Visual Co-Presence, Coordination Signals, and Partner Effects in Spontaneous Spoken Discourse

Susan E. Brennan

Psycholinguistic experiments have shed much light on the distinct processes of listening and speaking. However, less is known about how these processes are coordinated in the service of language use in conversation, both between individuals who alternate in their roles as speaker and hearer, and within the mind of the individual language user who is processing on many levels at once. This article addresses the coordination of language use between individuals. I will present a program of research that focuses on how paralinguistic cues and processes contribute to communication, including: how conversation is shaped by communication media through the process of grounding; how devices such as hedges, latencies to responses, and intonation may be used in coordination; how disfluencies are distributed in referential communication and how they affect the interpretation of utterances; and how the choice of particular expressions in conversation reflects partner-specific adjustments between speakers and addressees. These paralinguistic aspects of language use bear on how people manage to repair problems in speaking and understanding in order to converge on the same perspective.

Keywords: conversation, coordination, speech disfluencies, repair, paralinguistic cues, hedges

Introduction

Language use in conversation involves not only producing and comprehending utterances, but also coordinating these two sets of cognitive processes. Speaking and listening are coordinated in two ways: between interlocutors, as they take turns in a conversation and in so doing, seek and provide evidence that they understand one another, and within the mind of each interlocutor, who is simultaneously processing language on many levels at once. At a given moment, a speaker may be simultaneously articulating a constituent already planned, monitoring whether it is coming out as intended, choosing morphemes to fill in the next constituent in the plan, monitoring back-channels from the addressee to estimate his understanding and uptake, and halting to repair when she detects a problem with her own utterance. Her addressee is presumably interpreting the words in her ongoing utterance, trying to recognize her intentions, displaying precisely timed feedback responses, estimating whether his and the speaker’s intentions are converging sufficiently (and initiating a repair when they are not), planning his own next utterance, and determining when to begin articulating it. Both of them keep track of what has been discussed so far, attend to social needs such as face-management, and display, to some degree, a joint focus of attention upon the topic or task at hand.

Engaging all these processes simultaneously demands a remarkable degree of coordination. In trying to understand discourse processes, the field of experimental psycholinguistics has dis-
covered much about discourse comprehension, somewhat less about utterance production, a bit about how speaking and understanding are coordinated between pairs of interlocutors, and very little indeed about how the coordination of speaking and listening takes place at the same time within an individual mind. In this article, I will present findings on the coordination of discourse processes between pairs of interlocutors.

The Process of Grounding

Our studies of language processing typically have two characteristics: First, we observe speakers and addressees interacting, since our goal is to study language processes in situ.

Second, pairs of people communicate in order to do a specific task that has observable and incremental products. This gives the experimenters behavioral evidence about what people in conversation mean and understand at any given moment, independent from the linguistic forms of their utterances. The goal is to achieve a balance between naturalness and experimental control. Data typically include the taped transcripts of the conversations from which linguistic and paralinguistic events are coded, as well as performance measures such as time to complete the task, errors, and events such as pointing, reaching, moving objects, pressing buttons, typing, and gazing.

Consider the following exchange from one of our referential communication experiments. Two people, A and B, are visually separated by a barrier while doing a referential communication task, matching identical sets of cards displaying pictures of abstract geometric shapes. A is the director and her cards are arranged in the target order; B’s cards are arranged in a random order and they need to converse in order to get B’s cards arranged to match A’s.

(1) A: ah boy this one ah boy
all right it looks kinda like-
on the right top there’s a square that
looks diagonal
B: uh huh
A: and you have sort of another like rectan-
gle shape, the-
like a triangle, angled, and on the bottom
it’s uh
I don’t know what that is, glass shaped
B: all right I think I got it
A: it’s almost like a person kind of in a weird
way
B: yeah like like a monk praying or some-
ting
A: right yeah good great
B: all right I got it

(Stellmarn & Brennan, 1993)

Several things are apparent from this exchange. First, there are a number of interruptions of fluent speech; the director restarts her first turn twice and her second turn at least once. These disfluencies do not appear to prevent the matcher from understanding, however. Second, the director pauses occasionally, uses fillers like uh, and displays her commitment (or lack of commitment) to what she is saying with displays like ah boy this one ah boy and I don’t know what that is. In addition to displaying her degree of commitment to what she is saying and sometimes marks an utterance as provisional, dependent upon feedback from the matcher. Third, the director proposes a perspective on the object in a series of installments, punctuated with backchannels from the matcher. In this way, the director and matcher appear to establish meaning incrementally, producing utterances that are not fully planned at the time they begin speaking. Fourth, both parties appear to take responsibility for the perspective they establish. In this example, even though the director is the one who knows what the target picture is, it is the matcher who proposes the description that they both end up ratifying: like a monk praying or something. This example is quite typical; in this particular referential communication study, the 24 pairs of partners who discussed this object ended up with 24 different but mutually agreed-upon perspectives.
on it.

While there is much variability between conversations, there is little variability within a conversation. In the experiment from which this example was drawn, once the director and matcher matched each of the 12 picture cards in the set, the cards were shuffled, A and B's director/matcher roles were switched, and the task was repeated. In the very next round, the conversation went like this:

(2) B: nine is that monk praying
   A: yup

Later on, referring was even more efficient:

(3) A: three is the monk
   B: ok

A and B marked the fact that they had achieved a mutual perspective by reusing the same term, monk, in repeated references to the same object. In addition, they tended to shorten their referring expressions over time. In Brennan and Clark (1996), we showed that once people establish a perspective on an object, they tend to continue using the same terms to mark that perspective (e.g., the man's pennyloafer), even when they could use a shorter and more common basic-level term (e.g., the shoe, when the set of objects has changed such that it no longer needs to be distinguished from other shoes in the set). This process of lexical entrainment appears to be partner-specific to some degree. That is, upon repeated reference to the same object but with a new partner, speakers were more likely to revert to the basic level term when they no longer needed to distinguish the object from similar ones (Brennan & Clark, 1996). For that reason, we have proposed that people in conversation establish conceptual pacts, or temporary, flexible agreements to see an object in a particular way. A conceptual pact may be maintained perceptually through the feedback a speaker receives from an addressee, or through a specific mental representation by the speaker of what the addressee knows (these two possibilities are of course not mutually exclusive). The extent to which entrainment is truly partner-specific (or derived from a conceptual pact with a particular individual) as opposed to emerging from what is least effortful for the individual interlocutor has been a matter of debate, and I will return to this issue in the last section of this article.

