Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Memory and Language 50 (2004) 62–81
Journal of Memory and Language
www.elsevier.com/locate/jml

Speaking while monitoring addressees for understanding☆

Herbert H. Clark* and Meredyth A. Krych[1]

Department of Pyschology, Stanford University, Building 420, Jordan Hall, Stanford, CA 94305-2130, USA

Received 7 October 2001; revision received 11 August 2003

Abstract

Speakers monitor their own speech and, when they discover problems, make repairs. In the proposal examined here, speakers also monitor addressees for understanding and, when necessary, alter their utterances in progress. Addressees cooperate by displaying and signaling their understanding in progress. Pairs of participants were videotaped as a director instructed a builder in assembling 10 Lego models. In one group, directors could see the builders' workspace; in a second, they could not; in a third, they gave instructions by audiotape. Two partners were much slower when directors could not see the builders' workspace, and they made many more errors when the instructions were audiotaped. When their workspace was visible, builders communicated with directors by exhibiting, poising, pointing at, placing, and orienting blocks, and by eye gaze, head nods, and head shakes, all timed with precision. Directors often responded by altering their utterances midcourse, also timed with precision.

Keywords: Speaking; Language production; Dialogue; Monitoring; Gestures; Collaboration

Speaking and listening in dialogue have been viewed from two main perspectives. In *unilateral accounts*, speaking and listening are autonomous processes. Speakers determine the course of their utterances by themselves, and listeners try to understand those utterances on their own. In *bilateral accounts*, speaking and listening together form a joint activity. Speakers monitor not just their own actions, but those of their addressees, taking both into account as they speak. Addressees, in turn, try to keep speakers informed of their current state of understanding.

In this paper, we offer evidence for speaking as a bilateral process. We take our evidence from spontaneous dialogue. There people not only speak, but nod, smile, point, gaze at each other, and exhibit and place things. Gestural acts like these are often tied to what people are doing as they are talking. In the kitchen, people may point at utensils, show each other ingredients, and hand each other pots and pans. At the dinner table, they may point at salt shakers, pass food, and exhibit empty plates. It is the vocal and gestural acts together that comprise their talk, so both must be examined as evidence for how they speak.

The two accounts of speaking differ in what speakers monitor for. In unilateral accounts, speakers rely entirely on *self-monitoring*, whereas in bilateral accounts, they also rely on *other-monitoring*. If we assume that speaking is bilateral, what do speakers monitor others for, and how do they use that information in the course of speaking? These are the main questions addressed in this paper.

☆This research was supported in part by Grant N000140010660 from the Office of Naval Research. We are indebted to a host of colleagues for solicited and unsolicited advice on the research. We thank Adrian Bangerter, Eve V. Clark, Richard Gerrig, Zenzi Griffin, Anna Katz, Teenie Matlock, and Martin Pickering for comments on earlier versions of the paper.

* Corresponding author.

E-mail addresses: Clark@Stanford.edu (H.H. Clark), krychm@mail.montclair.edu (M.A. Krych).

[1] Present address. Department of Psychology. Montclair State University.

0749-596X/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2003.08.004

Speaking and listening in dialogue

Most accounts of language processing are implicitly unilateral. Models of production, for example, tend to focus on choosing messages, formulating expressions, and articulating those expressions, all treated as autonomous processes (see, e.g., Bock & Levelt, 1994; Ferreira, 2000; Garrett, 1980; Kempen & Hoenkamp, 1987; Levelt, 1989). Although speakers are known to monitor their own progress, making repairs when needed (Levelt, 1983; Schegloff, Jefferson, & Sacks, 1977), these models have no provision for monitoring addressees and using that information to change course on line. Models of listening, in turn, tend to focus on attending to, parsing, and interpreting utterances, also treated as autonomous processes (Clark & Haviland, 1977; Frazier & Clifton, 1996; Marslen-Wilson, 1987; Tanenhaus & Trueswell, 1995). These models have no provision for using that information to influence a speaker's current utterance. In truth, these models were not designed for dialogue, so it is not surprising they are unilateral.

Other models of language use are *explicitly* unilateral. In Searle's (1992) proposals, for example, speech acts are treated as autonomous acts by the speaker S toward a hearer H. "S goes up to H and cuts loose with an acoustic blast; if all goes well, ...then the speech act is successful and nondefective... In real life, speech [consists] of sequences of exchange speech acts in a conversation, where alternately S becomes H; and H, S" (p. 7). Searle follows Grice (1975) in assuming that speakers expect addressees to cooperate in interpreting their utterances. It is just that speakers design their utterances without the active participation of addressees, an assumption common to unilateral models (e.g., Clark & Haviland, 1977; Grice, 1975, 1991; Horton & Keysar, 1996; Sperber & Wilson, 1986).

Some accounts of speaking and listening are explicitly bilateral. According to Sacks, Schegloff, and Jefferson (1974), the length and shape of a turn is determined not by the current speaker alone, but by the current and potential next speakers working jointly. Following Sacks et al., there is a long tradition of research showing that speaking and listening in conversation are bilateral processes (e.g., Atkinson & Heritage, 1984; Button & Lee, 1987; Drew & Heritage, 1992; Schegloff et al., 1977). Other research has reached much the same conclusion with evidence from gestures (Bavelas, Chovil, Lawrie, & Wade, 1992; Bavelas, Coates, & Johnson, 2000; Engle, 1998, 2000; Streeck, 1993, 1994), reference (Clark & Wilkes-Gibbs, 1986), computer interfaces (Brennan, 1990; Clark & Brennan, 1991), and comprehension in general (Clark, 1997). Perhaps the clearest evidence for speaking as a bilateral process is found in grounding.

Grounding

In dialogue, speakers try to *ground* their communicative acts as they go along: They work with their partners to reach the mutual belief that the partners have understood them well enough for current purposes (Clark & Schaefer, 1989; Clark & Wilkes-Gibbs, 1986; Traum, 1994). Consider a spontaneous exchange (from Svartvik & Quirk, 1980, with pairs of asterisks marking overlapping speech):

Alan	were you there when they erected the new signs? -
Beth	th- w*hich* new *signs*?
Alan	*litt*le notice boards, indicating where you had to go for everything,
Beth	no,

For Alan and Beth to ground his question about "the new signs," they must deal with four levels of joint action, ordered from bottom to top (Clark, 1996).

Level 1. Alan must get Beth to *attend* to his vocalization. They would ordinarily try to establish this as common ground. If she had missed it, she might say "What?" or "Pardon?" and he would repeat it.

Level 2. Alan must get Beth to *identify* the words, phrases, and sentence he has presented. They would ordinarily try to establish her identification as common ground. If she was uncertain of "erected," she might ask "Did *what* to the new signs?" to which he would respond "erected."

Level 3. Alan must get Beth to *understand* what he means by those words. What does he mean by "there," and which signs is he referring to? They would ordinarily try to establish this, too, as common ground. In fact, Beth asks "*Which* new signs?" and Alan explains, "Little notice boards, indicating where you had to go for everything."

Level 4. Alan must get Beth to *consider* answering his question. Should she reveal she was there or not? Does she remember? They would ordinarily try to establish this as common ground as well. She could reply "I don't recall" or "I'll never tell." In fact, she answers "no."

People have many ways of grounding at these four levels (Clark, 1996). Addressees can use continuers such as *uh huh* and *yeah* (Schegloff, 1982), assessments such as *gosh* (Goodwin, 1986a), appropriate next contributions (e.g., answers to questions), echoic repeats, side-sequences (Jefferson, 1972), and other techniques. Many of these work via downward evidence (Clark, 1996). When Beth says "no," she demonstrates that she has understood what Alan meant by "Little notice boards, indicating where you had to go for everything" (level 3). But her answer also entails that she has identified Alan's words and phrases (level 2), which entails that she has attended to his vocalizations (level 1).

Speakers often change course because of what their partners say. Overlapping speech mid-utterance, for example, often interferes with grounding at level 1, and both the interrupting and the interrupted speakers have ways of repairing the problem (Sacks et al., 1974). When speakers are unsure if their partners will recognize a name or other reference, they often produce it with a *try marker*, a rising intonation followed by a slight pause, to request confirmation mid-utterance (Sacks & Schegloff, 1979). Speakers tend to produce other types of utterances in installments (e.g., telephone numbers, addresses, instructions, and recipes) and get confirmation on each installment before going on (Clark & Schaefer, 1987, 1989; Cohen, 1984; Geluykens, 1987, 1988; Goldberg, 1975). This is all evidence for bilateral accounts of speaking and listening (see Clark, 1996; Clark & Wilkes-Gibbs, 1986; Schegloff, 1991).

Other-monitoring

In bilateral accounts, speakers monitor at all four levels of joint action. *What* they monitor can be divided into five perceptual regions in and around their partners: (1) voices; (2) faces; (3) workspaces; (4) bodies; and (5) shared scenes. As illustration, we will refer to the moment at which Alan is speaking to Beth.

(1) *Voices*. People in dialogue pay attention to each other's vocal acts. When Alan produces an utterance to express what he wants, Beth tries to attend to it, and vice versa.

(2) *Faces*. In face-to-face conversation, people keep close track of each other's faces—especially the eyes and facial gestures. Alan must often keep track of where Beth is looking, and vice versa, and they can do so with great accuracy (Gale & Monk, 2000). They use mutual eye gaze to signal that they are attending to each other's speech (Argyle & Cook, 1976; Argyle, Lalljee, & Cook, 1968; Goodwin, 1981; Kendon, 1967). They also attend to each other's smiles, frowns, grimaces, and other facial gestures (Bavelas, Black, Lemery, & Mullett, 1986; Kraut & Johnston, 1979).

(3) *Workspaces*. While talking, Alan may perform actions in the region in front of his body—his *workspace*. There he produces manual gestures: pointing, or deictic, gestures; iconic gestures; so-called emblems (like thumbs-up); and beats (Ekman & Friesen, 1969; Kendon, 1993; McNeill, 1992). There he manipulates physical objects (Clark, 1996, 2003). In the kitchen, Alan may expect Beth to watch him cut and sauté vegetables as he talks about them, and at dinner, he may expect her to see him hold out a plate or pour the wine. Workspaces are essential to many games. To succeed in tennis, chess, and bridge, players must attend to each other's actions on the court, chess board, and card table.

(4) *Bodies*. Alan and Beth also take notice of the actions and orientations of each other's bodies—especially head and torso. Alan can signal Beth what he is attending to by orienting his head, torso, or both toward an object (Schegloff, 1998). He can also use his body for other gestures, such as shrugs, head shakes, and head nods.

(5) *Shared scenes*. Alan and Beth also track their joint attention in areas beyond their workspaces, such as pictures on walls, tennis matches, or cars on highways. Standing on the side-lines, they can refer to "that Cézanne," "that volley," or "those Toyotas" even without gestures.

