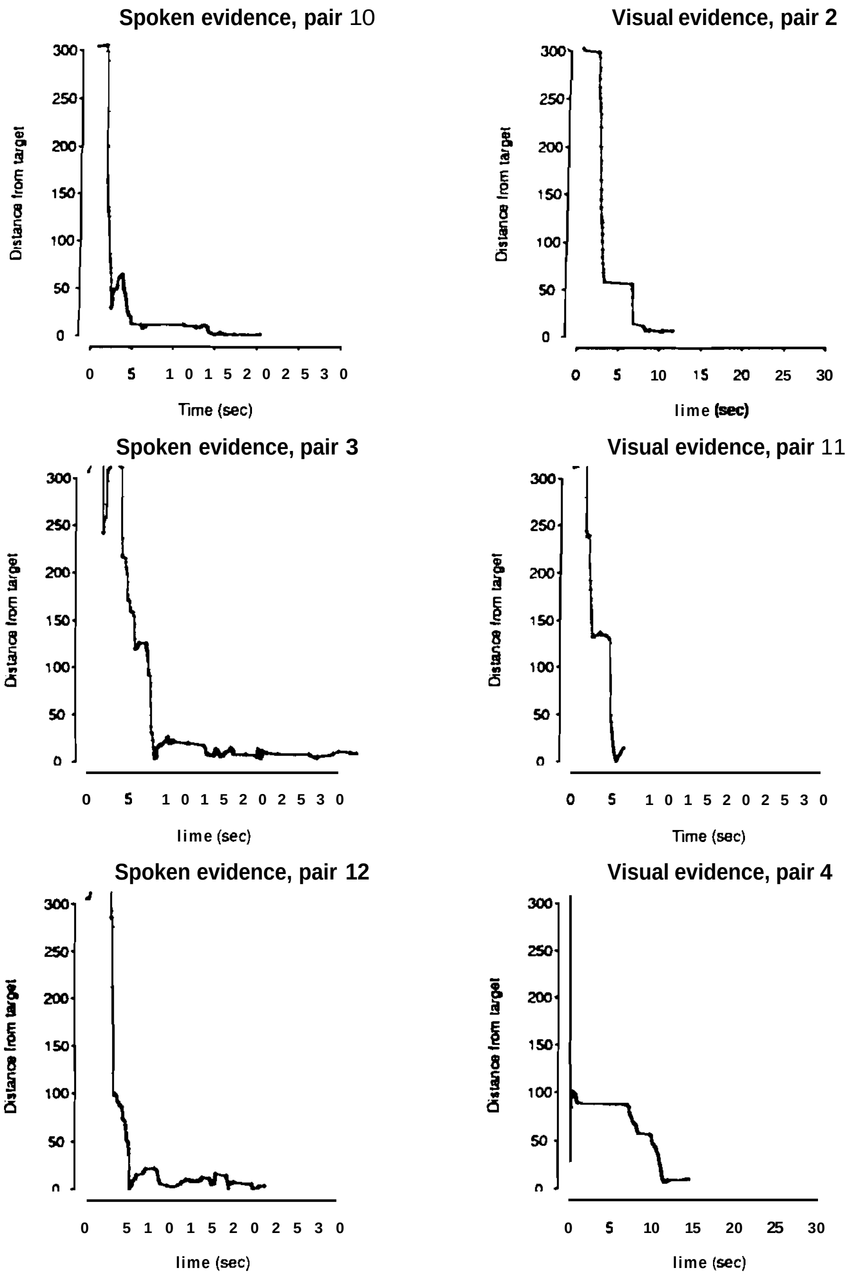


Figure 4.2

Time-distance plots for six matchers' progress toward the same target on the unfamiliar (Cape Cod) map. Those on the left were generated in the verbal-only evidence condition, and those on the right, in the visual evidence condition.