To refer successfully in conversation, people achieve a joint focus of attention that enables them to be confident that they both have similar meanings in mind. Grounding is the process by which people seek and provide evidence that their meanings are sufficiently aligned, that is, that they understand one another well enough for the purposes at hand. When they cannot see one another, such as over a telephone, they must use entirely verbal means to do this. Many referential communication studies focus on verbal and linguistic communication strategies, and so are conducted in the absence of visual co-presence — that is, directors and matchers are not able to see one another. When there is some degree of visual co-presence (such as when people converse face-to-face or when both can see relevant parts of a task they are doing together), grounding takes quite a different form. Visual co-presence can afford relatively direct information about what a partner means or understands, how the task is going, what the partner is attending to, what might be meant by a particular referring expression, what subgoals in a task are currently relevant, and when input into the conversation or task is expected, invited, or possible. With some degree of visual co-presence, partners can often use deictic (pointing) strategies to formulate and ground references to objects. For conversations about physical tasks, this is often faster and more efficient than using speech alone.

To consider the effects of visual co-presence on communication, it is useful to conceptualize communication as a set of activities that include getting a conversation started, formulating and producing utterances, receiving and understand-
### Table 1
Affordances of Four Communication Media (adapted from Clark & Brennan, 1991 and from Kraut, Fussell, Brennan, & Siegel, in press). Key: ++ means that the affordance is present in that medium, + means present to a limited degree, ?? means may or may not be present, depending on the particular instantiation, and -- means not available in that medium.

<table>
<thead>
<tr>
<th>Affordances of Media</th>
<th>Medium</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Face-to-face</td>
</tr>
<tr>
<td>1. Visual co-presence: Participants share the same physical environment, including a view of what each other is doing and looking at.</td>
<td>++</td>
</tr>
<tr>
<td>2. Visibility: Participants can see each other but not what each other is doing or looking at.</td>
<td>++</td>
</tr>
<tr>
<td>3. Audibility: Participants can hear one another.</td>
<td>++</td>
</tr>
<tr>
<td>4. Cotemporality: Messages are received close to the time that they are produced, permitting fine-grained interactivity.</td>
<td>++</td>
</tr>
<tr>
<td>5. Simultaneity: Multiple participants can send and receive messages at the same time, allowing backchannel communication to occur.</td>
<td>++</td>
</tr>
<tr>
<td>6. Sequentiality: Participants take turns in a fairly orderly fashion within a single conversation at a time; one turns’ relevance to another is signaled by adjacency.</td>
<td>++</td>
</tr>
<tr>
<td>7. Reviewability: Messages do not fade over time.</td>
<td>--</td>
</tr>
<tr>
<td>8. Revisability: Messages can be revised before being sent.</td>
<td>--</td>
</tr>
</tbody>
</table>

ing them, taking turns, displaying understanding, and repairing any problems that arise (see the framework presented in Clark & Brennan, 1991). Each of these activities has its own costs, and the costs differ according to the affordances of the particular communication medium. The affordances of four common communication media are shown in Table 1.

For instance, producing an utterance is more costly (for most people) when it must be typed than when it can be spoken aloud. Media that afford some degree of visual co-presence (such as face-to-face conversation, video teleconferencing, and electronic media that allow remote partners to access and manipulate the same document or web page simultaneously) offer advantages for tasks that involve manipulating objects. Cues that are easy to transmit or pick up face-to-face may be quite difficult to use electronically, such as facial expressions and expressive gestures, pointing, the direction and timing of eye gaze, eye contact, shared views of the task, partners’ actions produced in response to an utterance or instrumental to a task, and the ultimate results of a partner’s actions. Not all of this visual evidence is equally important; studies of electronic communication have shown that being able to see what a partner is looking at is more important than being able to see the partner’s face during a collaborative task (e.g., Whittaker & Geelhoed, 1993).

A substantial benefit of having full physical co-presence such as the sort afforded by face-to-face conversation is the relative ease of achieving and maintaining a joint focus of attention. Face-to-face, one person need only start speaking in order to get another’s attention. In comparison, start-up costs are significantly higher in teleconferencing if people must be summoned, or if the video setup is asymmetrical and one person has trouble knowing when another is on-line. Face-to-face, one person can see where another is looking and can often see precisely what the other person is looking at. Electronically, even if two partici-
pants have some visual co-presence in the form of shared objects or documents, they may not be able to tell exactly what the other is attending to or when one of them might be also attending to something outside of the shared electronic environment.

Next I will discuss a study that examined how verbal and visual evidence shapes conversation moment by moment, through the grounding process.

**Grounding and Visual Co-presence**

In a referential communication study, I compared task-oriented conversations in which one person either had or didn't have visual evidence about the other's progress (Brennan, 1990, in press). Directors and matchers in adjoining cubicles discussed various locations on identical maps displayed on networked computer screens. The task was for the matcher to get his car icon parked in the same target location as the car icon displayed on only the director's screen. In one condition, Visual Evidence, the director could see the matcher's car icon and its movements. In the other, Verbal Evidence Only, she could not. In both conditions, they could talk freely in order to do the task.