Much of the evidence speakers monitor for divides into signals versus symptoms. Signals are acts that are jointly construed as one person meaning something for others. Meaning here is speaker's meaning as characterized by Grice (1957, 1991), and joint construal is as described in Clark (1996, pp. 192–196, 212–216). Signals include not only vocal acts such as "uh-huh" and "yeah," but gestural acts such as pointing and head nods. Symptoms, in contrast, are acts that are *not* jointly construed as one person meaning something for others. These include self-talk and other actions that are not manifestly displayed to others. Still, speakers often use what their partners happen to be doing—symptoms—to infer what they are thinking.

People in face-to-face dialogue orchestrate their signals in several regions at once (see, e.g., Brennan, 1990; Clark, 1996; Engle, 1998, 2000; Streeck, 1993, 1994). Nowhere is this more evident than in grounding. Speakers monitor their addressees' eye gaze, and when the addressees are not gazing in return, they may alter the course of their utterances to obtain the return gaze (Goodwin, 1981). Speakers often elicit help from addressees mid-utterance by the use of eye-gaze, gestures, or the two in combination (Bavelas & Chovil, 2000; Goodwin & Goodwin, 1986; Goodwin, 1986b).

Vocal and gestural actions

How do people in conversation divide their efforts between vocal and gestural actions? They could, in principle, do everything vocally, as they do on the telephone. One reason they do not, we propose, is the *principle of least joint effort* (Clark, 1996; Clark & Brennan, 1991; Clark & Wilkes-Gibbs, 1986; Clark & Schaefer, 1989). According to that principle, people are *opportunistic*: they try to select from the available methods the ones they think take the least effort for the two of them jointly—the least cost in time, resources, errors, etc. (see Clark & Brennan, 1991). Face to face, they should exploit that combination of vocal and gestural actions they judge will take the least joint effort.

If people are opportunistic, they should generally opt for the grounding methods that are most efficient (Clark & Brennan, 1991). By efficiency, we mean speed with the least effort for a given accuracy. They should exploit

gestures when that would be more efficient. For example, they might point and exhibit things when their workspaces are visible, and nod and smile when their faces are visible. Visible workspaces are helpful in tasks with work objects—equations, blocks, and gears. People working collaboratively at a distance via computers are more efficient with shared visible workspaces, although they are no more efficient with visible faces with or without a shared workspace (Kraut, Gergle, & Fussell, 2002; Whittaker, 2003; cf. Boyle, Anderson, & Newlands, 1994). What grounding techniques do people use when? How efficient and accurate are these techniques, and why?

Finally, what if speakers cannot monitor their partners at all—what if they can neither hear nor see their partners? The simplest prediction is that they should make more errors, take longer, or both. Overhearers, listeners whose actions speakers do not monitor, make more errors in understanding than do addressees, whose actions speakers do monitor (Schober & Clark, 1989). And speakers who do not get feedback from addressees take longer and make more elaborate references (Krauss & Weinheimer, 1966).

The prediction of greater time and errors without feedback might also seem to follow from grounding (Clark & Schaefer, 1989; Clark & Wilkes-Gibbs, 1986)—establishing the mutual belief that addressees have understood the speaker *well enough for current purposes*. Whenever two partners cannot establish this belief, they should be more vulnerable to errors. And yet professional writers—newspaper reporters, novelists, script writers—do not get immediate feedback, and they seem to be understood. They apparently have learned, through training and experience, how to write and revise in a way that minimizes misunderstandings (Traxler & Gernsbacher, 1992, 1993). What about spontaneous speakers? Can they compensate for lack of feedback?

The experiment described here was designed to reveal how speakers monitor addressees for both vocal and gestural evidence and how they use this evidence in the course of their utterances. In the first part, we take up quantitative evidence about what speakers monitor their addressees for. We then examine the actual gestural and vocal techniques used and show how these account for the bulk of the quantitative results.

Methods

In this experiment, a *director* was asked to tell a *builder* how to assemble 10 simple Lego models. The director had a prototype for each model out of sight of the builder, and the builder assembled the model from a set of loose Lego blocks. In four *interactive* conditions, half of the partners could see the builder's workspace, and the other half could not. Half of the time the two partners could see each other's faces, and half the time they could not. We will refer to these two dimensions as *workspace visible* versus *workspace hidden* and *faces visible* versus *faces hidden*. In a fifth *non-interactive* condition, directors recorded their instructions blind to the builders, and builders later assembled their models from the recordings.

Procedure

In the *interactive* conditions, 28 pairs of people each worked together to assemble 10 Lego models. The director sat at one end of a table 2.0 m long, the builder at the other end. On the first trial, the experimenter placed a prototype behind a low screen in front and to the side of the director so that only the director could see it. The builder had dozens of miscellaneous Lego blocks scattered on the table or in a nearby box. The two participants were told that the builder was to assemble a model that was identical to the director's prototype. They could talk as much as they needed, and they were to let the experimenter know the moment they were finished ("We're done"). The experimenter then stepped outside the room, and when the pair had finished, she returned, checked the completed model for accuracy, showed the partners any discrepancies, disassembled the model, gave the director a second prototype, and left again. So it went for the 10 models. The participants gave us permission to make audio- and video-recordings of the sessions.

The 28 pairs were randomly assigned to one of two conditions, 14 per condition. In the workspace hidden condition, there was a low barrier across the middle of the table that kept the director from seeing the builder's blocks, hands, or model in progress, but not their faces. In the workspace visible condition, the low barrier was absent. For five of the 10 trials, there was a second, high barrier, also across the middle of the table, that kept the two of them from seeing each other's faces, but not the builder's blocks; for the other five trials, the high barrier was absent. These are the face hidden and face visible trials. The high barrier was present on the even numbered trials for half of the pairs in each condition, and on the odd numbered trials for the other half. In short, the two barriers created a 2×2 design by their presence or absence.

In the *non-interactive* condition, 10 pairs of participants each worked to assemble the same 10 Lego models much as in the interactive conditions, but without interacting. The 10 directors, sitting at the same table, were recorded as they told a future builder how to assemble the 10 Lego models. They were told that the builder would assemble the models a week later from their recording. They could look at the prototype as long as they wanted before starting their instruction. They were given the first prototype, and the experimenter left

the room. When they were finished, they said "Done," and the experimenter returned, gave them the next prototype, and left again. So it went for the 10 prototypes.

In a later session, the 10 builders, each yoked with a different director, sat at the same table with loose blocks on it and assembled the 10 models from the tape recording of the director's instructions. They were allowed to start, stop, and rewind the tape as often as they wanted. When they were finished with each model, they said "Done," and the experimenter returned, checked the model for accuracy, showed them any errors, and left again.

The 10 Lego prototypes each consisted of six to eight large Lego blocks (technically, Duplo blocks) three or four blocks high. They were designed so they could not be simply described as familiar objects such as bridges, animals, or buildings. Pilot testing showed that the 10 models took roughly equal time to assemble. The 10 models were completed in the same order by all pairs of participants.

The 76 participants were Stanford University students (34 male and 42 female) who received either payment or course credit as part of an introductory psychology course. The two partners of each pair, who were unacquainted with each other, were randomly assigned to be director and builder at the beginning of the session.

Video and audio analyses

All sessions were recorded on analogue videotape. Two small (5 cm by 5 cm by 5 cm), black-and-white, wide-angle video cameras were placed unobtrusively in the middle of the table but off center, one trained on the director and the other on the builder. The two outputs were fed through an image splitter onto a single videotape. Each session was also recorded on audiotape. It was impossible to analyze the video- and audiotapes fully, so we carried out selective analyses as follows.

Time and errors. For the interactive conditions, we measured the building times for each of the 10 models for each of the 38 pairs of participants. The building time for a model was measured from the moment two partners began speaking about a model to the moment they said "We're done." For the non-interactive conditions, we measured the description and building times separately. We also noted all models and blocks in error.

Words and turns. We made detailed transcripts of what people said for 16 pairs of participants in the interactive conditions, half in the workspace visible condition and half in the workspace hidden condition.

Gestures and other actions. In the most time-consuming analysis, we examined the videotapes of models 4 and 5 for eight pairs of participants in the workspace visible condition. We chose pairs for which the gestures and other actions were clearest on the videotapes. We converted these 16 videotape segments into 16 QuickTime digital movie clips, totaling 33.8 min, or 127 s per model. We analyzed the 16 clips using both the timing capabilities of QuickTime and the mark-up capabilities of MediaTagger, an application that allows one to mark events in QuickTime movies frame by frame in one or more tiers. With the director and builder on a split screen, we were able to identify what they were doing at identical times.

We report the results of this experiment in three parts—interactive partners, non-interactive partners, and grounding with gestures.

Interactive partners

The first issue is how efficiently and accurately people worked when the builder's workspace was mutually visible and when they could see each other's faces. We begin by characterizing how two people carried out this task.

Building a Lego model usually fell into six to eight *building cycles*, each with two main steps. *Step 1: identify block.* The director got the builder to find the next block (or blocks) to be placed. *Step 2: place block.* The director got the builder to put the block where it was supposed to go. Consider this exchange from the Legos-hidden condition (David is the director and Ben the builder):

David	And then you're gonna take a blue block of four.
Ben	M-hm.
David	And you're gonna put it on top of the four blocks—four yellow blocks farthest away from you.
Ben	Which are the ones closest to the green.
David	Yeah
Ben	Okay. But the green's still not attached.
David	Yeah. And then...

These seven turns cover a single building cycle. Step 1, identify block, is completed in the first two turns and step 2, place block, in the last five. The turns for each step fit a characteristic pattern for grounding in dialogue: a *presentation phase* followed by an *acceptance phase* (Clark & Schaefer, 1989). In turn 1, David presents Ben with an instruction ("And then you're gonna take a blue block of four") and in turn 2, Ben confirms that he understands with an acknowledgment ("M-hm"). In turn 3, David accepts Ben's confirmation by going on. Grounding the place-block step takes more turns, as Ben checks on his understanding several times before David goes on.

By contrast, consider this exchange for the same block and model from a workspace visible condition, with gestural acts enclosed in square brackets (Doris is the director and Betty the builder):

Doris	Take a short blue.
Betty	[Retrieves a short blue block.]
Doris	[Looks at Betty's block.] Put it at the end of the yellow close to the green.
Betty	[Places the blue block on the yellow block.]
Doris	[Looks at result.] Take a...

Step 1, identify block, is completed in lines 1 and 2 and step 2, place block, in lines 3 through 5. But with visual access to Betty, Doris confirms Betty's understanding by inspecting what she has done, and she implicates that Betty is right by continuing on ("Put it..."). Grounding is achieved *without Betty saying a word.*

The main way people coordinate in exchanges like these is with *adjacency pairs* (Sacks & Jefferson, 1992; Schegloff & Sacks, 1973), as in this exchange:

B	So the yellow is pointing off to the left and the green is pointing up?
D	Yeah.