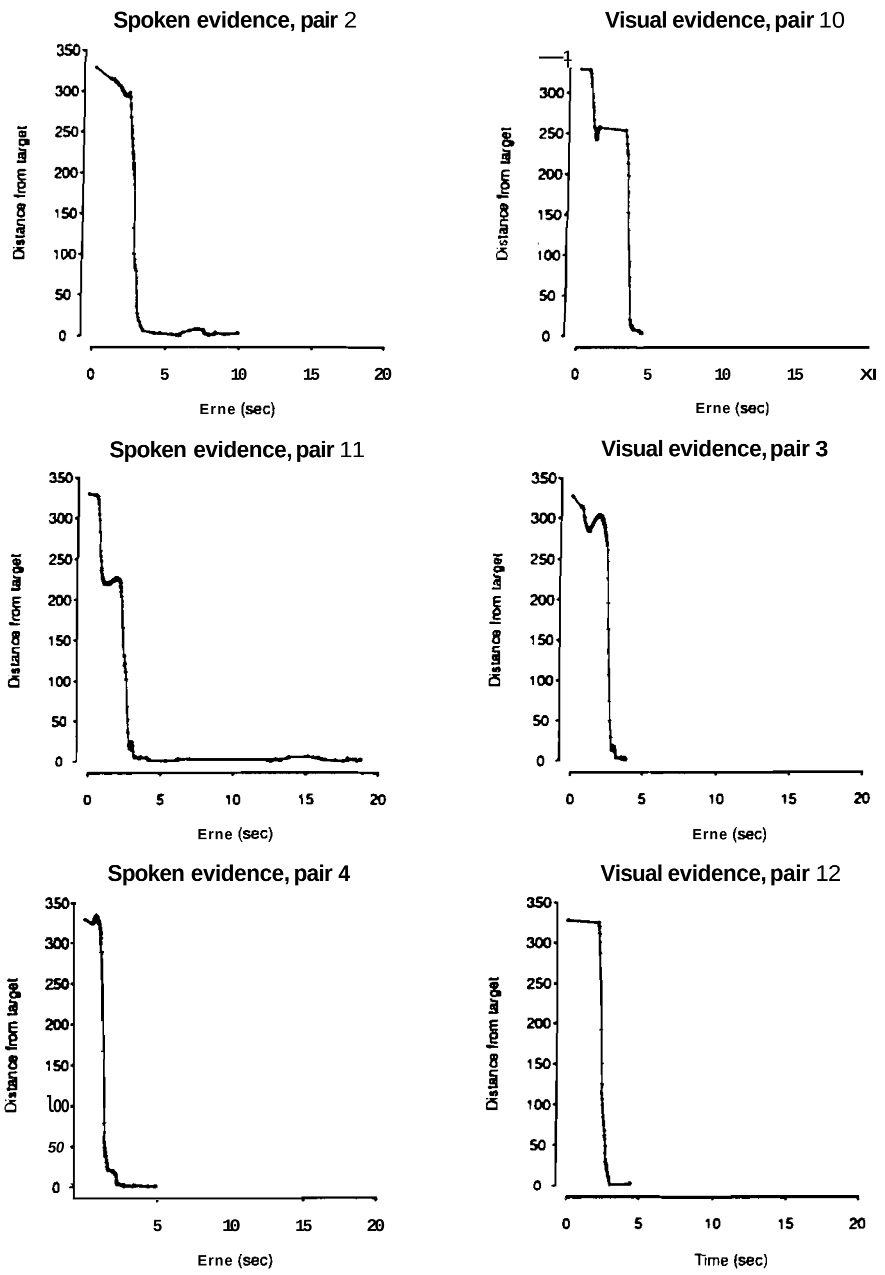


Figure 4.3
Time-distance plots for six matchers' progress toward the same target on the familiar (Stanford) map. Those on the left were generated in the verbal-only evidence condition, and those on the right, in the visual evidence condition.

shows that it took the matcher less time to conclude he was correct when the director had visual evidence. Consider the interval after the elbow, between $T_{reliable}$ and T_{final}. This is the interval in which the matcher already held and was testing the hypothesis that he ultimately accepted. For the randomly chosen subset of forty-eight trials, it took the matcher nearly four times longer to conclude that his location at $T_{reliable}$ was correct with verbal-only evidence alone than with visual evidence, 8.40 to 2.12 seconds, $min\,F'(1,17) = 14.37$, $p < .005$.

The interval from $T_{reliable}$ to T_{final} was where I predicted the director and matcher would reciprocally test their final meaning hypotheses; if this is so, then in the verbal-only evidence condition, this interval should include speech by not only the director, but also the matcher. Alternatively perhaps the director would just speak until she finished what she thought was an appropriate description and then wait while the matcher silently searched for the target (testing his hypothesis autonomously) before he eventually parked his icon. For the random subset of forty-eight trials, I examined the number of words spoken by the matcher from $T_{reliable}$ (the moment when the matcher reliably reached the target location without moving away again) to T_{final}. Matchers uttered a mean of 6.33 words from $T_{reliable}$ to T_{final} in the verbal-only evidence condition (different from zero at $t(12) = 5.28$, $p < ,001$). So in the verbal-only condition, the matcher was far from silent after reaching the target, just before parking his icon. In the visual condition, matchers took less responsibility, uttering only 2.13 words in the interval after $T_{reliable}$.

Even a highly accurate hypothesis at $T_{reliable}$ often did not stay *exactly* the same but got refined as the matcher made the final decision to park his icon. So it made sense to examine the impact of visual evidence after that, near the very end of a trial. In the contribution model, this point would correspond approximately to the end of the acceptance phase of a contribution. Consider the interval between T_{pause} and T_{final}, when (by definition) the matcher was motionless just before parking his icon. Here, the matcher could provide a final acknowledgement to the director, or perhaps even seek a final bit of evidence. This interval should be shorter in the visual condition than in the verbal-only condition because partners could use visual evidence for grounding. Indeed, it was less than half as long in the visual condition as in the verbal-only condition, 1.22 seconds to 3.00 seconds, $min\,F'(1,45) = 17.30$, $p < .001$. As it turns out, the matcher was far from silent during this interval in the verbal-only condition, uttering three times as many words as in the visual condition, 3.83 to 1.08.

Another way to look at this is by counting how many trials ended with verbal acknowledgments by the matcher. Clark and Wilkes-Gibbs's (1986) principle of mutual responsibility leads to the prediction that when directors do not have visual evidence,

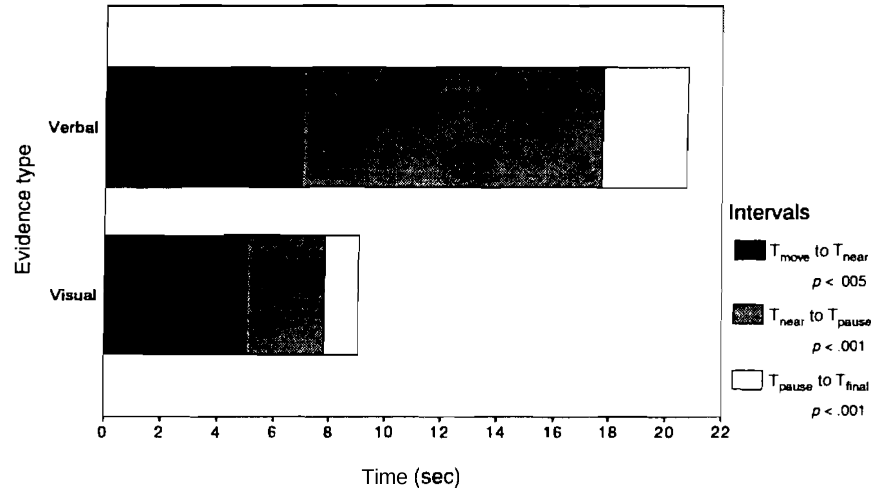


Figure 4.4
Effects of evidence on the time course of the action transcripts (955 trials).

matchers should give verbal acknowledgments, and when directors have visual evidence, matchers should withhold them. Every one of the verbal-only trials in the sample of forty-eight contained an acknowledgment by the matcher immediately before, during, or immediately after the moment he parked his icon, whereas only 29 percent of the ones in the visual condition contained such an acknowledgment, $\mathit{min}\,F'(1,17) = 32.00$, $p < .001$. That matchers adapted their own acknowledgments according to the perceptual evidence available to ***their*** partners supports the idea that backchannels are precise signals used for grounding (Brennan 1990), as opposed to responses emitted automatically or at a particular rate (as suggested by Duncan (1973) and Rosenfeld (1987)).

The type of evidence also affected the time course of the matcher's understanding during the presentation of the target-location description early in the trial (although not as dramatically as it did at the end of the trial, in the acceptance phase). With visual evidence, it took 5.0 seconds for the matcher's icon to first arrive within 20 pixels of the target (the interval from T_{move} to T_{near}); without visual evidence, it took 6.9 seconds, $\mathit{min}\,F'(1,42) = 9.48$, $p < .005$. The effect of the type of evidence on these intervals is summarized in figure 4.4.