Consider this spontaneous exchange from the Verbal Evidence Only condition (asterisks show overlap across turns, and periods, pauses):

(4) D: you're in the upper
    far far upper right-hand corner of the
    screen it says Sea Street?
M: *yah*
D: *way* at the top?
M: yeh
D: you're
    you're just a little bit on the road
and the
    corner of your car is touching A
of Sea
but you're mostly off the road
the road is to your right
just a- touching *the car*
    M: *the road* is to the right of the car?
    D: put the road-
        put the car right on the road
        and you'll overlap me.
    M: ok

In this example, D and M used entirely verbal means to establish a referent incrementally and opportunistically, and they shared the responsibility for doing so. M's icon was actually in the correct target location early on, by the fourth turn in this exchange, when he uttered his second acknowledgement, yeh. But they still needed to go through a lengthy phase during which they grounded their meaning verbally. It was up to M to propose when he thought he understood well enough for current purposes.

Now consider Example 5, taken from a different pair of speakers doing the same task, but in the Visual Evidence condition: M's cursor appeared superimposed on D's screen as M tried to move his cursor to the target location.

(5) D: ok
    now we're gonna go over to
        M-Memorial Church?
        and park right in Memor-
        right there
        that's *good.*
    M: *that's* rude
        to park in the church.
    D: heheh heh

In this condition, when D had moment-by-moment evidence about M's understanding, verbal descriptions were easier to ground, and the task was much more efficient. M, who knew D could see his cursor, often used its motion to replace his turns in the conversation. And D took responsibility for determining when M understood well enough to move on. In fact, she often interrupted herself mid-utterance to do so; in this example, M's icon reached the target location just before D interrupted herself at “Memor-”.

In this study, a total of twelve pairs of subjects matched the same 80 locations, distributed
between the two co-presence conditions; the exchanges for half of these locations were transcribed, and language-action transcripts were produced for a randomly chosen subset of 10% (48) of these. During each trial, the x and y coordinates of the matcher’s icon were recorded and time-stamped, to provide a moment-by-moment estimate of where the matcher thought the target location was. For the smaller sample of 48 trials, I plotted the distance between the matchers’ icon and the target (the director’s icon) over time, to provide a visible display of how their beliefs about the target location converged.

The sample time-distance plots for Examples (4) and (5) are shown in Figures 1 and 2, respectively. Matchers’ icons got closer to the target over time, but not at an even rate. Typically, distance diminished relatively steeply early in the trial, as the matcher interpreted the director’s initial description and rapidly moved the icon toward the target. Many of the plots then showed a distinct elbow shape, followed by a nearly horizontal region, showing that the matcher then paused or moved away only slightly before returning to park the car icon. Both the time-distance plots and the transcripts suggest that it wasn’t sufficient for the matcher to develop a reasonable hypothesis about what the director meant by the description she presented, but that they also had to ground their understanding by exchanging sufficient evidence in order to establish mutual belief. This process differed according to whether they had some degree of visual co-presence. That is, the region after the elbow appears to correspond to a process during which D and M checked whether their understanding of the location description had converged; this mostly horizontal region appears to correspond roughly to the acceptance phase for contributions to conversations that was proposed by Clark & Schaefer (1989). This grounding or acceptance phase was substantially shorter when directors had visual evidence (Fig. 2) than when they did not (Fig. 1). The accompanying speech transcripts showed that matchers gave verbal acknowledgements when directors did not have visual evidence and withheld them when directors did have visual evidence. That is, matchers adjusted to directors even though the information on the matchers’ own screen was the same in both conditions (conditions alternated or partners switched director/matcher roles after every 10 target locations).

These results document the grounding process by which two people in conversation get their hypotheses to converge, moment by moment. The grounding process is flexible; it is shaped by the affordances of the situation or medium. Part-
ners shift responsibility to whomever can pay a particular cost most easily, which enables them to minimize the effort they expend collectively (Clark & Wilkes-Gibbs, 1986).

**Implications.** This experimental investigation of conversation on-line has implications not only for the coordination between two interacting speakers in a conversation, but also for how speaking and monitoring are coordinated within the architecture of the speech production system. Cascaded models of speech production (e.g., Bock, 1986; Bock & Levelt 1994; Dell, 1986; Levelt, 1989; Hartsuiker & Kolk, 2001) include monitoring loops, with which the system monitors constituents for discrepancies from the ongoing utterance plan. Monitoring processes are both internal and external; they may discover a discrepancy while a given constituent is being planned or after articulation has begun (and the error is heard by both speaker and addressee). When a discrepancy is found before articulation, the constituent is repaired, often resulting in a pause in the ongoing utterance (a covert repair, according to Levelt, 1989). When a discrepancy is found in a constituent during articulation, the speaker typically interrupts the constituent and initiates a repair (Nooteboom, 1980; Levelt, 1983). The interval between the interruption and the repair may contain an editing expression such as *I mean* or *uh*. While there is some debate as to how quickly the system can interrupt itself for repair once a problem has been identified (see Blackmer & Mitton, 1991; Hartsuiker & Kolk, 2001), interruptions often occur within a problematic word itself (Levelt, 1983). I will return to disfluencies and repairs presently.

The point is that speech production architectures should include monitoring not only for whether the utterance being articulated matches the internally represented plan, but also for whether the addressee’s moment-by-moment response to the utterance matches the speaker’s expectations. Either sort of discrepancy affects how the speaker executes and repairs the ongoing utterance.