In the first turn, B *proposes* that D tell B whether the yellow and green blocks fit a particular pattern, and in the second turn, D *takes up* B's proposal and completes it by saying "yeah" (see Clark, 1996). To be an adjacency pair, however, the two parts must be utterances, and when the workspace was visible, one or both parts of similar exchanges were often gestural acts, as here:

B	Oh, on top of the yellow and the blue?
D	[Shakes head]

Although B asks a question by speaking, D answers with a head shake. We propose the term *projective pair* to cover both spoken and gestural versions of such pairs (Clark, in press). A projective pair consists of two actions in sequence, by two people, in which the first person is jointly construed as making a proposal, and the second, as taking it up. The pair is projective in that the first action is taken as *projecting* the second.

Efficiency

At least some gestural acts, we assume, are more efficient for grounding than the vocal acts that would be needed without them. If so, two partners should be faster when they can take advantage of gestural acts—as in Doris and Betty's exchange. To test this prediction, we examined the *building time* for each model—the time from the moment two partners began speaking about a model to the moment they said "We're done." The average building times for the four conditions are shown in Fig. 1. As predicted, they were much shorter when the workspace was visible than when it was hidden, 90–194 s, $F(1, 52) = 55$, $p < .001$. Building times were 8 s less when the faces were visible, but this difference was not significant, $F(1, 52) < 1$, nor was the interaction with visibility of the workspace. (The visibility of faces made no difference in any of our measures, so we will mention it no further.)

Number of words used, which is correlated with speaking time, shows that *both* partners contributed to efficiency. The mean for each model is shown for 16 pairs of participants in Fig. 2. As predicted, two partners together used fewer words in the workspace visible condition than in the workspace hidden condition, 213 words to 450 words, $F(1, 28) = 107$, $p < .001$. Not surprisingly, directors used over four times as many words as builders, 265 words to 66 words, $F(1, 28) = 38$, $p < .001$. The interaction between these two factors was not significant. Plainly, building a Lego model was more

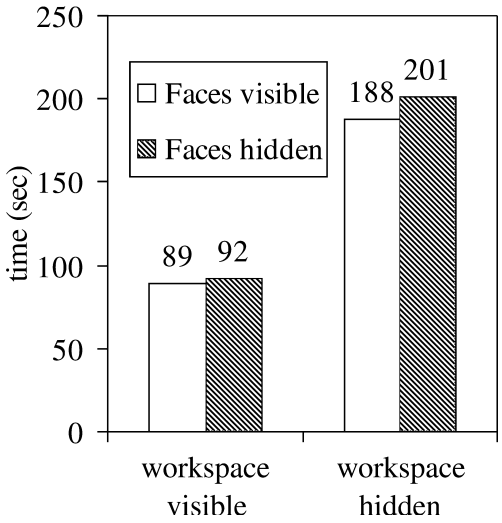

Fig. 1. Mean building times per model.

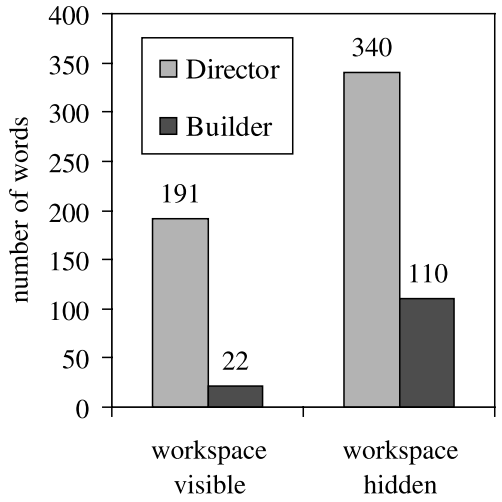

Fig. 2. Mean number of words per model.

efficient when the director could see the builder's workspace, although it was not measurably more efficient when they could see each other's faces.

Errors should be rare if two partners work on each building cycle until it is grounded to their satisfaction. Indeed, there were errors on only 12 of the 280 models (4% of the models)—five when the workspace was visible and seven when it was hidden—and each error had just one block out of place.

Turns

Grounding should take different forms when the workspace is visible and hidden. When the workspace was visible, Doris confirmed Betty's understanding by inspecting what Betty had done. She took one long turn, and Betty took none at all. But when the workspace was hidden, David sought spoken confirmation from Ben, which required both parties to take turns. These differences are reflected in the mean number of words per turn (coded as change of speaker) shown in Fig. 3.

Directors used over five times as many words per turn as builders, 16–3, $F(1,28) = 184$, $p < .001$. There were slightly more words per turn when the workspace was visible than when it was hidden, 10.9–8.4, $F(1,28) = 7.9$, $p < .01$. But directors used *more* words per turn when the workspace was visible than when it was hidden, 19.5–12.6, whereas builders used *fewer* words per turn, 2.2–3.8. The interaction is significant, $F(1,28) = 19$, $p < .001$. This fits the exchanges cited earlier. Because Doris's workspace was visible, she could manage without speaking; but because David's workspace was hidden, he required many words. Indeed, when the builder's workspace was visible, there were five models on which builders used only a single word. When it was hidden, all models required builders to use at least 12 words.

Sources of evidence

Visible evidence of understanding is generally taken to be more reliable than mere spoken claims of understanding (Clark & Marshall, 1981). Doris and Betty can be certain that Betty has placed a block in the right place because Doris can *see* it in the right place. In contrast, even once Ben gets a block in the right place, he still feels obliged to describe the result to David ("Which are the ones closest to the green" and "But the green's still not attached") and get his confirmation ("yeah" and "yeah"). Let us call this extra process *checking time*. Checking time should be greater when the workspace is hidden.

To test this prediction, we examined model number 4 (a representative model mid-task) on the videotapes of all interactive pairs. (The blocks were obscured for one pair, so we eliminated a balancing pair from the other condition.) We divided each building cycle into a base time and a checking time. The base time was the interval, measured on videotape, from the beginning of the director's description to the moment the builder got the right block into the final, objectively correct location. The checking time was the additional time two partners spent checking on the correctness of that location up to the moment the director went on to describe the next block.

As expected, the base and checking times were shorter when the workspace was visible than when it was hidden. The average times are shown in Fig. 4. Both the base and checking times dropped with a visible

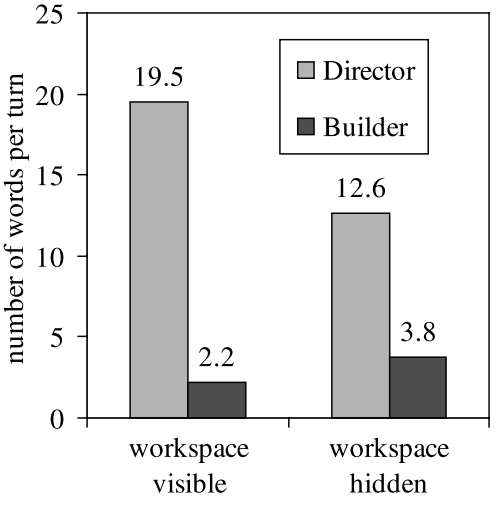

Fig. 3. Mean number of words per turn.

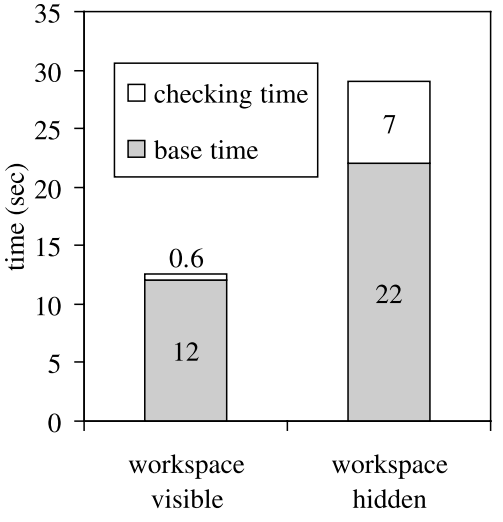

Fig. 4. Mean base and checking times per block in model 4.

workspace, $F(1,24) = 14.8$, $p < .001$, but the checking times dropped more. The interaction between hidden–visible and base–checking time was significant, $F(1,24) = 12.2$, $p < .002$. Viewed differently, the percentage of total time spent checking was reduced from 21 to 5%. These findings replicate observations by Brennan (1990). So in monitoring others, speakers make more efficient use of visible than of spoken evidence of understanding.

Non-interactive partners

The second issue to be examined is how efficiently and accurately two partners work when they cannot interact at all. Indeed, they have major difficulties.

Table 1 lists the percentage of model and block errors in the non-interactive condition and in the comparable interactive condition, the workspace hidden condition. A model was counted as in error whenever it did not match the prototype in every way. There were 5% model errors in the interactive condition, but 39% in the non-interactive condition. The difference is significant, $F(1,22) = 25.9$, $p < .001$. A block was counted as in error whenver it was the wrong color or size, or in the wrong location or orientation. Block errors were counted in relation to the first block, so with 69 blocks in the 10 models, there were 59 possible errors per builder. There were 0.8% errors (7 out of 826) in the interactive condition, but 12.5% (74 out of 590) in the non-interactive condition. The difference is significant, $F(1,22) = 8.7$, $p < .01$. So when monitoring was precluded, builders made eight times as many model errors, and 14 times as many block errors. These increases are dramatic by any standard.

Percentage of errors does not tell the whole story. In the interactive condition, there was little variation in errors from one pair to the next. Number of errors per pair ranged from 0 to 2 models and from 0 to 3.3% of the blocks. In the non-interactive condition, the variation was enormous: Number of errors per pair ranged from 2 to all 10 models and from 3.3 to 53% of the blocks. The most accurate pair in the non-interactive condition was only as good as the least accurate pair in the interactive condition. When we gave the instructions from the worst pair in the non-interactive condition (10 model errors, 53% block errors) to a second builder, he made about as many errors (9 model errors, 47% block errors) as the first builder. This suggests that the directors are largely responsible for this variation.

Table 1
Percentage of errors in interactive instructions (workspace hidden condition) and in non-interactive instructions

Condition	Model errors	Block errors
Interactive	5	0.8
Non-interactive	39	12.5

In the non-interactive condition, builders rewound or paused the tape an average of 7.7 times and took 245 s per model (range 162–374 s). Directors, who could delay as much as they wanted before speaking, took 274 s per model (range 171–475 s), which was not significantly longer than the builders. In the interactive condition, in contrast, pairs averaged only 183 s per model (range 111–289 s), which was significantly shorter than the non-interactive builders, $F(1,22) = 7.16$, $p < .02$.