How do these effects of evidence arise? Although the ultimate responsibility for parking the icon and ending the trial rested with the matcher (since the next trial could not be initiated until he parked his icon), the director was the one who took the responsibility for proposing to the matcher that his hypothesis was correct whenever she could see his icon. Sometimes the director did this by cutting herself off in mid-presentation as soon as she could see that the matcher had arrived:

D: ok,
now we're gonna go over to
M-Memorial Church?
and park right in Memor-
right there.
that's good.

The corresponding language-action transcript in figure 4.5 confirms that this matcher had just arrived at the target and stopped there at the moment the director cut herself off at "Memor-." In the smaller sample of forty-eight trials, twenty-two of the twenty-four trials in the visual condition ended with a deictic cue from the director, such as "You're there," which the matcher acknowledged simply by clicking to park his icon. Then the director initiated the next trial.

Impact of Familiar vs. Unfamiliar Maps on Grounding

Consider the familiar and unfamiliar maps. When the director and matcher mutually believed they were members of the same relevant community (in this case, Stanford), it should have taken them less time to establish a referent because they could make assumptions about each other's knowledge and build on their common ground. Indeed, the mean time it took to complete a trial was less for the Stanford map than for the Cape Cod map, 13.00 seconds to 16.71 seconds ($min\,F'(1, 100) = 14.97$, $p < .001$). As predicted, this difference was particularly strong very early in the trial, in the interval from T_{move} to T_{near}. The matcher got within 20 pixels of the correct target 2.26 seconds faster on the Stanford map than on the Cape Cod map ($min\,F'(1, 101) = 14.36$, $p < .001$). So most of the timing advantage of discussing a Stanford location happened within the first 5 seconds, as the director designed and presented a description. In contrast, the type of map made no difference at all at the end of each trial in the interval between T_{pause} and T_{final} ($min\,F'(1, 61) = 0.36$, n.~.)The moment-by-moment effects of maps on these intervals is summarized in figure 4.6.

The early advantage with the familiar (Stanford) map appears to have been due to the director's ability to present intelligible definite references earlier than with the unfamiliar (Cape Cod) map. Whenever the director began to describe a target location to a matcher, she had to choose a level of detail for the description, a choice that may

D: #ok[0],#
now we're gonna go over to
M-Memo[1]rial Church?[2]
and park [3]ri[illegible]in Manor-
right there.
that's *good.*

M: *that's* [5]rude,
to #park# in the [6]church.

D: hhhhhhheh heh

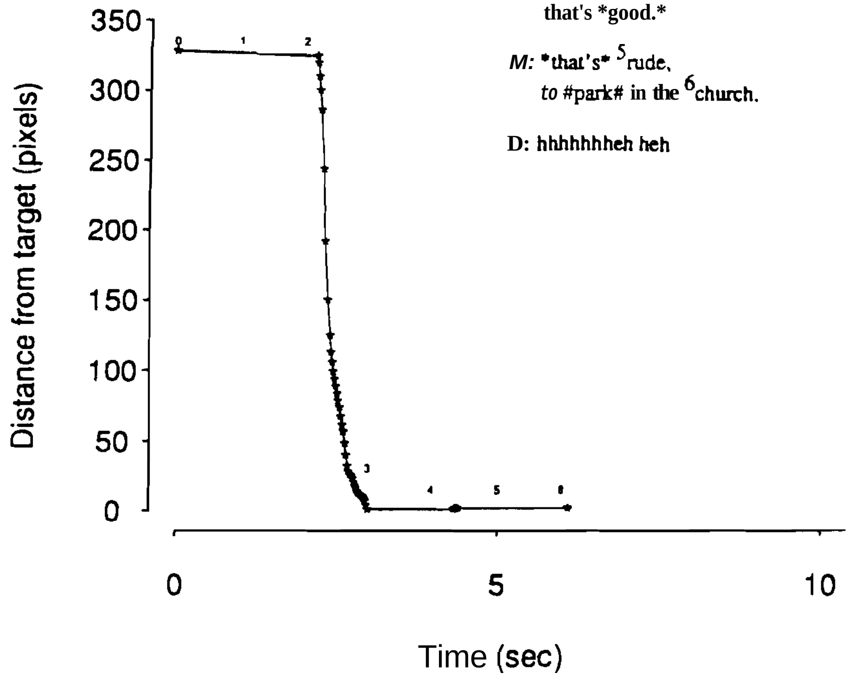


Figure 4.5

Using a deictic cue (right there) in the visual evidence condition to propose that hypotheses have converged. Superscripts in text correspond to points directly below numbers on the plot. Audible beeps at beginning and end of trial are represented within ##s. Stretches of overlapping speech are marked by pairs of **s.

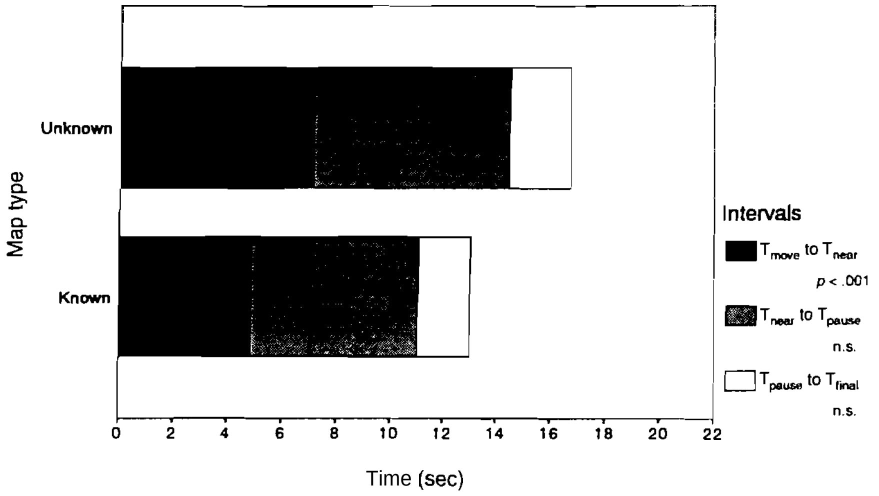


Figure 4.6
Mean effects of map on the time course of the action transcripts (955 trials).

have been guided by her estimation of the knowledge she shared with the matcher. The sooner she was able to present a detailed description, the sooner the matcher could form a good hypothesis about the target and move there. Consider this initial description of a target on the Cape Cod map (the levels of description are categorized at the right):

D:	uhh	
	go northeast	(general level)
	up to the comer,	(definite level)
	there's a little tiny street that has a three-letter name	(detailed level)

This presentation starts with a general directive, then increases in detail with a definite description and then a description of a more detailed landmark, which may have helped the matcher "zoom in" on the target. This strategy was common. In the sample of forty-eight trials, directors began with descriptions that were more specific for Stanford than for Cape Cod locations (in a sign test of the median, $p < .05$).