**Coordination Signals**

People in conversation rely on paralinguistic cues to coordinate their cognitive and linguistic processing. These cues, called *track 2* or *secondary* signals by Clark (1994; 1996), provide information about the ongoing utterance itself (distinct from *track 1*, the “official business” of the utterance). They may be used in a variety of ways, such as to signal the speaker’s commitment (or lack of commitment) toward what she is saying, to invite participation in the utterance being presented, to capture the addressee’s attention, to manage a repair, or to display that the speaker is aware she has just made (or is anticipating making) an error. For example, a speaker may signal that she is having difficulty by interrupting herself (sometimes mid-word), producing an editing expression such as *I mean*, *no*, or *uh*, and restarting the problematic phrase.

**Latency and intonation in question-answering**

When a speaker is answering a question, what she does in the moments before producing the answer is informative about her metacognitive processing. Consider the following two exchanges, from our corpus of spontaneous answers to general knowledge questions (Brennan & Williams, 1995):

(6) Q: Who founded the American Red Cross?

A: ........ *um........ Florence Nightingale*

(7) Q: Who founded the American Red Cross?

A: ................. *Clara Barton*

When speakers are not entirely committed to the answer they are giving, they often pause before responding and produce an interjection such as *uh* or mark the answer with rising intonation (Brennan & Williams, 1995; Smith & Clark, 1993); such answers are less likely to be correct (for the question in (6) and (7), the correct answer is *Clara Barton*). And when
speakers believe they know the answer but just can’t produce it (and are perhaps in a tip-of-the-tongue state), they may pause longer or say *uh* or *um* before answering *I don’t know* (Brennan & Williams, 1995; Smith & Clark, 1993). When such uncertainty cues precede *I don’t know*, they suggest the speaker probably *does* know; those speakers are more likely to recognize the correct answer later on a multiple-choice test (Smith & Clark, 1993). Delays and fillers may serve as an intentional display of the speaker’s retrieval difficulty, or else they may just be symptomatic (in the same sense that smoke is a symptom of fire). However, when listeners hear responses such as those in (6) or (7) and non-answers such as *I don’t know*, presented in isolation from the questions that evoked them, the listeners can interpret metacommunicative information in order to correctly estimate whether speakers do actually know the answer, which we have called the *Feeling of Another’s Knowing* (Brennan & Williams, 1995). People’s signals of their own uncertainty about or commitment to what they are saying, as well as their interpretation of this information in their partners, count as track 2 signals in Clark’s terminology.

**Hedging**

Another cue by which speakers and addressees coordinate is hedging. In referential communication tasks, speakers may use hedges when they first present referring expressions, especially when the objects being matched are novel, ambiguous, have no conventional, lexicalized labels, or lend themselves to many possible perspectives. Recall Example (1), where A proposed *it’s almost like a person kind of in a weird way*, and B responded with *yeah like like a monk praying or something*. Hedges such as *almost like and kind of* tend to drop out upon repeated referring, once the perspectives they accompany have been ratified by the partner (Brennan & Clark, 1996).

Not only are hedges devices for signaling a speaker’s commitment to the perspective she is proposing, but they also serve social needs. Memory, while usually studied as an *intra*personal event, is often an *inter*personal one (Ohaeri, 1998). When a group of people discusses an event they have experienced together, their task is not only to recall the event, but also to serve interpersonal needs such as saving face. By hedging, speakers can display their alignment with what they themselves or their partners are saying. We conducted an experiment (Brennan & Ohaeri, 1999; Ohaeri, 1998) in which we collected conversations generated by 26 three-person groups trying to recall the details of a movie they had watched together (John Sayles’s film, *The Secret of Roan Inish*). Half of the groups interacted face-to-face in front of a computer screen on which they reconstructed the movie, and half did so remotely from different rooms, using a chat program with which they composed messages privately on computer screens and sent the messages to the screens of the others in the group. In Example (8), the speaker used a final rise with a hedge (underlined) to elicit responses from the two partners, while the speaker in (9) elicited responses with a tag question (also underlined):

(8) We all agree it was in Ireland, I *think*?
(9) He was cold and almost dead... the women tied him up between two cows... he started to sweat, *right*?

Speakers may use both devices in the same utterance, as in this example from a text conversation:

(10) We all agree it was a *wreathy thingy* on his neck???

Hedges and phrase-final rising intonation are both devices that can mark an utterance as provisional and invite disagreement or counter-proposals from an addressee; during a disagreement, such responses appear to be less face-threatening since the speaker displays that she is not entirely committed to the utterance (Brennan & Ohaeri, 1999). Upon examining these two
devices, we found that the rate of questions per 100 words (either via intonation or punctuation) was equal in spoken conversation and text-based conversation, while the rate of hedging was substantially higher in spoken than in text conversation. This, we propose, is due to the fact that hedges require more effort to produce while typing than while speaking; for most people, typing an extra word or syllable is more costly than uttering one.

Finally, Example (11) shows how a three-person (face-to-face) group, M, S, and A, coordinate closely to recall the movie, with one completing the utterance another began and the third proposing a correction, mitigated with a tag question.

(11) M: He began telling the story
S: about when his father was young, when
A: Ireland was being ruled by the British
S: It still is, isn’t it?

It is not uncommon for people in conversation to complete one another’s utterances; spontaneous completions have been discussed in detail by Wilkes-Gibbs (1986).

**Self interruptions and joint visual attention**

Self-interruption has been proposed as a coordination device that can be used by a speaker to attract an addressee’s attention. Goodwin (1981) reports examples of speakers beginning to speak before capturing their addressees’ attention and then suspending the utterance until addressees turn their eye gaze to look at the speakers, at which point they backtrack and restart the utterance. In referential communication studies that have not prevented the director and matcher from seeing each other’s faces, eye gaze plays a special role; people tend to make eye contact at points of difficulty in the conversation (Boyle, Anderson, & Newlands, 1994).