Directors in this task, therefore, could not compensate fully for their inability to monitor builders. The Lego models we used are simple as such objects go, yet directors were unable to give effective instructions without monitoring in some way. The main reason, suggested by one example, was not that builders could not keep up with the instructions, but that directors were giving inadequate instructions. People speaking spontaneously may be able to compensate for lack of other-monitoring in straight-forward descriptions, but not in descriptions with certain complexities.

Gestures and grounding

The third issue to be examined is why grounding is more efficient when the builder's workspace is visible to both partners. From a close look at the videotapes, the answer seems obvious: when the workspace is visible, the partners ground what they say not only with speech, but with gestures and other actions. To see how, let us begin with deictic expressions, which often require gestures or other actions.

Deictic expressions

Deictic expressions such as *this*, *that*, *here*, and *there* are specialized for indexing things in the local surroundings. They require both speakers and addressees to establish that the things indexed are in their joint attention (see, e.g., Clark, 1996). That, in turn, requires speakers to monitor what their addressees are doing, and addressees to show what they are doing. If so, directors and builders in our task should have used deictic expressions differently when they could monitor each other visually than when they could not.

To examine this issue, we counted the number of turns with deictic *here*, *there*, *this* (including *these*), *that* (and *those*), *like this* (and *like these*), and *like that* (and *like those*) for the 16 pairs in the interactive conditions. (We were careful to exclude non-deictic uses of *that* and *there* as in "the blue's on the right—so that they make a square" and "and so there's a space uh underneath it.") The percentages of turns with these expressions are shown in Fig. 5. When a turn had more than one deictic

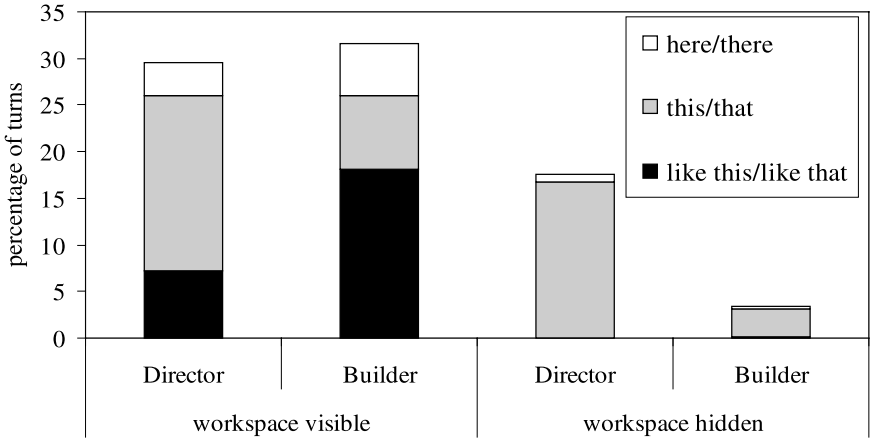

Fig. 5. Percentage of turns with deictic expressions.

expression, we classified it first by the presence of *like this* or *like that*, then *this* or *that*, and finally *here* or *there*.

As predicted, deictic expressions were used in more turns when the workspace was visible than when it was hidden, 31–11%. The difference is significant, $F(1,28) = 41$, $p < .001$. The percentages were about equal for directors and builders when the workspace was visible (30–32%), but greater for directors when the workspace was hidden (18–4%). The interaction between visible–hidden and director–builder was significant, $F(1,28) = 11$, $p < .005$.

When the builder's workspace was hidden, both partners used *this* and *that*, but in limited ways. Directors would say, "All right, um, on top of *that* is a red four by two piece," and builders would say, "No, I don't have *that*, " or ask, "So it's perpendicular to *that* ?" That is, both partners used *this* and *that* for objects just mentioned even though directors could not see them. In selecting *this* vs. *that*, the two partners represented the model-so-far from the *same* viewpoint—largely as distal from both. Directors favored distal *that* to proximal *this*, by 69–31% ($N = 396$), and so did builders, 64–36% ($N = 73$). The two partners used *here, there, like this*, and *like that* only 7, 7, 4, and 0 times.

When the builder's workspace *was* visible, the two partners added *here, there, like this*, and *like that*. They used *like that* and *like this*, for example, to point to particular arrangements of blocks, as when builders asked, "Like this?" But because they were able to see the workspace in relation to each other's location, the two partners took complementary viewpoints of the model-so-far: both treated the model as distal from the director and proximal to the builder. Directors favored *that* over *this* 90–10% ($N = 162$), *there* over *here* 88–12% ($N = 33$), and *like that* over *like this* 100–0% ($N = 59$). Builders showed the reverse preferences—20–80% ($N = 64$), 28–72% ($N = 43$), and 45–55% ($N = 140$). Statistically, the preference for distal expressions was greater with the workspace hidden than visible, $F(1,27) = 10.2$, $p < .005$, and greater for directors than for builders, $F(1,27) = 12.4$, $p < .005$. The interaction between the two accounted for the most variance, $F(1,27) = 42.8$, $p < .001$.

In brief, speakers cannot use deictic expressions without monitoring what their addressees are attending to—a bilateral process. This showed up in several ways. When directors could not see the builder's workspace, the two partners used only *this* and *that* for just mentioned objects and viewed them from the same perspective. But when directors *could* see the workspace, the two partners viewed the objects from complementary perspectives. And both partners used *here*, *there*, *like this*, and *like that* only when directors could visually monitor what the builders were doing.

Gestures by addressees

Builders regularly indicated things for directors with gestures. It was often these gestures that permitted the use of *this*, *that*, *here*, *like this*, and *like that*. What gestures did builders use and how?

For this analysis we created gestural transcripts for the 16 video clips of models 4 and 5 for eight pairs of participants in the workspace visible condition. To create the transcripts, we examined the video clips in several steps. First, with the sound off, we focused on the builders' hands. As it happens, builders tended to: (a) move their hands (usually with blocks in them) to a discrete location or orientation and then (b) hold their hands motionless for a period of time (if only a few frames). We call the beginning of this period the *onset* and the period itself a *hold*. At the end of the hold, they moved their hands to a new location or orientation. We recorded the frame at each onset and the duration of the hold in frames. Second, we classified what the builder

was doing at each onset and hold. And, finally, we added these annotations to the speech transcripts of the 16 movie clips. With the sound on, we identified the director's utterance just before and just after each onset, and the builder's utterance, if any, that accompanied the hold. We also noted the few gestures produced by directors. There were, in total, 332 holds in the 16 gesture transcripts.

The builders' actions were divided into the categories shown in Table 2. (All actions were classified by one coder; 13% were classified by a second coder with 93% agreement.) Two-thirds of these actions (66% of the total) were *explicit gestures*, actions that were not directly part of the building process except as communicative actions. *Manifest actions* (25% of the total) were direct parts of the building process and were made visually available to the directors. These included placing a block on the table, positioning or repositioning a block, attaching a block to the model-so-far, picking up a block, etc. *Manifest negative actions* (1%) were actions in which builders undid what they had already done—e.g., detaching a block from the model-so-far. *Explicit postponements* (9%) were actions in which builders manifestly refrained from taking the expected next action. They held a block instead of placing it, or placed their hand on a block without picking it up. We begin with the explicit gestures.

Table 2
Types of builders' actions and their percentages (Models 4 and 5 in workspace visible condition)

Action types	Percent	Percent
Explicit gestures		66.0
Exhibit, re-exhibit block	20.8	
Exhibit model-so-far	3.9	
Poise, re-poise block	39.5	
Point at, touch block	1.8	
Manifest actions		24.7
Position, re-position block	10.5	
Place block down	6.9	
Attach block (without poising)	3.3	
Arrange, re-arrange several blocks	1.8	
Pick up block	0.6	
Explicit postponements		8.7
Hold block motionless	3.9	
Poise block at distance	3.9	
Place hand on block	0.9	
Manifest negative actions		0.6
De-position block	0.3	
Detach block	0.3	
Total ($N = 332$)	100.0	100.0

(1) *Exhibiting*. Once directors described a block, builders had to retrieve a block and verify that it was the intended type. To do this, they often held up, or placed, a candidate block manifestly for the directors to look at, expecting them to confirm or deny that it was the right type. We will call this gesture *exhibiting*. That is, an exhibit is an action by which a person brings a thing into a conspicuous location and manifestly holds it there for inspection. The projective pair is this:

B [exhibits an object to D as candidate for D's referent]
D [confirms or denies that the object is D's referent]

Exhibiting, then, is an attempted signal (meaning, e.g., "Is it this one?"), the first part of a projective pair. Its joint construal as a signal is complete when directors signal their acceptance or rejection (meaning, e.g., "Yes it is"). Builders sometimes exhibited the model-so-far, which they would push forward or lift off the table for the director to see.

(2) *Poising*. Once directors described where a block was to go, builders often manifestly held, or poised, a block just above the location in the model where they thought it ought to go. Builders were trying to ask, in effect, "Does the block go here?" expecting their partners to signal yes or no. We will call this gesture *poising*.

(3) *Pointing*. Although pointing with the finger is common in many settings, it was used only five times in the 16 video clips. In one instance, a builder touched the location where she thought the block was to go. When builders pointed, it was as if to ask, "Does the block go here?" expecting directors to signal yes or no.

Most exhibits, poises, and points were jointly construed as signals, so the evidence suggests. First, builders accompanied 35% of their explicit gestures with expressions such as "Okay?" "Like this?" and "These two?" Just over half of these expressions contained deictic expressions. Ordinarily, gestures are considered *composite* parts of references made with deictic expressions, and that makes each of these gesture-plus-expressions a signal (see Clark, 1996). Second, directors responded to all but five of the builders' actions as signals. For 50% of them ($N = 109$), they replied with an explicit verification or rejection (e.g., "Okay," "No," "Exactly, perfect"), timed as a response to the builder's action. In the other replies timed as responses, they verified or rejected the actions by *implicating* that the builders were either correct (e.g., "[B exhibits block] and, and, uh, place it on the blue side") or incorrect (e.g., "[B poises block] but move it one notch [B re-poises block] towards the, towards yourself [B re-poises block again]"). Directors were more likely to give an explicit verdict when builders had made a query ("Like

this?") than when they had not, 75–25% of the time. (We return to the precision of this timing later.)

Most of the manifest actions, negative actions, and postponements were jointly construed as signals as well. Consider the 35 manifest actions of positioning and repositioning. Builders accompanied many of these ($N = 11$) with queries such as "like this?" "just those two?" and "here?" These expressions overlapped the onset or hold of the gesture, suggesting that the gestures-plus-expressions were designed as signals. And, in timed responses, directors replied to many of them ($N = 16$) with explicit verifications or rejections, and to the rest ($N = 19$) with assertions that implicated acceptance or rejection. In one example, D says, "All the way up toward me," at which point B repositions the block and asks "like that?" and D replies "exactly, yeah." In repositioning the block, B indicated the new position, just as he would have by pointing at the new position. That is, his repositioning plus "like that?" was taken as a signal.