In contrast to map type, the evidence type a director had made no difference in the initial level of description, but it did affect subsequent descriptions. Closer inspection of the forty-eight individual language-action transcripts showed that directors "zoomed out" again when they had visual evidence that matchers were having trouble. Consider this description of a location on the Stanford map in the visual evidence condition:

D:	we're movin:g	
	south,	(*general level*)
	we're in Mem Chu,	(*definite level*)
	right in the center of Mem Chu,	(*detailed level*)
	which is right on the Quad,	(*definite level*)
	right there.	(*dcictic*)
	stop.	

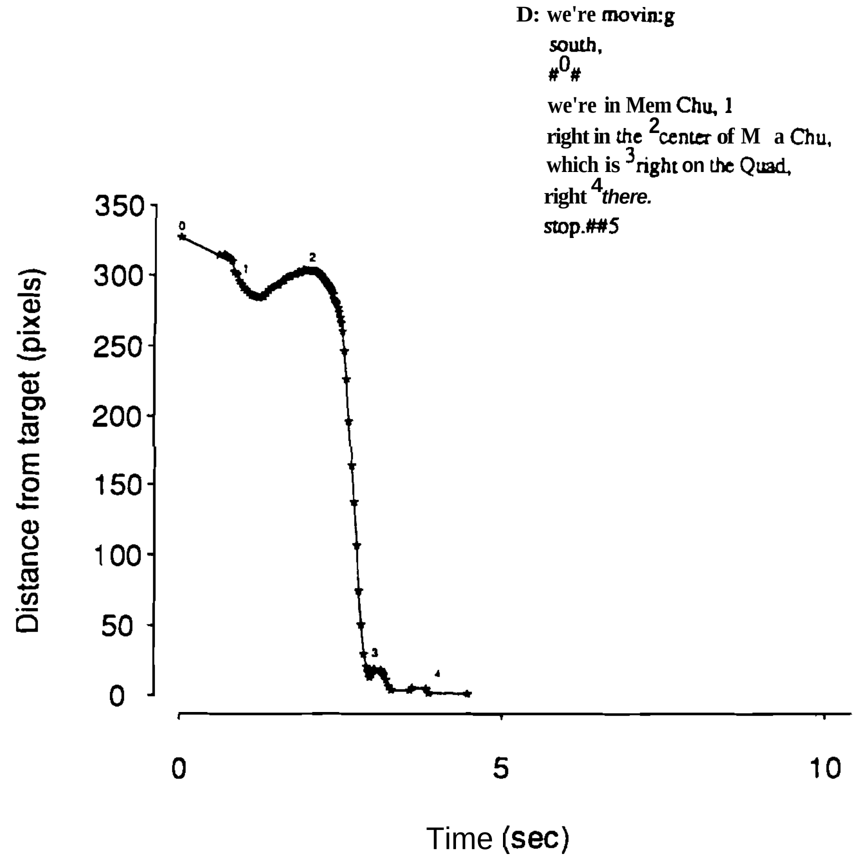
Here, the director got to a relatively high level of detail with "right in the center of Mem Chu" and then zoomed out or backed off a level by mentioning "the Quad" as a landmark relative to Memorial Church. At the moment the detailed description was uttered, the matcher had actually just gone past the target (see figure 4.7); the director may have changed her level of description in response to this direct evidence about the matcher's understanding. Finally, the director let the matcher know when his icon reached the target with a deictic cue.

Impact of Evidence on How Precisely Speakers' and Addressees' Hypotheses Converged

As predicted, the grounding process involves satisficing; directors and matchers actually got their hypotheses to converge more closely when the director *lucked* visual evidence of what the matcher was doing. Without visual evidence, matchers were, on average, 4.6 pixels off from dead center on the target, whereas with visual evidence, they were 5.6 pixels off ($minF'(1,73) = 4.00$, $p < .05$). This supports the prediction from the contribution model that people set their grounding criteria to ensure the desired degree of convergence, rather than the simple intuition that more evidence yields more convergence (see Karsenty 1999). How *precisely* two people could set their grounding criteria depended on how well they could use the available evidence. With visual evidence, they were only as accurate as they needed to be. Without visual evidence, they needed to adopt a higher criterion in order to be sure to reach an equivalent level of performance, and this took more effort overall.

What is remarkable is how flexible each pair was able to be, adjusting their grounding criteria whenever the evidence available to the director changed.

There were no more errors without visual evidence than with visual evidence; performance in both evidence conditions was nearly at ceiling. An error was counted whenever a pair failed to reach the criterion described in the task instructions — that is,



D: we're movin:g
south,
#^{0}#
we're in Mem Chu, 1
right in the ^{2}center of M a Chu,
which is ^{3}right on the Quad,
right ^{4}*there.*
stop.##5

Figure 4.7

Levels of description, synchronized with the matcher's icon movements (visual condition). Just before point 3, the matcher goes past the target and the director responds by issuing a more general locative.

whenever the matcher's parked icon would not have overlapped the director's, if the two screens were superimposed. There were only 11 errors in 955 trials, made by 9 different matchers and distributed evenly across the verbal-only condition (5 errors) and the visual condition (6 errors). So even though they were doing a spatial task in which visual evidence was highly relevant, people did just as well at reaching the task criterion when directors could not see matchers' icons as when they could; it just took more than twice as long when the director lacked visual evidence.

Discussion

These data serve two broad purposes. First, they exemplify a technique—earlier than eye tracking and more precise than the videotaped card-matching task used by Schober and Clark 1989—for measuring, moment by moment, what people understand and intend during task-oriented conversation (see also Clark and Krych 2002 for a later version of a task that generated a language-action transcript). This sort of evidence aims to combine the control and reliability valued by researchers in the language-as-product tradition with the ecological validity valued by those in the language-as-action tradition. Both are necessary in order to understand the intra- and interpersonal processing that underlies human language ability.

