Processes that Shape Conversation and their Implications for Computational Linguistics

Susan E. Brennan
Department of Psychology
State University of New York
Stony Brook, NY, US 11794-2500
susan.brennan@sunysb.edu

Abstract

Experimental studies of interactive language use have shed light on the cognitive and interpersonal processes that shape conversation; corpora are the emergent products of these processes. I will survey studies that focus on under-modelled aspects of interactive language use, including the processing of spontaneous speech and disfluencies; metalinguistic displays such as hedges; interactive processes that affect choices of referring expressions; and how communication media shape conversations. The findings suggest some agendas for computational linguistics.

Introduction

Language is shaped not only by grammar, but also by the cognitive processing of speakers and addressees, and by the medium in which it is used. These forces have, until recently, received little attention, having been originally consigned to "performance" by Chomsky, and considered to be of secondary importance by many others. But as anyone who has listened to a tape of herself lecturing surely knows, spoken language is formally quite different from written language. And as those who have transcribed conversation are excruciatingly aware, interactive, spontaneous speech is especially messy and disfluent. This fact is rarely acknowledged by psychological theories of comprehension and production (although see Brennan & Schober, in press; Clark, 1994, 1997; Fox Tree, 1995). In fact, experimental psycholinguists still make up most of their materials, so that much of what we know about sentence processing is based on a sanitized, ideal form of language that no one actually speaks.

But the field of computational linguistics has taken an interesting turn: Linguists and computational linguists who formerly used made-up sentences are now using naturally- and experimentally-generated corpora on which to base and test their theories. One of the most exciting developments since the early 1990s has been the focus on corpus data. Organized efforts such as LDC and ELRA have assembled large and varied corpora of speech and text, making them widely available to researchers and creators of natural language and speech recognition systems. Finally, Internet usage has generated huge corpora of interactive spontaneous text or "visible conversations" that little resemble edited texts.

Of course, ethnographers and sociolinguists who practice conversation analysis (e.g., Sacks, Schegloff, & Jefferson, 1974; Goodwin, 1981) have known for a long time that spontaneous interaction is interesting in its own right, and that although conversation seems messy at first glance, it is actually orderly. Conversation analysts have demonstrated that speakers coordinate with each other such feats as achieving a joint focus of attention, producing closely timed turn exchanges, and finishing each another's utterances. These demonstrations have been compelling enough to inspire researchers from psychology, linguistics, computer science, and human-computer interaction to turn their attention to naturalistic language data.

But it is important to keep in mind that a corpus is, after all, only an artifact—a *product* that emerges from the *processes* that occur between and within speakers and addressees. Researchers who analyze the textual records of conversation are only overhearers, and there is ample evidence that overhearers experience a conversation quite differently from addressees and from side participants (Schober & Clark, 1989; Wilkes-Gibbs & Clark, 1992). With a corpus alone, there is no independent evidence of what people actually intend or understand at

different points in a conversation, or why they make the choices they do. Conversation experiments that provide partners with a task to do have much to offer, such as independent measures of communicative success as well as evidence of precisely when one partner is confused or has reached a hypothesis about the other's beliefs or intentions. Task-oriented corpora in combination with information about how they were generated are important for discourse studies.

We still don't know nearly enough about the cognitive and interpersonal processes that underlie spontaneous language use—how speaking and listening are coordinated *between* individuals as well as *within* the mind of someone who is switching speaking and listening roles in rapid succession. Hence, determining what information needs to be represented moment by moment in a dialog model, as well as how and when it should be updated and used, is still an open frontier. In this paper I start with an example and identify some distinctive features of spoken language interchanges. Then I describe several experiments aimed at understanding the processes that generate them. I conclude by proposing some desiderata for a dialog model.

Two people in search of a perspective

To begin, consider the following conversational interchange from a laboratory experiment on referential communication. A director and a matcher who could not see each another were trying to get identical sets of picture cards lined up in the same order.