Finally, consider the postponements ($N = 29$)—distant poising, holding, and putting hands on a block. Builders appeared to use these to signal that they had too little information to proceed, and in every case, directors responded with more information. In one example, D says, "Put it on top of the yellow." At that point B poises the block high above the model, implicating that she does not yet know where to put it. That prompts D to go on, "But facing, uh, out, outwards. [B poises block on model] Exactly." In no instance did directors respond with an explicit "yes," "no," or "exactly."

In brief, most of the gestural actions in Table 2 were jointly construed as signals. Many were taken to mean "Do you mean this one, or like this, or here?" and a few to mean "I need more information." When the workspace was hidden, in contrast, the manual actions by the builders were *never* treated as signals. They were never accompanied by "like this?" "just those two?" or "here?" They were never accepted or rejected with precisely timed "yeah," "no," "exactly," or continuations. And they were never displayed at the hidden directors—not even at imaginary, non-visible partners. The contrast is clear-cut, further evidence that most of the gestural actions in the visible workspace were taken as signals.

Cross-timing of actions

If speaking is bilateral, speakers should respond to certain of their partners' actions in the course of their utterances. For evidence that they do, let us consider the cross-timing of speaker and partner actions.

We will illustrate cross-timing by means of *action graphs*. Fig. 6, our first example, represents 9 s of interaction between Sam and Ted. To create the graph, we used MediaTagger to mark frames in the videotape in several tiers. In tier 1, we represented Sam's speech, "kay now get | a-uh eight green piece | and join the two | so it's all symmetric | yeah right in the center." The vertical lines mark natural breaks in the speech, usually silences, which divide the speech into five parcels. We then marked the frames on which each parcel began and ended. In tier 2, we represented what Ted did with his

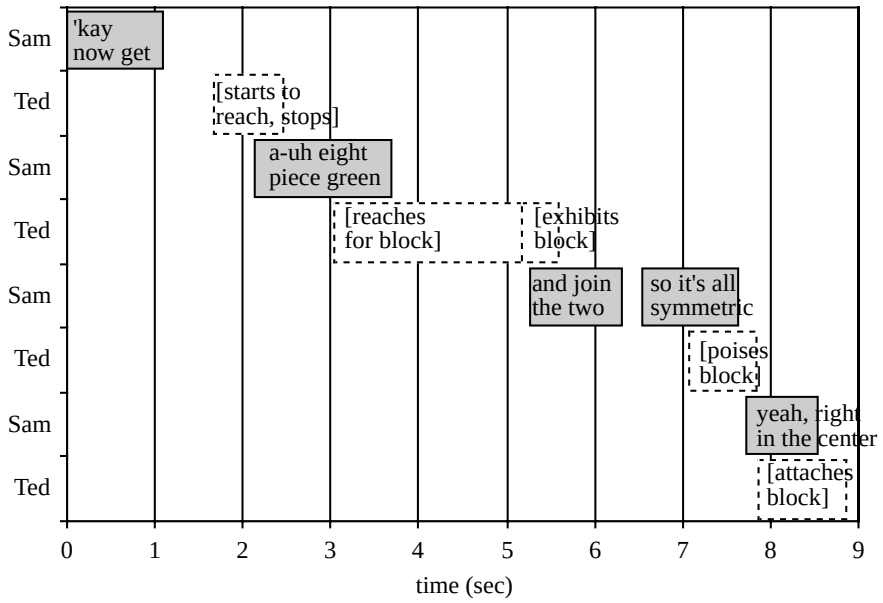

Fig. 6. Action graph with exhibiting and poising.

hands. We identified five distinct actions ("begins to reach," "retrieves block," "exhibits block," "poises block," and "attaches block") and marked the frames on which each action began and ended. In creating tier 1, we ignored the side of the screen with Ted on it and in creating tier 2, we turned off Sam's voice. Fig. 6 represents the two tiers. For other graphs, we created separate tiers for the director's voice, builder's voice, director's hands, builder's hands, director's face and eyes, and builder's face and eyes.

Here we consider five timing phenomena, most of which are illustrated in Fig. 6. We begin with the builders' actions.

Gestural signals. As noted earlier, addressees generally try to keep speakers informed of their current state of understanding. In our task, builders often did that with gestural signals that reflected their understanding *at that moment*. In Fig. 6, Ted initiates an exhibit (line 4) and a poise (line 6) apparently at the moment he understands which block is intended and where it is to go.

Overlapping signals. Although people generally try to avoid speaking while their partners are speaking (Sacks et al., 1974), they do not try to avoid gesturing. In Fig. 6, Ted begins his poise (in line 4) just as soon, apparently, as he believes he knows where the block goes, even though it overlaps with Sam's "and join the two..." In our 16 video clips, poises, exhibits, points, and other builder gestures often overlapped with directors' speech. We examined the 73 initial poises in the 16 video clips and measured the time interval between the start of the builder's moving into a poise and the end of the director's description of the location. The average overlap was .58 s (range –1.4–2.7 s). This was significantly greater than 0.00, $t(7) = 4.40$, $p < .005$. That is, builders regularly initiated their gestures before directors had finished their descriptions.

Projecting understanding. Listeners usually do not wait for the end of an utterance before acting on what they have heard. They begin some actions as soon as they have enough information to begin (e.g., Spivey, Tanenhaus, Eberhard, & Sedivy, 2002, Spivey-Knowlton, Tanenhaus, Eberhard, & Sedivy, 1998). In our task, builders often began to act as soon as they could merely *project* what their partners would say.

The point is illustrated in Fig. 6. In line 2, Ted begins reaching into his storehouse of blocks just after Sam has said "and now get." He has apparently projected that Sam is about to describe the type of block to get next. When Sam delays 1 s to formulate that description, Ted, in response, stops reaching and waits (his right hand resting on his left arm, his eyes still on the storehouse). Once Sam starts speaking again, "a-uh eight piece green," Ted starts reaching again. He starts at the words "piece green," once again before he could know what he is reaching for. He apparently has projected that he will have understood Sam's description by the time he gets to the blocks.

Initiation time. Speakers can *in principle* initiate an utterance as soon as they have formulated enough of it to start (Clark & Wasow, 1998; Levelt, 1989). Still, they rarely do that. In dialogue, people normally wait their turn, engineering their next utterance to start at the end of the current speaker's turn (Jefferson, 1973; Sacks et al., 1974). They do not simply blurt when ready. The condition is this (Clark, 1996; Goodwin, 1981): try to speak when your partner is prepared to attend to, parse, and understand what you say.

The point is illustrated in Fig. 6. In the videotape, Sam seems prepared to initiate his next instruction right after "eight piece green," and yet he waits 1.5 s—a long interval in spoken dialogue (Jefferson, 1989). Why? Throughout this interval, Sam has been watching Ted retrieve the right block, and he starts "and join the two" precisely as Ted begins to exhibit the block. That is, Sam does not start when *he* is prepared to speak, but only when he believes *Ted* is prepared to attend. That is a bilateral process.

As evidence for this suggestion, we examined every speech delay over 2 s in the eight video clips with both workspaces and faces visible. There were 27 such delays, all by directors, ranging up to 9 s. In 19 of these, directors waited for builders to find a block ($N = 10$), attach a block ($N = 7$), or detach a block ($N = 2$) before giving the next instruction. In 3 cases, they waited for builders to attach a block before confirming it. In the remaining five cases, they spent the time looking at their own prototype in preparation for their next instruction. So 22 of the 27 delays were by directors waiting for the attention of the builders (at level 1).

Timing uptake. As noted earlier, speakers often take up their partners' gestural signals in projective pairs. In Fig. 5, Ted exhibits a block and .2 s later, Sam takes up the exhibit by going on—a standard method of grounding (Clark & Schaefer, 1989). Next, Ted poises a block over the model-so-far (meaning "Does it go here?") and 0.5 s later, Sam replies "Yeah."

The uptake of a gestural signal, we propose, is signaled in part by its timing. Once Ted has initiated a projective pair by poising a block, Sam must complete the pair while the poise is still in place. To do this, he must respond immediately. If he does not (and if other conditions are right), he will *implicate* that the poise is incorrect. But what is "immediate"? We assume that two partners are fairly good at estimating how long it should take directors to examine a builder's poise and decide whether or not it is correct. If directors exceed this limit, they implicate the answer "no." In our 16 video clips, the limit seemed to range between .3 and 1.0 s, though what builders perceived the limit to be is difficult to measure (see Jefferson, 1989). Let us call this the *immediacy constraint*.

The immediacy constraint is real even if immediacy is difficult to measure. As we illustrate later (e.g., Figs. 7–9),

directors generally responded within a second after the builder's initiation of a gestural signal or other action. They often anticipated the arrival at the apex (or greatest extension) of those gestures, as Sam appears to do in Fig. 6. And as we also illustrate later, builders were as attentive to this constraint as directors. Within 1–2 s of a director's response, they would begin attaching a block if its location was confirmed, or begin moving it if it was not. In Fig. 6, Ted begins attaching his block 0.2 s after Sam's "yeah."

Consider one instructive example of timing. At one point Ted says "like this?" and, 0.13 s after initiating the speech, begins poising a block over location 1; however, 0.17 s later, he leaves location 1 to re-poise the block over location 2. As it happens, just 0.10 s *before* he reaches location 2, Sam initiates, "yeah that end." Ted apparently notices the timing and interprets "yeah" as referring not to location 2, his current location, but to location 1; we infer this because he returns and attaches the block to location 1. Ted could not have got the placement correct without noticing the *precise* timing of Sam's uptake.

Timing strategies

Two partners exploited cross-timing in several interactive strategies. Two of these are self-interruption and collaborative reference.

Self-interruption. Speakers are known to interrupt themselves in order to make self-repairs (Blackmer & Mitton, 1991; Levelt, 1983, 1989), repeat words (Clark &

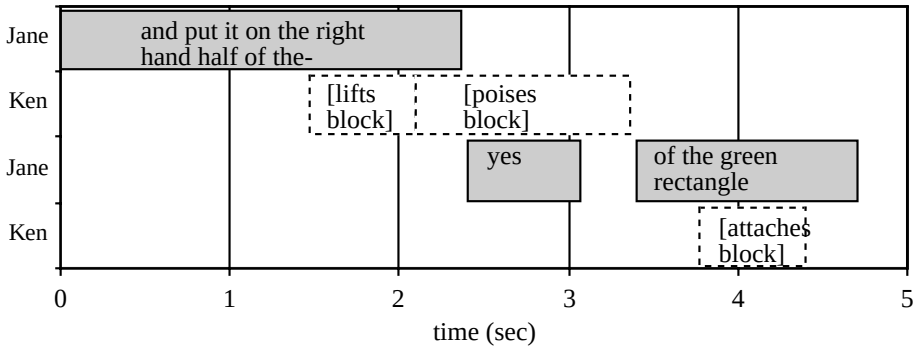

Fig. 7. Action graph of self-interruption to confirm a poised block.