Joint visual attention occurs not only by achieving eye contact, but also by achieving mutual gaze upon a particular object. As mentioned earlier, being able to gaze at the same object and being mutually aware of this is arguably more useful in a collaborative task than being able to see the partner’s face. Pointing, a form of referring, can be done with a finger, object, or cursor; it can also be done with the eyes. Young children can follow an adult’s line of regard to an object, and this contributes to word-learning (Baldwin, 1995). In referential communication experiments in which the participants have been able to see one another over an arrangement of objects, directors have had to wear sunglasses and carefully control head orientation so as not to signal which object they were referring to by where they were looking (Hanna, 2001; Metzing & Brennan, 2002). Currently we are investigating the extent to which and the time course by which eye gaze is used as a visual cue during reference resolution (Hanna & Brennan, 2002).

**Speech disfluencies**

Spontaneous speech is remarkably disfluent; on average, there appear to be about 6 disfluencies per 100 words (Fox Tree, 1995). Although disfluencies such as restarts, repetitions, and fillers (e.g., *um, uh, er*) are usually considered by psycholinguists to be personalities of an overburdened production system, some have argued that such interruptions of fluent speech may serve an interpersonal coordination function (Bortfeld, Leon, Bloom, Schober, & Brennan, 2001; Brennan & Schober, 2001; Clark & Wasow, 1998; Fox Tree & Clark, 1997). If a speaker takes a long time to produce an utterance, she risks losing her addressee’s attention or her speaking turn; but if she rushes to produce one that is defective, she risks being misunderstood (Clark & Brennan, 1991). So she may warn her addressee of an upcoming delay in producing a word or phrase by uttering a filler such as *um, uh* (or British *er*) (Smith & Clark, 1993; Clark, 1994). In this way a filler may serve to hold the conversational floor, unless the addressee can help with the
speaker's word-finding problem (in which case the filler may invite the addressee to complete the speaker's utterance).

One thing that is clear is that disfluency rates vary somewhat across different speech corpora or situations (Bortfeld et al., 2001; Oviatt, 1995; Shriberg, 1996). For instance, people talking on the telephone are more disfluent than those talking face-to-face, producing 8.83 to 5.50 disfluencies per 100 words (Oviatt, 1995). Using the grounding framework, we have argued that differences in disfluency rates in conversations conducted over different media may be influenced by the resources these media offer for coordination (Bortfeld et al., 2001). That is, when eye contact and other visual cues are available, there are likely to be alternative ways of signaling such things as the intention to continue speaking or difficulty with an utterance in progress, leading to lower rates of fillers in particular.

Indirect evidence that certain disfluencies may serve (or be somehow associated with) interpersonal coordination processes comes from their relative distributions in speech with and without interacting addressees. For instance, conversational speech is more disfluent than monologue speech; in a study by Oviatt (1995) there were more disfluencies in dialogues (5.50–8.83 per 100 words) than in monologues (3.60 per 100 words). That study did not consider distributions of fillers separately from distributions of other types of disfluencies.

We conducted a study of the rates and distributions of disfluencies in Schober and Carstensen's (2001) referential communication corpus (Bortfeld et al., 2001). This corpus crossed age (pairs of young, middle-aged, or older adults), task roles (each speaker alternated as director and matcher), topic difficulty (speakers described and matched abstract geometric figures as well as photographs of children), gender (each pair consisted of a male and a female speaker), and relationships between speakers (half the pairs were married to each other and half were strangers, married to someone else). Levels of education and socioeconomic status was balanced across gender. As for results, disfluency rates were higher in long utterances than in short ones, in directors' speech rather than in matchers', and in discussions of abstract figures rather than pictures of children. This confirms the general association of disfluencies with increased planning load. However, fillers do not seem to work like repetitions and restarts. When utterance length was controlled for, the director/matcher role no longer led to differences in repetition and restart rates, suggesting these depend mainly on planning load. However, director/matcher role still mattered for filler rates; directors produced higher filler rates than matchers in utterances matched for length. This supports Shriberg's (1996) suggestion that fillers (but not repetitions or restarts) are related to processes of interpersonal coordination. Since directors generally take more of the initiative in a referential communication task, they may use fillers to signal upcoming delays in producing utterances, to solicit help from matchers, or to hold on to the conversational floor if a matcher cannot help with a referring expression.

Consistent with earlier findings (e.g., Albert, 1980), older speakers in the Bortfeld et al. study produced slightly higher disfluency rates overall than young and middle-aged speakers. The interesting age-related finding concerns the positions of fillers within utterances. With increasing age of speakers, there were higher rates of fillers within constituents but no increase in rates of fillers between constituents (Bortfeld et al., 2001). This finding is consistent with the fact that older speakers have more word-finding problems than younger speakers (Obler & Albert, 1984), but that any coordination-related uses may occur regardless of age.

Cues for coordinating repairs during comprehension of disfluent speech

I have discussed opportunities for coordination
and metacommunication presented by cues in disfluent speech; next I will consider some of the consequences upon comprehension. According to Levelt (1989), disfluent speech presents a *continuation problem* for listeners; disfluencies must be edited out in order for utterances to make sense during comprehension. Although disfluencies are common in spontaneous speech, psycholinguistic theories tend to ignore them. Most computational approaches focus on parsing idealized utterances rather than naturally disfluent ones (notable exceptions being Core & Schubert, 1999; Hindle, 1983; Nakatani & Hirschberg, 1994; Shriberg, Bear, & Dowding, 1992). Speech production and speech processing are accomplished incrementally (see, e.g., Tanenhaus et al., 1995), and such incremental processing requires quite a different architecture and mechanism for ambiguity resolution and repair than one that begins processing only at the end of a complete utterance, or that breaks down when it encounters anything but grammatically well-formed speech. The idea is that speakers can mark problems with their utterances in ways that not only don’t slow their addressees down, but that enable listeners to recover from any misparsed or premature commitment to the wrong interpretation.