Second, and more specifically, these data are consistent with the contribution model and shed light on the time course by which two individuals in conversation coordinate their processing. The shapes of the language-action transcripts showed, on average, a distinct phase between the time when a matcher first had a correct meaning hypothesis and the time when he accepted it by ending the trial. This grounding or acceptance phase was shortened considerably when the director had visual evidence of the matcher's progress toward the target. The documentation of a distinct phase after the matcher arrives at an adequate meaning hypothesis suggests that the mental representations of two individuals in conversation do not converge simply because of similarity between the individuals or between the processes of speaking and listening.

These findings also confirm that backchannels are specific, timely signals for coordinating individuals' mental states, as opposed to habitually or automatically emitted signals showing a general level of engagement in the conversation. Even though what the matcher saw on his display was exactly the same in both evidence conditions, he provided verbal acknowledgments when the director could not see his icon movements and withheld them when she could. He rapidly adapted this behavior as the experimental conditions alternated between visual and verbal-only evidence.

The difference between the two evidence conditions concerning which of the two partners took responsibility for concluding that their individual hypotheses had converged is consistent with Clark and Wilkes-Gibbs's (1986) principle of mutual responsibility. When the only possible evidence from the matcher was in spoken form, the director typically relied on the matcher's judgment that their hypotheses had converged. With ostensive visual evidence from the matcher, the director could judge the likelihood of convergence for herself. The responsibility for concluding that their hypotheses had converged fell to whoever was in a position to have the strongest evidence. In a task-based conversation such as this one, seeing the matcher's icon counts as strong evidence, and so the director in the visual condition took on most of that responsibility. This saved not only the matcher's effort, but their collective effort as well. The distribution of responsibility between partners in conversation turned out to be quite flexible.

Trials with visual evidence and with the familiar map were fastest, but these advantages were distributed quite differently over the course of a trial. The mutually familiar Stanford map led to an advantage early in the trial, while the bulk of the descriptions of the target locations were being presented, but not late in the trial. In contrast, visual evidence led to an advantage throughout the trial that was especially large late in the trial, during the phase in which the descriptions were being grounded.

It is striking that so many of the exchanges in the visual evidence condition contained no speech at all by the matcher. In these cases, what the matcher did with icon movements substituted for what he did with speaking turns in the verbal-only condition. Obviously, icon moves are instrumental to doing the task. So did the matcher really intend certain icon moves to function, in addition, like utterances? That is, did he use ostensive actions deliberately, intending the director to recognize this? Ostensive evidence can function in two ways: as a mere *symptom* of what was understood (as "natural" evidence in the sense that "smoke means fire") or *intentionally* where one person expects the other to recognize it as communicative, in the sense of nonnatural meaning or *meaning*$_{nn}$ (Grice 1957, 1989). When the matcher and the director were mutually aware that the director could see the matcher's icon, the matcher adapted by withholding verbal acknowledgments, and his icon move counted as his turn or presentation in the conversation. This enabled the matcher's presentations (in the form of icon moves) to overlap the director's (in the form of verbal descriptions), so that they could ground continually instead of by discrete turns, coordinating their meaning hypotheses in finer increments.

Consider the interchange in figure 4.8, during which the director presented a description in many installments separated by pauses (indicated by new lines). As we see

#^{0}# ^{1}

D: ok
now
go^{2} down ^{3}
to about
go^{4} down about athird of the way,
an:d^{6}
^{7}right
to the ^{8}edge.
to ^{9}Terman Engineering.
^{10}right in the corner
in the ^{11}inside comer,
^{12}right next to the ^{13}number five
in the ^{14}insider comer of that ^{1}'building.

M: *there?*

D: ^{16}*go down.* ^{17}
ok, ^{18}there.
^{19}

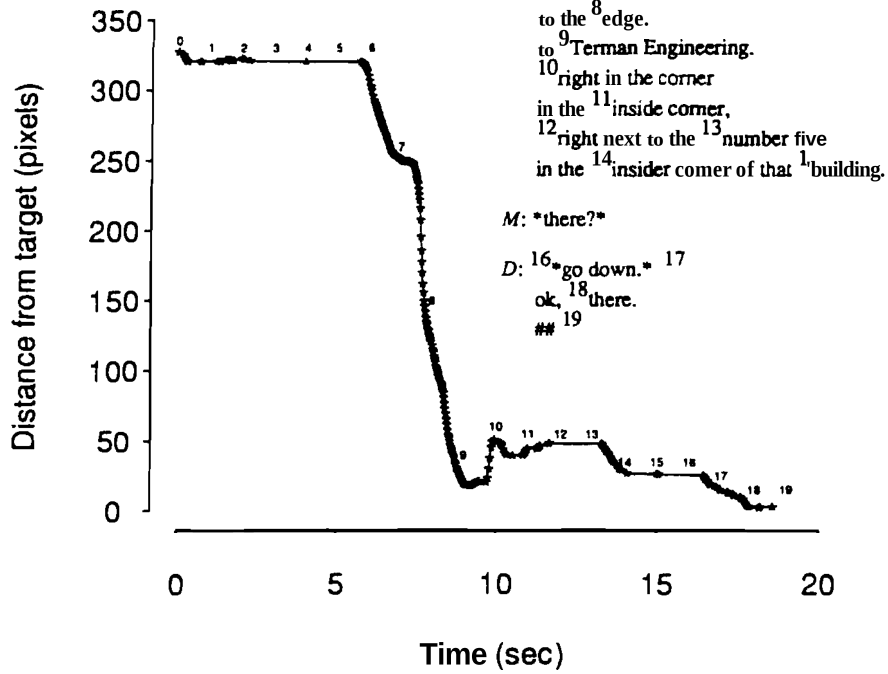


Figure 4.8
Repairing a matcher's incorrect meaning hypothesis between points 14–18 (visual evidence condition).

from figure 4.8, the matcher reached a location about 30 pixels above the target after 14 seconds and stopped. The director did not respond explicitly to this evidence, but kept on presenting installments. The matcher did not say anything at first, seemingly waiting for the possibly inattentive director to stop talking and notice his icon. After the director produced quite a few descriptive phrases, there was a pause. Then the matcher made an explicit verbal proposal that their hypotheses had converged: "there?" (superscripts 15–16). This suggests that he may have originally intended for the director to accept his icon position as an intentional presentation. As it turned out, the director was not being inattentive; the matcher's hypothesis was simply not close enough. It appears that the director had been responding to the incorrect icon position by continuing to present descriptions. When the matcher finally demanded to know whether his position was correct, the director opted, at the same time, to provide explicit deictic evidence in the form of an overlapping utterance, "go down—ok, there." This exchange demonstrates that collaborative hypothesis testing requires positive evidence (of convergence) as well as negative evidence (of misunderstanding). A lack of response will not do when positive evidence is expected.