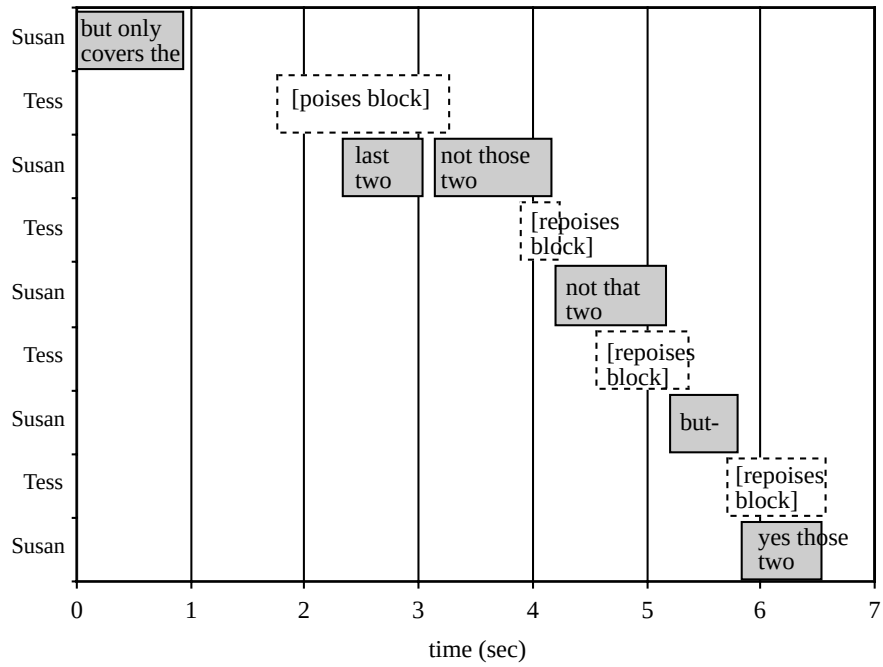

Fig. 8. Action graph of four poises with immediate responses.

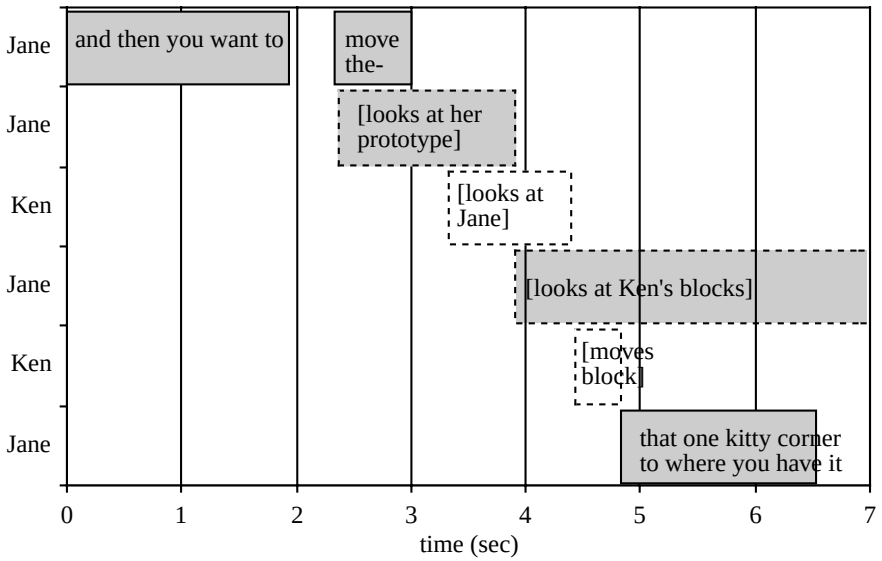

Fig. 9. Action graph of establishing a reference mid-utterance.

Wasow, 1998), or pause (Clark & Fox Tree, 2002; Goldman-Eisler, 1968). These self-interruptions are initiated to deal with the speakers' own problems. But self-interruptions should also be initiated to deal with evidence from addressees. By the immediacy constraint, speakers are expected to respond to this evidence immediately, so if the evidence comes mid-utterance, they must interrupt themselves to make their responses immediate. Indeed, for the 332 builders' actions in Table 2, directors interrupted themselves at least 44 times, marked in our examples by a trailing hyphen (-). They may also have interrupted themselves at locations we could not distinguish from phrase completions.

Fig. 7 shows Jane interrupting her utterance to deal with Ken's gesture. As Jane speaks, she monitors Ken's actions, and the moment he poises his block at the right location, she interrupts her fluent utterance-so-far to complete the projective pair:

Ken	[poises block, meaning "Does it go here?"]
Jane	Yes

She then returns to complete her sentence, "of the green rectangle." Jane is put into a predicament the moment Ken poises his block. If she does not say "yes" immediately, he may infer that the location is wrong. To prevent that, she suspends speaking 0.3 s after his poise, moving from "of the-" to "yes" without a pause.

In Fig. 7, Jane returns to complete her sentence, but most directors did not. For the 44 clear self-interruptions in our data, directors returned to complete their sentences only 36% of the time. Often, they stopped to accept or reject an exhibit, poise, or placement, as here:

Donna	They're not on top of each other they're | right next-
Beth	[exhibits the model-so-far]
Donna	Yeah, like that.

Beth initiates her exhibit coincident with "right next-" and .47 s after its initiation, Donna interrupts herself to confirm it, "Yeah like that." There is no pause between "right next-" and "yeah like that." Other times, directors abandoned their original construction and made a fresh start, as here: "put them so that- [B positions two blocks] at a right angle, but the yellow four hanging over the green four."

Collaborative reference. Gestural signals are necessarily bilateral. In one instance, June points at two raised dots on a Lego block and asks Kay, "These two?" June's reference is a composite of "these two" plus pointing. In performing the gesture, she must make sure Kay sees it, or continue until she *does* see it. Many deictic references in our transcripts relied on the addressees' prior gesture. In one example, Donna (the director) says "so it takes up the center-" and when Beth (the builder) poises a block, Donna interrupts herself to reply, "right there." Donna's deictic reference relies on Beth's current gesture.

One strategy that exploits gestural collaboration is the *extended* collaborative reference. Fig. 8 represents an example in which Susan is telling Tess where to put the next block. They accomplish this as Tess poises her block over four successive locations and as Susan rejects each location except the last one. Tess begins each new poise only 0.6, 0.3, and 0.4 s after Susan has rejected the previous one. Likewise, Susan begins her rejection only 0.45, 0.2, and 0.5 s after Tess's first three poises. Within

just 4 s, the two of them ground what Susan is saying by means of four projective pairs—poise plus uptake—performed fluidly and efficiently.

Extended collaborative references were frequent in our 16 video clips. Builders poised blocks for the first or only time on 73 occasions, but went on to *re-poise* the blocks on 61 occasions. In all, they poised a block an average of 1.8 times (range 1–6). Also, once builders had positioned one or two blocks on the table, they *re-positioned* them 45 times. Not all pairs proceeded as Susan and Tess did. Other builders let directors give fuller descriptions before trying to poise or position their blocks the first time.

Visual monitoring

If speaking is bilateral, speakers should monitor addressees both auditorily and visually. In our task, when the workspace was visible, directors were usually checking as they spoke on what the builders were doing. In the eight videotapes with both faces and workspaces visible, directors spent 55% of the time looking at the builders' workspace (or face), and the rest of the time examining their own prototype. (At a distance of 2 m, directors could see the partner's workspace and face with little change in gaze.) They gazed an average of 2.56 s at the workspace and 1.88 s at the prototype. The difference is not significant, $t(7) < 1$. For their part, builders spent 83% of the time focused on their own blocks and only 17% of the time looking at directors. They averaged 5.56 s gazing at their own blocks and only 1.00 s gazing at directors. This difference is significant, $t(7) = 4.84$, $p < .002$. So directors and builders paid most attention to the regions with the most information for their work: their workspaces.

When builders glanced up at directors' faces, it was often to see why directors were having problems. In one instance, Jane tells Ken to put a block "so it'll go exactly diagonal to where you had it," and Ken places it where he believes it should go. But after 1.4 s, Jane still has not confirmed the location, so Ken looks up at her. She seizes the opportunity and shakes her head no. She does not shake her head until she sees he can attend to it.

A more striking example is represented in Fig. 9. While Jane is looking at Ken's blocks, she begins "and then you want to," pauses 0.4 s, continues "move the-," and then suspends speaking, looking down at her prototype. All this is evidence that she is having trouble formulating a description for the next block. Ken appears to notice her difficulty and tries to deal with it. First, he looks up at her .33 s after her suspension. Then, .53 s after she looks back at his blocks, he, too, looks back at his blocks and moves the one he apparently thinks she is trying to refer to. Whether or not it was the right block, she seizes the opportunity and says "that one." That is, she abandons her previous noun phrase, which might have come out "the blue one on the left," and *instantly* formulates a new deictic expression, "that one," exploiting their just-altered state of knowledge.

In brief, people take others' eye gaze as evidence of what they are attending to (level 1 in the four levels of joint actions) and thinking about (e.g., Goodwin, 1981). They use that evidence in determining the course of their current utterance or action.

Discussion

People engaged in joint activities have to work together to succeed. In our task, two people built Lego models together. The director knew what to build and the builder did the assembling. The two of them were fastest when the director could see the builder's workspace. They took twice as long when the director could not see it. And they made eight times as many errors when they could not monitor each other at all. How are we to account for these findings?

The argument is that people ordinarily try to ground what is said, and grounding is often most efficient when they can monitor each other's voices, faces, gestures, and workspaces (Clark & Brennan, 1991). People can sometimes compensate when prevented from monitoring each other, but at a cost. Monitoring addressees' faces, it has been shown, does not usually make grounding more efficient in task-oriented dialogues (see Whittaker, 2003, for a review). Although people do make use of eye-gaze and head gestures when visible, as our results showed, that did not lead to measurably greater efficiency in our task. Monitoring the addressees' workspaces, on the other hand, *is* critical, and in our task, preventing it doubled the time needed. And preventing all monitoring of others led to eight times as many errors. So in tasks like ours, speakers can compensate, usually at a time cost, when prevented from monitoring the workspaces of their addressees, but they cannot fully compensate when prevented from monitoring addressees altogether.

Our findings have general implications for models of speaking. Perhaps the most basic is that speakers and listeners do not use the same processes in dialogue—the primary site for language—as they do when they are alone. When speakers are alone, speaking is necessarily unilateral. But when they are in a dialogue, speaking is normally bilateral. Let us look at what makes the bilateral process different.