Consider the spontaneously disfluent utterance *Move to the yell- uh, orange square.* Here, the speaker halted midword and initiated a repair of the color word, marking the edit interval between the interruption point and the onset of the repair with a filler. The part of the utterance before the interruption point that needed to be excised (in this case, *yell-*) is known as the reparandum. Lickley and Bard (1996) found evidence that listeners are relatively deaf to words in a reparandum, and Shriberg and Lickley (1993) found that fillers such as *um* or *uh* may be produced with distinctive intonation that may help listeners distinguish (and perhaps filter) them from the part of the utterance that counts as the primary message or track 1, to use Clark’s (1996) terminology. Fox Tree (1995) found that while previous restarts in an utterance may slow monitoring for a subsequent word, repetitions don’t seem to hurt, and some fillers, such as *uh*, seem to speed monitoring.

It has been proposed that speakers halt speaking as soon as they detect a problem (Levelt, 1989; Nootboon, 1980). This proposal is supported by production data from Levelt’s (1983) corpus; 69% of the time, speakers interrupted themselves within or right after the problematic word. If the problem is detected after a troublesome word has already been uttered, the speaker is somewhat more likely to finish the current word (but not the current phrase) before stopping to repair (Levelt, 1983, 1989). In Levelt’s corpus study, when an interruption occurred midword, the problem was with the interrupted word itself (Levelt, 1983, 1989), which led to the proposal that an interrupted word signals what a speaker does not mean. Thirty percent of the repairs in Levelt’s corpus marked the editing interval with a filler, and the earlier an error was repaired, the more likely it was to be so marked.

From Levelt’s corpus-based evidence, Michael Schober and I developed and tested two hypotheses (not mutually exclusive) concerning the impact of disfluencies on utterance comprehension in reaction time studies (Brennan & Schober, 2001). The first is that mid-word interruptions (like *yell-orange*) are better signals than between-word interruptions (like *yellow-orange*) that a word has been produced in error and that the speaker intends to replace it. The second is that interruptions marked by *uh* (like *yell-uh-orange*) are better error signals than interruptions without *uh* (like *yell-orange*). We reasoned that if listeners can use the regularities in the forms of disfluencies to repair their interpretations, then they should be faster to comprehend and respond to color words in utterance where such cues were present than when they were absent.

---

1) However, it has been argued that when interruptions are followed immediately by repairs, the trouble has been detected before the interruption (Blackmer & Mitton, 1991).
We had a naive speaker issue commands to an addressee who carried them out; the commands concerned a display that sometimes changed as the speaker was speaking, causing him to produce both fluent and disfluent utterances. Some of these we then edited electronically so that we ended up with four different kinds of edited controls in addition to three different types of naturally occurring disfluencies, of the following sorts:

(12) a. Move to the orange square
   b. Move to the | orange square
   c. Move to the yellow- orange square
   d. Move to the ye- orange square
   e. Move to the ye- uh, orange square
   f. Move to the orange square
   g. Move to the ye- orange square
   h. Move to the uh, orange square

Utterance (12a) was naturally fluent, while (b) shortened it by having the disfluency electronically excised. Utterances (c), (d), and (e) were spontaneous (unedited) disfluencies, and (f), (g), and (h) were edited versions of (e) that replaced original material from the reparandum and/or editing interval with pauses of equal length in order to control for timing. In utterances (c)–(h), the reparandum began after the word the and continued until the interruption point (in (f) and (h) these points were the same, as the repairs were covert). The edit interval in (c)–(h) began at the interruption site, included silence or a filler, and ended at the start of the repair color word. This enabled us to test the relative contributions of mid-word interruptions, fillers, and the time that elapsed during these phonological elements of the utterance.

We played large numbers of fluent utterances along with smaller numbers of disfluent ones and edited controls to listeners who used keys to select colored geometric objects on a screen as quickly and accurately as possible. Response times were calculated relative to the onset of the target color word (in these examples, orange). As we predicted, mid-word interruptions and fillers did not disrupt comprehension. In our first two experiments, we found that the fastest responses, along with relatively low error rates, occurred for utterances like (e), above, that contained mid-word interruptions marked by fillers; the color word following these disfluencies was interpreted even faster than when it appeared in a fluent utterance such as (a). Next, we conducted two more experiments to determine whether this "disfluency advantage" was due to the actual phonological forms of the interrupted words and fillers, or else to the amount of time that elapsed in the edit interval, presumably preparing the listener to make a repair before the onset of the target color word. It turned out that both cues achieved their advantages due to their effects on the timing of the reparanda and the repairs. That is, listeners made fewer errors, the less incorrect information they heard (that is, the shorter the reparanda); mid-word interruptions such as (d) presented less misleading information than between-word interruptions such as (c).

And finally, the phonological form of the filler did not matter; with the timing precisely controlled, (e) was no faster than (g), and (h) was no faster than (f). Mid-word interruptions contributed to faster response times since they provided information about what the speaker didn't intend, with utterances of types (e) and (g) faster than (f) and (h). We concluded that, at least for this comprehension task, fillers served as useful cues due to their impact on utterance timing rather than their phonological form.

**Partner Effects in the Comprehension and Production of Utterances**

In this final section, I will briefly describe two of our studies that show how coordination in conversation can be partner-specific. No one denies that speakers engage in audience design, or adapting their utterances to their addressees. However, there has been some debate concerning how and why audience design occurs. The first issue concerns whether speakers make ad-
justments to addressees due to taking their perspectives, or else because what is easiest for addressees to interpret just happens to coincide with what is easiest for speakers to say. The second issue concerns whether partner adjustments (either by speakers to particular addressees or by addressees to particular speakers) occur due to representing and maintaining a mental model of the partner’s knowledge, including common ground, or else due to moment-by-moment feedback from the partner.