It particularly interesting that directors were so attuned to the evidence of understanding from matchers that they sometimes interrupted themselves midword or mid-constituent in order to provide a timely deictic cue (as in the exchange in figure 4.5). We are currently investigating such self-interruptions by directors who can monitor matchers' eye movements in a referential communication task, in order to understand how closely the processes of speech planning and monitoring may be attuned to feedback from addressees (Brennan and Lockridge 2004).

Tracking Eye-Gaze vs. Mouse Movements in a Referential Communication Task

The online measure developed for this study presents both advantages and disadvantages in comparison to head-mounted eye tracking, an important new technique in the study of spoken-language use (see, e.g., Metzing and Brennan 2003 and chapters 1, 3, 5, and 6 in this volume). The usefulness of eye tracking is based on the "eye-mind assumption"—that the object a person is looking at reflects what is being processed at that moment. Eye movements to locations in space are ordinarily accompanied by shifts in attention to those locations (Deubel and Schneider 1996; Hoffman and Subramaniam 1995; Kowler et al. 1995); however, the reverse is not necessarily true—one may shift attention without shifting gaze (Posner 1980).

For studying online processing in conversation, how does eye tracking compare to the spatial technique used in this study? Tracking eye gaze is obviously far more temporally precise as an indicator of mental processes because the time to plan and launch a saccade is shorter than the time to plan and execute the motor movements for positioning a cursor with a mouse. Unlike mouse movements, eye movements need not be intentional. Eye tracking is also more spatially precise than mouse tracking; note that in my task, when the mouse was still (reflected in the time-distance plots as a perfectly flat horizontal line), the matcher may have been searching the display for the referent of the director's description, and this evidence is not recorded in the action transcripts.

On the other hand, the precision afforded by eye tracking may actually present disadvantages for answering the sorts of questions 1 have asked here. Not all eye fixations represent a hypothesis about what a speaker is referring to in a task like this one. People often make irrelevant saccades, as well as fixations whose purpose it is to take in information while searching for a potential referent. The map displays in my task were complex enough that they were not completely encoded by matchers before the task, so any eye fixations to a location could have represented information gathering about what was in the display rather than a likely hypothesis about its status as a potential referent. In this spatial task, mouse movements were instrumental in dragging the icon to the target, and so these movements were actually more directly coupled to the matcher's intentions and hypotheses than eye movements would have been. The matcher's mouse movements indicated at least some level of confidence that he had narrowed down his hypothesis about the target location. For a location-finding task such as this one, the language-action transcripts provide a useful window into the time course by which two people's beliefs converge.

Implications for Mediated Communication

Although theories of discourse structure have focused mainly on linguistic utterances, visual evidence can be as much a part of discourse as can verbal evidence. As Grice (1957, 388) argued, "linguistic intentions are very like nonlinguistic intentions." A theory of discourse should be able to account not only for a conversation's linguistic structure, but also for its visually presented elements. Of particular interest in the current study was how conversations in the two evidence conditions differed in their turn-taking structure. Within the visual condition, grounding did not have to be done in discrete verbal turns, but was done continuously and in parallel, because the expression of one partner's hypothesis provided the other partner's evidence. People in this experiment appeared to adjust quite flexibly and rapidly to the degree of perceptual copresence they had. Similarly, in another study of remote communication using a shared electronic whiteboard, we found that utterances produced by typing were sometimes presented and accepted in parallel, and that when they lacked spoken evidence and had *only* visual evidence, people used spatial rather than temporal contiguity to ground utterances (Whittaker, Brennan, and Clark 1991).

Understanding the process of grounding in the detail presented here enables us to better predict how a medium will shape conversation and collaboration. Many studies have described differences in tasks conducted over different media without any theoretical framework to explain these differences (e.g., Chapanis et al. 1972; Cohen 1984; Ochsman and Chapanis 1974; Williams 1977). For instance, Cohen studied telephone and keyboard conversations in which one person directed another to assemble a pump. On the telephone, people would first get their partners to identify a part, and only then tell them what to do with it, whereas with keyboards, they would do all this in a single turn. He concluded that "speakers attempt to achieve more detailed goals in giving instructions than do users of keyboards" (Cohen 1984, 97). The reason this should be so follows logically from the grounding framework, in which people make different trade-offs in different media in order to minimize their collective effort (Clark and Brennan 1991). Acquiring evidence about an addressee's understanding is less costly in spoken conversation than with text messages (which lack prosodic cues and take more time and effort to produce), so speakers take more frequent turns and ground smaller constituents than do typists in chat conversations. Also, the fact that speech is ephemeral and text is not makes grounding larger constituents easier for text messages and grounding small constituents more cost-effective for speech, particularly when there is visual copresence between partners.

In closing, when people are physically copresent, actions can stand in for utterances, and one partner's feedback and task-related actions can affect another's utterance planning, moment by moment. In this way, the techniques that a medium affords for grounding shape both the products and the processes of spontaneous language use.

Notes

This chapter presents previously unpublished data from Brennan 1990. I thank Herbert Clark for his insights at all stages of this project and many other colleagues, including Richard Gerrig, Michael Schober, Joy Hanna, and Eric Stubbs, for helpful input on earlier drafts. Thanks also to Stephen Lowder for software assistance, to Kathryn Henniss for coding assistance, and to Caroline Collins, Martin Kay, and Peter Sells for loaning their equipment. This material is based on work supported by the National Science Foundation under grants 0082602, 9980013, and 9202458. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation. Address correspondence to Susan Brennan at the Department of Psychology, State University of New York, Stony Brook, NY 11794-2500.

1. Understanding is not the same as agreement or uptake. When speakers and addressees have incompatible intentions, they might understand one another perfectly well but "agree to disagree" (see Bly 1993). The task I present here provides speakers and addressees with a shared goal, so this study focuses on cases where speakers and addressees with shared goals attempt to make their hypotheses converge.