Updating common ground

In dialogue, common ground is updated continuously. When Alan and Beth are face to face, they ordinarily have continuous access to each other's voices, faces, gestures, and workspaces. They can take most of this information as common ground because it is visually or

auditorily present to the two of them, and they are openly co-present (Clark & Marshall, 1981). So when both Alan and Beth can see Beth's workspace, they can immediately take Beth's moving a block as common ground. Updating common ground does not come at intervals, or at the ends of utterances. Within the limits of processing, it is both instantaneous and continuous.

The updating of common ground was exploited at many points in our task. In Fig. 7, Ken poises a block for Jane to verify, and she interrupts herself just 0.3 s later to do that ("yes"). In Fig. 8, Susan and Tess carry out four such exchanges, each within 0.5 s of the previous action. So Ken's and Tess's poisings are *instantly* taken to be common ground. In Fig. 9, Ken looks at Jane to see that she is looking at his blocks before he moves them 0.4 s later. He makes sure that his move will be visually co-present to the two of them. She, in turn, takes his move to be common ground when she refers to the block a mere 0.3 s later as "that one." Deictic expressions such as *that one* or *like this?* require speakers to monitor what addressees are attending to. Most of the techniques we have described depend on the instant updating of common ground.

Speaking and listening are incremental processes, and in dialogue, many of the increments are determined jointly. Speaking is incremental at many levels (Levelt, 1989). In one study (Griffin, 2001), people were observed as they described a picture of, for example, a clock and television set above a needle. They gazed at the clock and began producing "the clock" even before gazing at the television set, after which they produced "and the TV..." They formulated the utterance in increments. Listening is also incremental at many levels (Marslen-Wilson, 1987; Spivey et al., 2002, Spivey-Knowlton et al., 1998, Tanenhaus & Trueswell, 1995). In one study (Tanenhaus & Spivey-Knowlton, 1996), people sat at a table with, for example, a candy and a box on it and were asked "Pick up the candy." They began looking at the candy even before the end of the word "candy." When the box was replaced with a candle, they took longer because they had to hear more of "candy" to distinguish it from "candle." Listeners generally try to resolve interpretations at the earliest opportunity, a thoroughly incremental process.

The thrust of our findings is that in dialogue many of these increments are determined jointly by speakers and addressees. In Fig. 8, Susan tells Tess "but only covers the [Tess poises block] last two, not those two, [Tess re-poises block] not that two, [Tess re-poises block] but, [Tess re-poises block] yes those two." Susan cannot *in principle* formulate "not those two" or "not that two" until Tess has poised her block, yet she initiates her speech within 0.4 s after the onset of Tess's two poises. In just 4 s, Susan formulates her utterance in at least six increments, many contingent on Tess's actions. In the same 4 s, Tess visibly revises her understanding four times, each time contingent on Susan's actions. The same point could be made for Figs. 6, 7, and 9.

Bilateral processes

In dialogue, addressees are normally expected to let speakers know as they go along about their understanding of the current utterance. Addressees can in principle take one of three actions while they are listening. (1) They can *tell* speakers about the current state of their understanding whenever they think it would be useful. (2) They can *allow* speakers moment-by-moment access to evidence of their current understanding. Or (3) they could be *indifferent* to speakers. Unilateral models such as Searle's (1992) make no mention of what addressees might try to do. Bilateral models assume that addressees take an active part both: (1) by telling speakers about their understanding and (2) by giving them access to evidence of understanding.

Our video analyses show that addressees do both (1) and (2). Builders actively told addressees about the current state of their understanding by exhibiting and poising blocks, by pointing, by nodding or shaking their heads, and by using eye gaze—all while utterances were in progress. They also provided directors with access to what they were doing. As directors spoke, builders would manifestly move, place, or reach for blocks at the expected time, or *not* do so, and directors took this as evidence of understanding, or the lack of it. Directors and builders apparently *expected* the builders to take active part with feedback of types (1) and (2).

Speakers work bilaterally whenever they can. To speak bilaterally is to rely on both self- and other-monitoring, and to speak unilaterally is to rely on self-monitoring alone. When it is impossible for people to monitor each other, as in our non-interactive condition, speakers have to work unilaterally. But in everyday conversation, people can almost always monitor each other in *some* way. On the telephone, they can monitor each other's voices, and face to face, they can also monitor each other's faces, bodies, workspaces, and shared scenes. And we show that they monitor others at all levels of joint action—from attention up to considering what to do next.

Not only are people *able* to monitor each other, but they are expected to do so when possible. Ordinarily, it is uncooperative, even rude, not to do so. In Fig. 8, Susan would have been uncooperative if she had ignored the options Tess offered. In Fig. 7 Jane would have been uncooperative if she had ignored Ken's poise just because it came mid-utterance. In Fig. 6, Sam would have been uncooperative if he had continued his instruction before Ted was ready. One example illustrates the consequences of not cooperating. Helen starts, "And take the green piece," to which Molly responds "m-hm" but does not move to pick up the green piece directly in front

of her. This leads Helen to repeat, "the rectangular green piece," and although Molly nods and says "m-hm," she still makes no move. Only 6.2 s later, after the next instruction, does Molly reach for the block. Molly's non-response led Susan to make an unnecessary repeat.

Opportunistic processes

Speech planning is opportunistic. If speaking is incremental, speakers should take advantage of opportunities that arise mid-utterance, and they do (see also Clark & Schaefer, 1989). We have identified four strategies in dialogue that depend on these opportunities.

1. *Offering options*. In Fig. 8, as just noted, Tess offers Susan four interpretations, four poisings, in succession. Susan says no to the first three, but "yes those two" to the fourth. It is only when Susan and Tess jointly settle on an interpretation that they stop. This strategy depends on the two working together: one offers plausible options, and the other seizes the opportunity when the right option comes up, which stops the process.
2. *Self-interruptions*. In Fig. 7, Ken poises a block over the right location halfway through Jane's utterance, and she seizes the opportunity by interrupting herself to say yes: "and put it on the right hand half of the- *yes* of the green rectangle." She would have lost the opportunity if she had not.
3. *Waiting*. In Fig. 6, Sam waits until Ted is has retrieved and exhibited the right block before he continues with his instruction ("and join the two"). He waits for the right opportunity to proceed.
4. *Instant revisions*. In Fig. 9, Jane is having trouble describing a block when Ken moves a particular block. At that moment she abandons her current noun phrase ("the..."), seizes the opportunity offered by Ken's move, and formulates the deictic expression "that one." She could not have done that if he had not created the opportunity.

In all four strategies, speakers changed course by seizing opportunities made available by their partners—whether the opportunities were made intentionally or not. Speakers made these alterations instantly, typically initiating them within a half a second of the opportunities becoming available.

These are four of many opportunistic strategies that have been identified in the literature (see Brennan, 1990; Clark, 1996; Clark & Wilkes-Gibbs, 1986; Goodwin, 1986b; Goodwin & Goodwin, 1986; Sacks et al., 1974; Sacks & Schegloff, 1979; Schegloff et al., 1977). What makes these different is that the opportunities arise not in speech, but in gestural acts and other visible actions.

Speakers reformulate what they say mid-utterance to deal not only with their own problems, but with issues originating in their addressees. Spontaneous speech is replete with self-repairs that deal with problems in planning and production (Levelt, 1983, 1989; Schegloff et al., 1977). Here is an example from the workspace hidden condition:

> Duncan *You are going to um* okay you're going to make both of them *into yel-* into the L shape.

Duncan begins "you are going to," apparently runs into planning problems, and starts again, replacing the words in italics. Later he replaces "into yel-" with "into the L shape."

But many phenomena that *look* like self-repairs are actually revisions based on other-monitoring—*other-revisions*. Consider Jane's utterance in Fig. 9:

> Jane And then you want to move *the-* that one kitty corner to where you have it.

With only an audio recording, we might conclude that Jane was making a self-repair. But as shown in Fig. 9, she reformulates what she is saying mid-utterance in response to Ken's moving of a block. It is an other-revision. Or consider this example:

> Dawn Put it at *the end of the red that's o-* the other end.

The video clip shows Dawn reformulating her utterance in response to her partner Betty's pointing (see also Goodwin, 1981). Speakers, therefore, may reformulate what they are saying mid-utterance either because of problems discovered in self-monitoring or because of information acquired in other-monitoring.

Multi-modal processes

In dialogue, the participants use vocal and gestural modalities in parallel. People in conversation normally take turns speaking (Sacks et al., 1974). They do so apparently because it is difficult to hear and understand two lines of speaking at once, because it is difficult to listen and speak at once, or both. With important exceptions, people in many cultures prefer to speak in the clear.

These same people, however, seem perfectly happy to *gesture* while others are speaking. In Figs. 6–8, builders begin to poise or exhibit Lego blocks while directors are still speaking. Not only do the gestures and speech overlap, but directors often respond while continuing to speak. That is, people are able to communicate, to some degree at least, by speech and gesture in parallel: separately *and* simultaneously.

The visual modality is faster and more secure than the auditory modality for certain types of communication. In the workspace hidden condition, when directors described a block or location, they and their partners often spent much time grounding that description. Mike, for example, described one block as "a green one that's like

four, like the two rows of two." In response, Nancy asked, "Two rows of two?" to which Mike replied, "Yeah, like the square one," and only then did Nancy say "'Kay." Grounding took 10 words in three turns—and this is one of the briefer examples. When the workspace was visible, grounding was much faster. In one case, Hannah initiates grounding by poising a block, and 0.36 s later, Jeff replies "exactly." Grounding took one move and one word in one speaking turn. Recall that the building cycle was 2.5 times as long when the grounding had to be done verbally—the checking time itself was 11 times as long. The visual modality is highly effective for establishing the identity or placement of material objects (see Clark & Marshall, 1981).

In dialogue, then, the participants work together in determining the course of each utterance. They rely not only on each other's vocal signals, but on each other's gestural signals such as exhibiting, poising, pointing at, and placing physical objects, nodding and shaking heads, and directing eye gaze, and on other mutually visible events. They use the signals to create projective pairs by which they ground what they are currently saying. Dialogues are the artful orchestration of these actions. Models of language use that are limited to only part of this process are necessarily incomplete and, for many purposes, incorrect.

References

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

Argyle, M., Lalljee, M., & Cook, M. (1968). The effects of visibility on interaction in a dyad. *Human Relations, 21*, 3–17.

Atkinson, J. M.& Heritage, J. (Eds.). (1984). *Structures of social action: Studies in conversation analysis*. Cambridge: Cambridge University Press.

Bavelas, J. B., Black, A., Lemery, C. R., & Mullett, J. (1986). I show you how you feel: Motor mimicry as a communicative act. *Journal of Personality and Social Psychology, 50*, 322–329.