This first issue was raised originally by Brown and Dell (1987; Dell & Brown, 1991) for many things that speakers do in conversation, such as pronouncing unpredictable words more clearly than predictable words (Lieberman, 1963), shortening words that are given compared to words that are new (Bard et al., 2000; Fowler & Housum, 1987; Samuel & Troicki, 1998), preferring to use common, available words that happen to be easy for listeners to process, or tending to mention unusual information more often than information that is easily inferable by addressees. Brown and Dell (1987) argued that these choices do not actually represent adjustments to the needs of addressees, but simply reflect what is easiest for speakers. In an experiment, they had subjects retell stories that included target actions performed with either typical or atypical instruments. The stories were illustrated for the speakers by pictures that the addressees either could or could not see. The addressees were confederates who heard each story about 40 times. The speakers mentioned the atypical instruments more often and earlier (within the same clause as the target verb) than they did the typical instruments. However, whether the addressees could see the pictures was irrelevant to the form of instrument mention, leading Brown and Dell to conclude that speakers were not making adjustments to the knowledge needs of particular addressees. Similar conclusions have been drawn by Keysar and his colleagues (Horton & Keysar, 1996; Keysar et al., 2000; for debate, see also Keysar & Horton, 1998 and Polichak & Gerrig, 1998).

However, these studies all used confederates as conversational partners, which throws into question whether speakers can adjust to addressees’ needs when the addressees in question may not have actual needs (as with Brown & Dell’s confederates). We (Lockridge & Brennan, in press) used Brown and Dell’s paradigm in which speakers retold stories containing actions that involved typical or atypical instruments (e.g., Adolph stabbed the man with a knife vs. an icepick) to addressees who could either see a picture illustrating the action and instrument, or not. The main difference between our study and Brown and Dell’s was that our addressees were naive subjects to whom the stories were new. The addressees knew that they would receive a memory test about the stories later, and this helped motivate them to be engaged in the stories.

Unlike Brown and Dell, we found a reliable audience design effect: Speakers were most likely to mention instruments and to package them early in the stories (e.g., in the same clause as the action verb) when addressees lacked pictures and when the instruments were atypical, that is, not inferable from the story’s context. Upon first mention, instruments were most often marked as indefinite or new (i.e., He stabbed him with an icepick) when addressees could infer them from neither actions nor pictures. Instruments that were typical and inferable from the target action tended to receive initial mentioned as given and relatively late in the story, when they were mentioned at all (as in He wiped the blood off the knife). We concluded that speakers who have visual co-presence with addressees are able to take their needs into account early in the syntactic choices they make in utterances (Lockridge & Brennan, in press). It may be difficult for a confederate addressee to provide the same sort of feedback as a naive addressee; ordinarily, addressees are not informed again and again of the
same information by speaker after speaker. It is possible that either the partner’s feedback was simply uninformative or else informed the storytellers that the partner understood the story all too well. In the latter case, such feedback could be as subtle as an acknowledgment spoken just a bit too quickly (recall that in Brennan & Williams, 1995, we demonstrated that listeners are indeed sensitive to such cues). Moreover, the degree to which naive addressees are engaged by stories (as opposed to doing a distracting secondary task) affects not only the feedback they provide but also the details storytellers present (Bavelas, Coates, & Johnson, 2000; Pasupathi, Stallworth, & Murdoch, 1998).

Earlier, I introduced the idea of a conceptual pact, which is reached when two interlocutors reach a temporary, flexible agreement to view an object in a particular way. In previous studies, we found that speakers who have entrained on particular expressions together tended to continue to use these expressions even when they could have used simpler terms (Brennan & Clark, 1996). However, when such speakers changed partners, they were more likely to revert to the unadorned basic level terms than they did with original addressees. This amounted to clear evidence that speakers engage in audience design. However the previous studies did not establish whether lexical entrainment emerges due to one partner’s representation in memory of what the other partner knows — a representation that might be used from the start in producing or interpreting utterances — or whether it emerges due to feedback from partners about the acceptability of referring expressions. If audience design is mainly a reaction to a partner’s immediate feedback, speakers and addressees might not need much of a model of one another’s knowledge at all (for discussion, see Schober & Brennan, in press).

Next I will present evidence concerning whether the precedents addressees entrain upon with a particular speaker represent pacts that are truly partner-specific. A previous study by Barr & Keysar (2001) had naive addressees entrain on perspectives with (confederate) speakers who produced referring expressions according to a script. Once precedents for referring expressions were established between the partners, the addressees heard the expressions again, from the original speaker or else from a new speaker (prerecorded and presented through an earpiece) who happened to use the same referring expression that had been produced by the first speaker. Their prediction was that if entrainment is indeed partner-specific, addressees should be slower to gaze at and reach for the referent when the referring expression is delivered by the new speaker than by the old speaker. However, addressees were equally fast in gazing and reaching toward the objects, regardless of who the speaker was. Barr and Keysar concluded that speakers and addressees use precedents simply because these are available and active in memory, and not because they represent anything specific about a partner’s knowledge.

The problem with this interpretation is that in order for a partner-specific effect to have emerged in this study (that is, in order for an addressee to have processed the original, entrained-on expression more slowly when it was spoken by a new speaker than when it was spoken by the original speaker), the addressees would have to have represented in memory what the new speaker did not know. But this is not computationally feasible, as Polichak and Gerrig (1998) have pointed out; the information not known to (or else not in common ground with) another person is potentially boundless. Furthermore, it would not even be adaptive for a comprehension system to be wired to make these sorts of computations! Typically, conventions are shared by groups of people larger than two, many of whom have never interacted directly or in person. It is not surprising (nor should it slow comprehension for an addressee) when a stranger happens to produce the same expression that the addressee has heard.
in conversation with another speaker.