2. For expository convenience, I will refer to the director as female and the matcher as male, even though subjects were run in single-sex pain and they switched director/matcher roles halfway through the session.

3. Note that these points provide more detail about the time course of a contribution than does Schober and Clark's (1989) "conjecture point". A particular "conjecture point" could ambiguously correspond to when the matcher first reached a correct hypothesis (T_{near}), when the matcher reli-***ably*** reached a correct hypothesis ($T_{reliable}$), or when the matcher reached a ***final*** hypothesis ($T_{completion}$).

4. Speech was transcribed in segments that corresponded roughly to one phonemic clause per line—that is, a short sequence of words separated by a pause, and generally containing one primary pitch accent (Rosenfeld 1987; see also Boomer 1978; Dittman and Llewellyn 1967). Each line was punctuated according to its clause-final prosody: "." for final pitch lowering, "?" for final rising, "," for the end of a tone unit (if midclause) or else for listlike intonation (when at the end of a clause), "-" for a sudden self-cutoff on a level pitch, and no punctuation for level pitch. Slowed speech or drawled syllables were denoted by ":" following the letter that most closely matched the sound being drawn out (ye:s for "yeeees," versus yes: for "yesss"). Overlapping speech was transcribed using single or double asterisks to enclose the beginning and ending of the simultaneous talk. Unintelligible speech was enclosed in "⟨ ⟩". All transcripts were checked for accuracy.

5. A linguistics graduate student naive to the experimental hypotheses listened to the videotapes of each trial with the language transcripts in front of her, using a videotape player equipped with a counter and a shuttle knob. For each tial she zeroed the counter at the start of the initial beep and recorded an integer at every 1.0-second interval over the text version of its corresponding spoken syllable on the transcript. It took many passes over the tapes to record these intervals and to check the synchronization of each trial.

6. Five action trials were eliminated because matchers inadvertently clicked twice while picking up the icon, inadvertently parking it and ending the tial early. The results presented here are based on the 955 automatically logged trials, unless otherwise stated. There was more variability in elapsed time for trials with verbal-only evidence alone than with visual evidence added (in a standard ratio-of-variances test, $F(2,78) = 5.81$, $p < .01$), so I transformed all time-interval lengths as a function of log(time). I then analyzed the transformed data using two-way ANOVAs with map and evidence condition as fixed factors, treated pairs of subjects and items (map locations) as random factors, and computed the statistic *min* F' as recommended by Clark (1973). There were no interactions between map and evidence.

7. *Min F′* combines into a single statistic the by-subjects and by-items ANOVAS (F_1 and F_2) traditionally reported in psycholinguistics and memory studies, but in a more conservative fashion (if a result is significant by *min F′*, then it is significant by both F_1 and F_2). The degrees of freedom are recalculated as a combination of those for F_1 and F_2 (see Clark 1973).

References

Berkovits, R. 1981. Are spoken surface structure ambiguities perceptually unambiguous? *Journal of Psycholinguistic Research, 10,* 41–56.

Bly, B. 1993. Uncooperative Language and the Negotiation of Meaning. Unpublished doctoral dissertation, Stanford University.

Boomer, D. S. 1978. The phonemic clause: Speech unit in human communication. In A. W. Siegman and S. Feldstein, eds., *Nonverbal Behavior and Communication.* Hillsdale, NJ: Erlbaum.

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

Brennan, S. E. 1990. Seeking and Providing Evidence for Mutual Understanding. Unpublished doctoral dissertation, Stanford University.

Brennan, S. E., and Clark, H. H. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory and Cognition, 6,* 1482–1493.

Brennan, S. E., and Lockridge, C. B. 2004. How visual copresence and joint attention shape speech planning. Unpublished manuscript.

Brown, P. M., and Dell, G. S. 1986. Adapting production to comprehension: The explicit mention of instruments. *Cognitive Psychology,* 19, 441–472.

Bruner, J. S. 1985. The role of interactive formats in language acquisition. In j. P. Forgas, ed., *Language and Social Situations,* 31–46. New York: Springer-Verlag.

Cahn, J. E., and Brennan, S. E. 1999. A psychological model of grounding and repair in dialog. *Proceedings, AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems,* 25–33. North Falmouth, MA: American Association for Artificial Intelligence.

Chapanis, A., Ochsman, R. B., Parrish, R. N., and Weeks, G. D. 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.

Clark, H. H. 1973. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12,* 335–359.

Clark, H. H. 1983. Making sense of nonce sense. In G. B. Flores d'Arcais and R. Jarvella, eds., *The Process of Language Understanding,* 297–331. New York: Wiley.

Clark, H. H. 1994. Managing problems in speaking. *Speech Communication, 15,* 243–250.

Clark, H. H. 1996. *Using Language.* Cambridge, MA: Cambridge University Press.

Clark, H. H., and Brennan, S. E. 1991. Grounding in communication. In L. B. Resnick, J. Levine, and S. D. Behrend, eds., *Perspectives on Socially Shared Cognition,* 127–149. Washington, DC: American Psychological Association Books.

Clark, H. H., and Gerrig, R. J. 1983. Understanding old words with new meanings. *Journal of Verbal Learning and Verbal Behavior, 22,* 591–608.

Clark, H. H., and Krych, M. A. 2004. Speaking while monitoring addressees for understanding. *Journal of Memory and Language, 50,* 62–81.

Clark, H. H., and Marshall, C. R. 1981. Definite reference and mutual knowledge. In A. K. Joshi, B. Webber, and I. A. Sag, eds., *Elements of Discourse Understanding,* 10–63. Cambridge, UK: Cambridge University Press.

Clark, H. H., and Schaefer, E. F. 1987. Collaborating on contributions to conversations. *Language and Cognitive Processes, 2,* 19–41.

Clark, H. H., and Schaefer, E. F. 1989. Contributing to discourse. *Cognitive Science, 13,* 259–294.

Clark, H. H., and Wilkes-Gibbs, D. 1986. Referring as a collaborative process. *Cognition,* 22, 1–39.

Cohen, P. R. 1984. The pragmatics of referring and the modality of communication. *Computational Linguistics, 10,* 97–146.

Cook, M., and Lalljee, M. 1972. Verbal substitutes for visual signals in interaction. *Snniotica, 6,* 212–221.