Bavelas, J. B., & Chovil, N. (2000). Visible acts of meaning. An integrated message model of language in face-to-face dialogue. *Journal of Language and Social Psychology, 19*, 163–194.

Bavelas, J. B., Chovil, N., Lawrie, D. A., & Wade, A. (1992). Interactive gestures. *Discourse Processes, 15*, 469–489.

Bavelas, J. B., Coates, L., & Johnson, T. (2000). Listeners as co-narrators. *Journal of Personality and Social Psychology, 79*(6), 941–952.

Blackmer, E. R., & Mitton, J. L. (1991). Theories of monitoring and the timing of repairs in spontaneous speech. *Cognition, 39*(3), 173–194.

Bock, K., & Levelt, W. J. M. (1994). Language production: Grammatical encoding. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 945–984). San Diego: Academic Press.

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

Brennan, S.E. (1990). *Seeking and providing evidence for mutual understanding*. Unpublished Ph.D. dissertation, Stanford University.

Button, G.& Lee, J. R. (Eds.). (1987). *Talk and social organisation*. Clevedon, Avon; Philadelphia: Multilingual Matters.

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

Clark, H. H. (1997). Dogmas of understanding. *Discourse Processes, 23*(3), 567–598.

Clark, H. H. (2003). Pointing and placing. In S. Kita (Ed.), *Pointing. Where language, culture, and cognition meet* (pp. 243–268). Hillsdale, NJ: Lawrence Erlbaum.

Clark, H.H. (in press). Pragmatics of language performance. In L.R. Horn & G. Ward, (Eds.), *Handbook of pragmatics*. Blackwells, Oxford.

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

Clark, H. H., & Fox Tree, J. E. (2002). Using *uh* and *um* in spontaneous speaking. *Cognition, 84*(1), 73–111.

Clark, H. H., & Haviland, S. E. (1977). Comprehension and the given-new contract. In R. O. Freedle (Ed.), *Discourse production and comprehension* (pp. 1–40). Hillsdale, NJ: Erlbaum.

Clark, H. H., & Marshall, C. R. (1981). Definite reference and mutual knowledge. In A. K. Joshi, B. L. Webber, & I. A. Sag (Eds.), *Elements of discourse understanding* (pp. 10–63). Cambridge: Cambridge University Press.

Clark, H. H., & Schaefer, E. F. (1987). Collaborating on contributions to conversations. *Language & Cognitive Processes, 2*(1), 19–41.

Clark, H. H., & Schaefer, E. R. (1989). Contributing to discourse. *Cognitive Science, 13*, 259–294.

Clark, H. H., & Wasow, T. (1998). Repeating words in spontaneous speech. *Cognitive Psychology, 37*(3), 201–242.

Clark, H. H., & 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.

Drew, P.& Heritage, J. (Eds.). (1992). *Talk at work: Interaction in institutional settings*. Cambridge: Cambridge University Press.

Ekman, P., & Friesen, W. (1969). The repertoire of nonverbal behavior: Categories, origins, usage and coding. *Semiotica, 1*, 49–98.

Engle, R. A. (1998). Not channels but composite signals: Speech, gesture, diagrams, and object demonstrations are integrated in multimodal explanations. In M. A. Gernsbacher & S. J. Derry (Eds.), *Proceedings of the twentieth annual conference of the cognitive science society*. Mahwah, NJ: Erlbaum.

Engle, R.I. (2000). *Toward a theory of multi-modal communication: Combining speech, gestures, diagrams, and demonstrations in instructional explanations*. Unpublished Ph.D. dissertation, Stanford University.

Ferreira, F. (2000). Syntax in language production: An approach using tree-adjoining grammars. In L. Wheeldon (Ed.), *Aspects of language production* (pp. 291–330). Philadelphia, PA: Psychology Press/Taylor & Francis.

Frazier, L., & Clifton, C. (1996). *Construal*. Cambridge, MA: MIT Press.

Gale, C., & Monk, A. F. (2000). Where am I looking? The accuracy of video-mediated gaze awareness. *Perception & Psychophysics, 62*(3), 586–595.

Garrett, M. F. (1980). Syntactic processes in sentence production. In B. Butterworth (Ed.), *Speech production* (pp. 170–220). New York: Academic Press.

Geluykens, R. (1987). Tails (right-dislocations) as a repair mechanism in English conversation. In J. Nuyts & G. de Schutter (Eds.), *Getting one's words into line: on word order and functional grammar* (pp. 119–129). Dordrecht: Foris.

Geluykens, R. (1988). The interactional nature of referent-introduction. *Papers from the 24th regional meeting of the Chicago Linguistic Society*. Chicago Linguistic Society, Chicago, pp. 151–164.

Goldberg, J. (1975). A system for the transfer of instructions in natural settings. *Semiotica, 14*, 269–296.

Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in spontaneous speech*. New York: Academic Press.

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

Goodwin, C. (1986a). Between and within: Alternative sequential treatments of continuers and assessments. *Human Studies, 9*, 205–217.

Goodwin, C. (1986b). Gestures as a resource for the organization of mutual orientation. *Semiotica, 62*, 29–49.

Goodwin, M. H., & Goodwin, C. (1986). Gesture and coparticipation in the activity of searching for a word. *Semiotica, 62*, 51–75.

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

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics, Vol. 3: Speech acts* (pp. 113–128). New York: Seminar Press.

Grice, H. P. (1991). *In the way of words*. Cambridge, MA: Harvard University Press.

Griffin, Z. M. (2001). Gaze durations during speech reflect word selection and phonological encoding. *Cognition, 82*(1), B1–B14.

Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition, 59*(1), 91–117.

Jefferson, G. (1972). Side sequences. In D. Sudnow (Ed.), *Studies in social interaction* (pp. 294–338). New York, NY: Free Press.

Jefferson, G. (1973). A case of precision timing in ordinary conversation: Overlapped tag-positioned address terms in closing sequences. *Semiotica, 9*, 47–96.

Jefferson, G. (1989). Preliminary notes on a possible metric which provides for a standard maximum silence of approximately one second in conversation. In D. Roger & P. Bull (Eds.), *Conversation* (pp. 166–196). Clevedon: Multilingual Matters.

Kempen, G., & Hoenkamp, E. (1987). An incremental procedural grammar for sentence formulation. *Cognitive Science, 11*(2), 201–258.

Kendon, A. (1967). Some functions of gaze direction in two-person conversation. *Acta Psychologia, 16*, 22–63.

Kendon, A. (1993). Human gesture. In K. R. Gibson & T. Ingold et al. (Eds.), *Tools language and cognition in human evolution* (pp. 43–62). Cambridge: Cambridge University Press.

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

Kraut, R.E., Gergle, D., & Fussell, S.R. (2002). *The use of visual information in shared visual spaces: Informing the development of virtual co-presence*. Paper presented at the conference on computer supported cooperative work.

Kraut, R. E., & Johnston, R. E. (1979). Social and emotional messages of smiling: An ethnological approach. *Journal of Personality and Social Psychology, 37*, 1539–1553.

Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition, 14*, 41–104.

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

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. Special issue: Spoken word recognition. *Cognition, 25*(1–2), 71–102.

McNeill, D. (1992). *Hand and mind*. Chicago: University of Chicago Press.

Sacks, H., & Jefferson, G. (1992). *Lectures on conversation*. Oxford, UK; Cambridge, MA: Blackwell.

Sacks, H., & Schegloff, E. (1979). Two preferences in the organization of reference to persons in conversation and their interaction. In G. Psathas (Ed.), *Everyday language: Studies in ethnomethodology* (pp. 15–21). New York: Irvington Publishers.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking in conversation. *Language, 50*, 696–735.

Schegloff, E. A. (1982). Discourse as an interactional achievement: Some uses of uh huh and other things that come between sentences. In D. Tannen (Ed.), *Georgetown University Roundtable on Languages and Linguistics 1981*, *Analyzing discourse: Text and talk* (pp. 71–93). Washington, DC: Georgetown University Press.

Schegloff, E. A. (1991). Conversation analysis and socially shared cognition. In L. B. Resnick & J. M. Levine et al. (Eds.), *Perspectives on socially shared cognition* (pp. 150–171). Washington: American Psychological Association.

Schegloff, E. A. (1998). Body torque. *Social Research, 65*(3), 535–596.

Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language, 53*, 361–382.

Schegloff, E. A., & Sacks, H. (1973). Opening up closings. *Semiotica, 8*, 289–327.

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

Searle, J. R. (1992). Conversation. In J. R. Searle, H. Parret, & J. Verschueren (Eds.), *(On) Searle on conversation* (pp. 7–29). Amsterdam, Philadelphia: J. Benjamins Pub. Co.

Sperber, D., & Wilson, D. (1986). *Relevance*. Cambridge, MA: Harvard University Press.

Spivey, M. J., Tanenhaus, M. K., Eberhard, K. M., & Sedivy, J. C. (2002). Eye movements and spoken language compre-

hension: Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology, 45*(4), 447–481.

Spivey-Knowlton, M. J., Tanenhaus, M. K., Eberhard, K. M., & Sedivy, J. C. (1998). Integration of visuospatial and linguistic information: Language comprehension in real time and real space. In P. Olivier & K.-P. Gapp (Eds.), *Representation and processing of spatial expressions* (pp. 201–214). Mahwah, NJ: Lawrence Erlbaum Associates.

Streeck, J. (1993). Gesture as communication: I. Its coordination with gaze and speech. *Communication Monographs, 60*(4), 275–299.

Streeck, J. (1994). Gesture as communication II: The audience as co-author. *Research on language and social interaction special issue: Gesture and understanding in social interaction, 27*(3), 239–267.

Svartvik, J.& Quirk, R. (Eds.). (1980). *A corpus of English conversation*. Sweden: GleerupLund.

Tanenhaus, M. K., & Spivey-Knowlton, M. J. (1996). Eye-tracking. *Language & Cognitive Processes, 11*(6), 583–588.

Tanenhaus, M. K., & Trueswell, J. C. (1995). Sentence comprehension. In J. L. Miller & P. D. Eimas (Eds.), *Handbook of perception and cognition (2nd ediition): Speech language and communication* (pp. 217–262). San Diego: Academic Press.

Traum, D. (1994). *A computational theory of grounding in natural language conversation*. Unpublished Ph.D. dissertation, Department of Computer Science, University of Rochester.

Traxler, M. J., & Gernsbacher, M. A. (1992). Improving written communication through minimal feedback. *Language & Cognitive Processes, 7*(1), 1–22.

Traxler, M. J., & Gernsbacher, M. A. (1993). Improving written communication through perspective-taking. *Language & Cognitive Processes, 8*(3), 311–334.

Whittaker, S. (2003). Things to talk about when talking about things. *Human Computer Interaction, 18*(1–2), 149–170.