We tested whether conceptual precedents are partner-specific by having speakers actually break the pacts that had been established with addressees (Metzing & Brennan, 2002). In a referential communication task, two confederates acted in succession as directors, with the subject performing as matcher. The matchers were told that the purpose of the experiment was to study how they took direction from different people, and that they would be working with 2 different directors who were both research assistants in our laboratory. The matcher wore a lightweight head-mounted camera and eye-tracker and faced the director above an array of cubbyholes that contained small objects visible to both partners. The task was for the director to instruct the matcher about how to move objects around so that the array matched a target picture that only the director could see. The first director and the matcher entrained on referring expressions for eight sets of 12 objects, each containing one target object. The matcher’s looks to the target object were coded in detail as each set of objects was matched 4 times in succession. The first director interacted with the matcher in the first three trials, introducing referring expressions in the first trial in a natural fashion, with hedges and descriptions, and then re-using these in shortened form on the next two trials. Then, for the fourth trial, this director left the room and either returned, or a new director entered. For the critical object on the fourth trial, the original or new director used either the original expression (e.g., the shiny cylinder) or an equally descriptive new expression (the silver pipe) and we measured the speed of the matcher’s first look to the target object. The original and new expressions had been normed by having another group of subjects name each object in the context of each set and were produced equally often.

Consistent with Barr and Keysar’s (2001) findings, we found that matchers were equally fast to gaze at the object, whether the familiar expression was uttered by either the original speaker (with whom it had been entrained), or by the new speaker, or if the new expression was used by the new speaker. However, when the original speaker uttered a new expression, breaking the conceptual pact that he had established with the matcher, the matcher was not only slower to touch the target object after the onset of the referring expression, but her very first look to the target object also took significantly longer than when the very same new expression was uttered by the new partner (Metzing & Brennan, 2002). This suggests that the effect of a partner’s knowledge has impact very early in reference resolution, rather than as a late adjustment or repair, as Keysar and colleagues have claimed. Upon hearing the unexpected expression from the original director, matchers seemed to launch a rapid search over the entire display, as if scanning for a new object that had not yet been referred to. The existing conceptual pact interfered with mapping a new expression onto the object — but only with the original speaker, not for the new speaker. This is the first experiment to demonstrate that having a conceptual pact (or common ground) with a speaker can actually interfere with interpreting the speaker’s utterances that break the pact. It demonstrates that in these circumstances, addressees represent what their partners know (as opposed to what they don’t know).

Conclusions

I have presented some findings concerning how language use is coordinated between individuals. Clearly, holding a conversation represents quite a feat of coordination involving processes both simultaneous and successive: People listen while planning utterances, speak while monitoring addressees’ responses, and perform speaking turns with remarkably few overlaps or pauses in between. The product is remarkable as well; people manage to align their distinct mental states sufficiently to come to the belief that they have
understood one another well enough for the purposes at hand.

For those of us interested in advancing the science of linguistic communication, what needs to be explored further is how these psycholinguistic and interpersonal processes are integrated within the cognitive architecture of the individual. Many questions come to mind: To what extent do processes such as comprehension and production share the same resources and the same constraints upon accessing those resources? Does the speech production system monitor responses from an addressee in the same time frame and at the same granularity that it uses to determine whether the utterance being articulated matches the evolving speech plan? As for repair processes, both for repairs of comprehension (whether recovery from temporarily ambiguous sentences or recovery from speech disfluencies) and repairs of production (whether overt or covert), which parts of the system do they disrupt? For instance, from the speaker’s perspective, does message planning or monitoring addressee feedback come to a halt while an articulation error is repaired? From the addressee’s perspective, which sorts of comprehension or reception errors are hardest to recover from, and does the need to repair have any impact upon planning or executing the next response or utterance?

Of course, psycholinguists study those aspects of language production and comprehension about which they can achieve sufficient degrees of experimental control. Using unobtrusive devices such as head-mounted eye-trackers, it is possible to measure moment-by-moment processing in a conversational setting. However, getting enough control to look at the big picture — at cognitive processes operating simultaneously — is a challenge. Until we can place the psycholinguistic findings about individual processes into a more comprehensive architecture, we cannot be certain about how these results “scale up” into a working language system.

Acknowledgments

The research program summarized in this paper is based on data and experiments that have been reported in detail elsewhere. I am grateful to all of my collaborators for their contributions, as well as many colleagues for helpful discussion. I thank Joy Hanna for helpful comments on the paper. This material is based upon work supported by the National Science Foundation under Grants No.9402167, 9711974, 9980013, and 0082602. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation. Please address correspondence to Susan E. Brennan, Department of Psychology, State University of New York, Stony Brook, NY, 11794-2500 (susan.brennan@sunysb.edu).

References


ing, Memory, & Cognition, 12, 575–586.
Hanna, J. E. (2001). The effects of linguis-


(Received 18 Dec. 2001)
(Accepted 23 Jan. 2002)

**Susan E. Brennan**

Susan Brennan is Associate Professor of Psychology at the State University of New York at Stony Brook, with joint appointments in the Departments of Linguistics and Computer Science. She received her Ph.D. in Cognitive Psychology from Stanford University in 1990, advised by Herbert H. Clark. She received her M.S.V.S. from MIT’s Architecture Machine Group, where she worked on computer-generated caricatures and teleconferencing interfaces under the direction of Nicholas Negroponte. She also holds a B.A. in anthropology from Cornell University. She has done research in dialogue and human-computer interaction at Atari, Apple Computer, and Hewlett-Packard Labs, including five years on HP’s Natural Language Project. She serves as Associate Editor of Discourse Processes and has served on the editorial board on Computational Linguistics. She is currently using eye-tracking and behavioral techniques to study the comprehension and production of spontaneous speech and paralinguistic cues.