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

(Stellmann & Brennan, 1993)

Several things are apparent from this exchange. First, it contains several disfluencies or interruptions in fluent speech. The director restarts her first turn twice and her second turn once. She delivers a description in a series of installments, with backchannels from the matcher to confirm them. She seasons her speech with fillers like *uh*, pauses occasionally, and displays her commitment (or lack thereof) to what she is saying with displays like *ah boy this one ah boy* and *I don't know what that is*. Even though she is the one who knows what the target picture is, it is the matcher who ends up proposing the description that they both end up ratifying: *like a monk praying or something*. Once the director has ratified this proposal, they have succeeded in establishing a *conceptual pact* (see Brennan & Clark, 1996). En route, both partners hedged their descriptions liberally, marking them as provisional, pending evidence of acceptance from the other. This example is typical; in fact, 24 pairs of partners who discussed this object ended up synthesizing nearly 24 *different* but mutually agreed-upon perspectives. Finally, the disfluencies, hedges, and turns would have been distributed quite differently if this conversation had been conducted over a different medium—through instant messaging, or if the partners had had visual contact. Next I will consider the proceses that underlie these aspects of interactive spoken communication.

1 Speech is disfluent, and disfluencies bear information

The implicit assumptions of psychological and computational theories that ignore disfluencies must be either that people aren't disfluent, or that disfluencies make processing more difficult, and so theories of fluent speech processing should be developed before the research agenda turns to disfluent speech processing. The first assumption is clearly false; disfluency rates in spontaneous speech are estimated by Fox Tree (1995) and by Bortfeld, Leon, Bloom, Schober, and Brennan (2000) to be about 6 disfluencies per 100 words, not including silent pauses. The rate is lower for speech to machines (Oviatt, 1995; Shriberg, 1996), due in part to utterance length; that is, disfluency rates are higher in longer utterances, where planning is more difficult, and utterances addressed to machines tend to be shorter than those addressed to people, often because dialogue interfaces are designed to take on more

initiative. The average speaker may believe, quite rightly, that machines are imperfect speech processors, and plan their utterances to machines more carefully. The good news is that speakers can adapt to machines; the bad news is that they do so by recruiting limited cognitive resources that could otherwise be focused on the task itself. As for the second assumption, if the goal is to eventually process unrestricted, natural human speech, then committing to an early and exclusive focus on processing fluent utterances is risky. In humans, speech production and speech processing are done incrementally, using contextual information from the earliest moments of processing (see, e.g., Tanenhaus et al. 1995). This sort of processing requires quite a different architecture and different mechanisms for ambiguity resolution than one that begins processing only at the end of a complete and well-formed utterance. Few approaches to parsing have tried to handle disfluent utterances (notable exceptions are Core & Schubert, 1999; Hindle, 1983; Nakatani & Hirschberg, 1994; Shriberg, Bear, & Dowding, 1992).

The few psycholinguistic experiments that have examined human processing of disfluent speech also throw into question the assumption that disfluent speech is harder to process than fluent speech. Lickley and Bard (1996) found evidence that listeners may be relatively deaf to the words in a reparandum (the part that would need to be excised in order for the utterance to be fluent), and Shriberg and Lickley (1993) found that fillers such as *um* or *uh* may be produced with a distinctive intonation that helps listeners distinguish them from the rest of the utterance. Fox Tree (1995) found that while previous restarts in an utterance may slow a listener's monitoring for a particular word, repetitions don't seem to hurt, and some fillers, such as *uh*, seem to actually speed monitoring for a subsequent word.

What information exists in disfluencies, and how might speakers use it? Speech production processes can be broken into three phases: a *message* or semantic process, a *formulation* process in which a syntactic frame is chosen and words are filled in, and an *articulation* process (Bock, 1986; Bock & Levelt, 1994; Levelt, 1989). Speakers monitor their speech both internally and externally; that is, they can make covert repairs at the point when an internal monitoring loop checks the output of the formulation phase before articulation begins, or overt repairs when a problem is discovered after the articulation phase via the speaker's external monitor—the point at which listeners also have access to the signal (Levelt, 1989). According to Nooteboom's (1980) Main Interruption Rule, speakers tend to halt speaking as soon as they detect a problem. Production data from Levelt's (1983) corpus supported this rule; speakers interrupted themselves within or right after a problem word 69% of the time.

How are regularities in disfluencies exploited by listeners? We have looked at the *comprehension* of simple fluent and disfluent instructions in a constrained situation where the listener had the opportunity to develop expectations about what the speaker would say (Brennan & Schober, in press). We tested two hypotheses drawn from some suggestions of Levelt's (1989): that "by interrupting a word, a speaker signals to the addressee that the word is an error," and that an editing expression like *er* or *uh* may "warn the addressee that the current message is to be replaced," as with *Move to the ye— uh, orange square.* We collected naturally fluent and disfluent utterances by having a speaker watch a display of objects; when one was highlighted he issued a command about it, like "move to the yellow square." Sometimes the highlight changed suddenly; this sometimes caused the speaker to produce disfluencies. We recorded enough tokens of simple disfluencies to compare the impact of three ways in which speakers interrupt themselves: immediately after a problem word, within a problem word, or within a problem word and with the filler *uh*.