Dell, G. S., and Brown, P. 1991. Mechanisms for listener adaptation in language production. In D. J. Napoli and J. A. Kegl, eds., *Bridges between Psychology and Linguistics: A Swarthmore Festschrift for Lila Gleitman,* 105–129. Hillsdale, NY: Erlbaum.

Deubel, H., and Schneider, W. 1996. Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research, 36,* 1827–1837.

Deutsch, W., and Pechmann, T. 1982. Social interaction and the development of definite descriptions. *Cognition, 11,* 159–184.

Dittman, A. T., and Llewellyn, L. G. 1967. The phonemic clause as a unit of speech decoding. *Journal of Personality and Social Psychology, 6,* 341–349.

Duncan, S. D. 1973. Toward a grammar for dyadic conversation. *Semiotica,* 9, 29–47.

Duncan, S. 1975. Interaction units during speaking turns in dyadic, face-to-face conversations. In A. Kendon, R. M. Hams, and M. R. Key, eds., *Organization of Behavior in Face-to-Face Interaction.* The Hague: Mouton.

Goodwin, C. 1979. *Conversational Organization: Interaction between Speakers and Hearers.* New York: Academic Press.

Grice, H. P. **1957.** Meaning. *Philosophical Review*, 66, 377–388.

Grice, H. P. **1975.** Logic and conversation (from the William James lectures, Harvard University, **1967).** In P. Cole and J. Morgan, eds., *Syntax and Semantics 3: Speech Acts*, **41–58.** New York: Academic Press.

Grice, H. P. **1989.** Meaning revisited. In *Studies in the Way of Words*, **283–303.** Cambridge, MA: Harvard University Press.

Heath, C. **1984.** Talk and recipiency: Sequential organization in speech and body movement. In J. M. Atkinson and J. Heritage, eds., *Structures of Social Action*, 247–265. Cambridge, UK: Cambridge University Press.

Hess, L.. J., and Johnston, J. R **1988.** Acquisition of back channel listener responses to adequate messages. *Discourse Processes*, *11*, 319–335.

Hoffman, J., and Subramaniam, B **1995.** The role of visual attention in saccadic eye movements. *Perception and Psychophysics*, *57*, 787–795.

lsaacs, E. A., and 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 visual context: An empirical study of comprehension problems in side-by-side and remote help dialogues. *Human-Computer Interaction*, *14*, 283–315.

Kendon, A. **1970.** Movement coordination in social interaction: Some examples described. *Acta Psychologica*, *32*, **101–125.**

Kowler, E., Anderson, E., Dosher, B., and Blaser, E. **1995.** The role of attention in the programming of saccades. *Vision Research*, *35*, 1897–1916.

Krauss, R M. **1987.** The role of the listener: Addressee influences on message formulation. *Journal of Language and Social Psychology*, *6*, **81–98.**

Krauss, R. M., and Fussell, S. R **1991.** Constructing shared communicative environments. In L B. Resnick, J. Levine, and S. D. Behrend, eds., *Perspectives on Socially Shared Cognition*, 172–200. Washington, DC: American Psychological Association Books.

Krauss, R. M., and Glucksberg, S. **1969.** The development of communication: Competence as a function of age. *Child Development*, 40, **255–256.**

Krauss, R M., and Weinheimer, S. **1964.** Changes in reference phases as a function offrequency of usage in social interaction: A preliminary study. *Psychonomic Science*, *1*, **113–114.**

Krauss, R M., and Weinheimer, S. **1966.** Concurrent feedback, confirmation, and the encoding of referents in verbal communication. *Journal of Personality and Social Psychology*, *4*, 343–346.

Krauss, R M., and Weinheimer, S. **1967.** Effect of referent similarity and communication mode on verbal encoding. *Journal of Verbal Learning and Verbal Behavior*, *6*, 359–363.

Kraut, R., Lewis, S., and Swezey, L. 1982. Listener responsiveness and the coordination of conversation. *Journal of Personality and Social Psychology, 43,* 718–731.

Ochsman, R. B., and Chapanis, A. 1974. The effects of 10 communication modes on the behavior of teams during co-operative problem-solving. *International Journal of Man-Machine Studies, 6,* 579–619.

Posner, M. 1980. Orienting of attention. *Quarterly Journal of Experimental Psychology, 32,* 3–25.

Rosenfeld, H. M. 1987. Conversational control functions of nonverbal behavior. In A. W. Siegman and S. Feldstein, eds., *Nonverbal Behavior and Communication,* 563–601. Hillsdale, NJ: Erlbaum.

Rumelhart, D. E. 1980. *Understanding Understanding.* Technical Report No. 100. San Diego: Center for Human Information Processing, University of California.

Russell, A. W., and Schober, M. F. 1999. How beliefs about a partner's goals affect referring in goal-discrepant conversations. *Discourse Processes,* 27, 1–33.

Schober, M. F., and Clark, H. H. 1989. Understanding by addressees and overhearers. *Cognitive Psychology,* 21, 211–232.

Simon, H. 1981. *The Sciences of the Artificial* 2nd ed. Cambridge, MA: MIT Press

Sperber, D., and Wilson, D. 1986. *Relevance.* Cambridge, MA: Harvard University Press.

Svartvik, J., and Quirk, R. 1980. *A Corpus of English Conversation.* Lund, Sweden: Gleerup

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

Whittaker, S. J., Brennan, S. E., and Clark, H. H. 1991. Coordinating activity: An analysis of interaction in computer-supported cooperative work. *Proceedings, CHI '91, Human Factors in Computing Systems.* 361–367. New Orleans, LA: Addison Wesley.

Wiemann, J. M., and Knapp, M. L. 1975. Turn-taking in conversations. *Journal of Communication,* 25, 75–92.

Wilkes-Gibbs, D. 1986. Collaborative Processes of Language Use in Conversation. Unpublished doctoral dissertation, Stanford University.

Wilkes-Gibbs, D. 1992. Individual goals and collaborative actions: Conversation as collective behavior. Unpublished manuscript.

Wilkes-Gibbs, D., and Clark, H. H. 1992. Coordinating beliefs in conversation. *Journal of Memory and Language,* 31, 183–194.

Williams, E. 1977. Experimental comparisons of face-to-face and mediated communication. *Psychological Bulletin,* 84, 963–976.

Yngve, V. H. 1970. On getting a word in edgewise. In *Papers from the Sixth Regional Meeting of the Chicago Linguistic Society,* 567–578. Chicago: Chicago Linguistic Society.