We reasoned that if a disfluency indeed bears useful information, then we should be able to find a situation where a target word is faster to comprehend in a disfluent utterance than in a fluent one. Imagine a situation in which a listener expects a speaker to refer to one of two objects. If the speaker begins to name one and then stops and names the other, the way in which she interrupts the utterance might be an early clue as to her intentions. So the listener may be faster to recognize her intentions relative to a target word in a disfluent utterance than in an utterance in which disfluencies are absent. We compared the following types of utterances:

a. Move to the orange square *(naturally fluent)*
b. Move to the |orange square *(disfluency excised)*
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

Utterances c, d, and e were spontaneous disfluencies, and f, g, and h were edited versions that replaced the removed material with pauses of equal length to control for timing. In utterances c—h, the reparandum began after the word *the* and continued until the *interruption site* (after the unintended color word, color word fragment, or location where this information had been edited out). The *edit interval* in c—h began with the interruption site, included silence or a filler, and ended with the onset of the repair color word. Response times were calculated relative to the onset of the repair, *orange*.

The results were that listeners made fewer errors, the less incorrect information they heard in the reparandum (that is, the *shorter* the reparandum), and they were faster to respond to the target word when the edit interval before the repair was *longer*. They comprehended target words after mid-word interruptions *with* fillers faster than they did after mid-word interruptions *without* fillers (since a filler makes the edit interval longer), and faster than they did when the disfluency was replaced by a pause of equal length. This filler advantage did *not* occur at the expense of accuracy—unlike with disfluent utterances without fillers, listeners made no more errors on disfluent utterances with fillers than they did on fluent utterances. These findings highlight the importance of timing in speech recognition and utterance interpretation. The form and length of the reparandum and edit interval bear consequences for how quickly a disfluent utterance is processed as well as for whether the listener makes a commitment to an interpretation the speaker does not intend.

Listeners respond to pauses and fillers on other levels as well, such as to make inferences about speakers' alignment to their utterances. People coordinate both the content and the process of conversation; fillers, pauses, and self-speech can serve as displays by speakers that provide an account to listeners for difficulties or delays in speaking (Clark, 1994; Clark, 1997; Clark & Brennan, 1991). Speakers signal their *Feeling-of-Knowing* (FOK) when answering a question by the displays they put on right before the answer (or right before they respond with *I don't know*) (Brennan & Williams, 1995; Smith & Clark, 1993). In these experiments, longer latencies, especially ones that contained fillers, were associated with answers produced with a lower FOK and that turned out to be incorrect. Thus in the following example, A1 displayed a lower FOK than A2:

Q*: Who founded the American Red Cross?*
A1: *.....um..... Florence Nightingale?*
A2*: ..... Clara Barton.*

Likewise, non-answers (e.g., *I don't know*) after a filler or a long latency were produced by speakers who were more likely to recognize the correct answers later on a multiple choice test; those who produced a non-answer immediately did not know the answers. Not only do speakers display their difficulties and metalinguistic knowledge using such devices, but listeners can process this information to produce an accurate *Feeling-of-Another's-Knowing*, or estimate of the speaker's likelihood of knowing the correct answer (Brennan & Williams, 1995).

These programs of experiments hold implications for both the generation and interpretation of spoken utterances. A system could indicate its confidence in its message with silent pauses, fillers, and intonation, and users should be able to interpret this information accurately. If machine speech recognition were conducted in a fashion more like human speech recognition, timing would be a critical cue and incremental parses would be continually made and unmade. Although this approach would be computationally expensive, it might produce better results with spontaneous speech.

2 Referring expressions are provisional until ratified by addressees.

Consider again the exchange in Example (1). After some work, the director and matcher eventually settled on a mutual perspective. When they finished matching the set of 12 picture cards, the cards were shuffled and the task was repeated several more times. 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, who switched roles on each round, marked the fact that they had achieved a mutual perspective by reusing the same term, *monk*, in repeated references to the same object. These references tend to shorten over time. In Brennan and Clark (1996), we showed that once people coordinate a perspective on an object, they tend to continue to use the same terms that mark that shared perspective (e.g., *the man's pennyloafer*), even when they could use an even shorter 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 *conceptual entrainment* appears to be partner-specific—upon repeated referring to the same object but with a *new* partner, speakers were more likely to revert to the basic level term, due in part to the feedback they received from their partners (Brennan & Clark, 1996).

These examples depict the interpersonal processes that lead to conceptual entrainment. The director and matcher used many hedges in their initial proposals and counter-proposals (e.g., *it's almost like a person kind of in a weird way*, and *yeah like like a monk praying or something*). Hedges dropped out upon repeated referring. We have proposed (Brennan & Clark, 1996) that hedges are devices for signaling a speaker's commitment to the perspective she is proposing. Hedges serve social needs as well, by inviting counter-proposals from the addressee without risking loss of face due to overt disagreements (Brennan & Ohaeri, 1999).

It is worth noting that people's referring expressions converge not only with those of their human partners, but also with those of computer partners (Brennan, 1996; Ohaeri, 1995). In our text and spoken dialogue Wizard-of-Oz studies, when simulated computer partners used deliberately different terms than the ones people first presented to them, people tended to adopt the computers' terms, even though the computers had apparently "understood" the terms people had first produced (Brennan, 1996; Ohaeri, 1995).

The impetus toward conceptual entrainment marked by repeated referring expressions appears to be so compelling that native speakers of English will even produce non-idiomatic referring expressions (e.g., *the chair in which I shake my body*, referring to a rocking chair) in order to ratify a mutually-achieved perspective with non-native speakers (Bortfeld & Brennan, 1987).

Such findings hold many implications for utterance generation and the design of dialogue models. Spoken and text dialogue interfaces of the future should include resources for collaboration, including those for negotiating meanings, modeling context, recognizing which referring expressions are likely to index a particular conceptualization, keeping track of the referring expressions used by a partner so far, and reusing those expressions. This would help solve the "vocabulary problem" in human-computer interaction (Brennan, to appear).

3 Grounding varies with the medium

Grounding is the process by which people coordinate their conversational activities, establishing, for instance, that they understand one another well enough for current purposes. There are many activities to coordinate in conversation, each with its own cost, including:

- getting an addressee's attention in order to begin the conversation
- planning utterances the addressee is likely to understand
- producing utterances
- recognizing when the addressee does not understand
- initiating and managing repairs
- determining what inferences to make when there is a delay
- receiving utterances
- recognizing the intention behind an utterance
- displaying or acknowledging this understanding
- keeping track of what has been discussed so far (common ground due to linguistic co-presence)
- determining when to take a turn
- monitoring and furthering the main purposes or tasks at hand
- serving other important social needs, such as face-management

(adapted from Clark & Brennan, 1991)

Most of these activities are relatively easy to do when interaction is face-to-face. However, the affordances of different media affect the costs of coordinating these activities. The actual forms of speech and text corpora are shaped by how people balance and trade off these costs in the context of communication.

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). Pairs of people discussed many different locations on identical maps displayed on networked computer screens in adjoining cubicles. The task was for the matcher to get his car icon parked in the same spot as the car 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-Only Evidence, she could not. In both conditions, they could talk freely.

Language-action transcripts were produced for a randomly chosen 10% of 480 transcribed interchanges. During each trial, the x and y coordinates of the matcher's icon were recorded and time-stamped, as a moment-by-moment estimate of where the matcher thought the target location was. For the 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.

Sample time-distance plots are shown in Figures 1 and 2. Matchers' icons got closer to the target over time, but not at a steady rate. Typically, distance diminished relatively steeply early in the trial, while the matcher interpreted the director's initial description and rapidly moved his icon toward the target location. Many of the plots then showed a distinct elbow followed by a nearly horizontal region, meaning that the matcher then paused or moved away only slightly before returning to park his car icon. This suggests 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, or exchange sufficient evidence in order to establish mutual belief. The region after the elbow appears to correspond to the acceptance phase proposed by Clark & Schaefer (1989); the figures show that it was much shorter when directors had visual evidence than when they did not. The accompanying speech transcripts, when synchronized with the time-distance plots, showed that matchers gave verbal acknowledgements when directors did not have visual evidence and withheld them when directors did have visual evidence. Matchers made this adjustment to directors even though the information on the matchers' own screen was the same for both conditions, which alternated after every 10 locations for a total of 80 locations discussed by each pair.

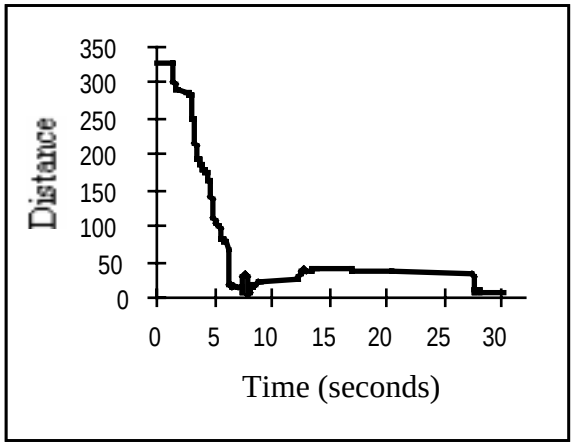

Figure 1: Time-Distance Plot of Matcher-Director Convergence, ***Without*** Visual Evidence of the Matcher's Progress

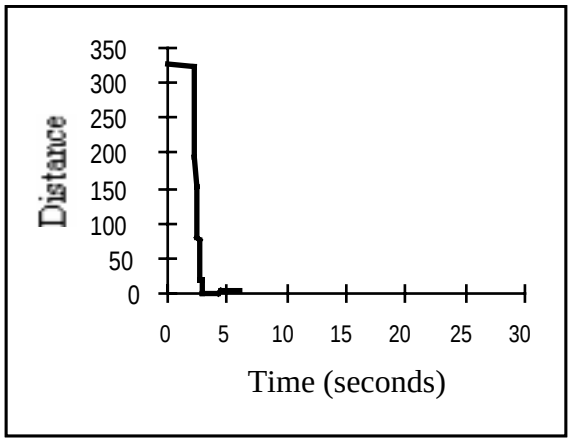

Figure 2: Time-Distance Plot of Matcher-Director Convergence, ***With*** Visual Evidence of the Matcher's Progress

These results document the grounding process and the time course of how directors' and matchers' hypotheses converge. The process is a flexible one; partners shift the responsibility to whomever can pay a particular cost most easily, expending the least collaborative effort (Clark & Wilkes-Gibbs, 1986).

In another study of how media affect conversation (Brennan & Ohaeri, 1999; Ohaeri, 1998) we looked at how grounding shapes conversation held face-to-face vs. via chat windows in which people sent text messages that appeared immediately on their partners' screens. Three-person groups had to reach a consensus account of a complex movie clip they had viewed together. We examined the costs of serving face-management needs (politeness) and

looked at devices that serve these needs by giving a partner options or seeking their input. The devices counted were hedges and questions.

Although both kinds of groups recalled the events equally well, they produced only half as many words typing as speaking. There were much lower rates of hedging (per 100 words) in the text conversations than face-to-face, but the same rates of questions. We explained these findings by appealing to the costs of grounding over different media: Hedging requires using additional words, and therefore is more costly in typed than spoken utterances. Questions, on the other hand, require only different intonation or punctuation, and so are equally easy, regardless of medium. The fact that people used just as many questions in both kinds of conversations suggests that people in electronic or remote groups don't cease to care about face-management needs, as some have suggested; it's just harder to meet these needs when the medium makes the primary task more difficult.

Desiderata for a Dialogue Model

Findings such as these hold a number of implications for both computational linguistics and human-computer interaction. First is a methodological point: corpus data and dialogue feature coding are particularly useful when they include systematic information about the tasks conversants were engaged in.

Second, there is a large body of evidence that people accomplish utterance production and interpretation incrementally, using information from all available sources in parallel. If computational language systems are ever to approach the power, error recovery ability, and flexibility of human language processing, then more research needs to be done using architectures that can support incremental processing. Architectures should *not* be based on assumptions that utterances are complete and well-formed, and that processing is modular.

A related issue is that timing is critically important in interactive systems. Many models of language processing focus on the propositional content of speech with little attention to "performance" or "surface" features such as timing. (Other non-propositional aspects such as intonation are important as well.)

Computational dialogue systems (both text and spoken) should include resources for collaboration. When a new referring expression is introduced, it could be marked as provisional. Fillers can be used to display trouble, and hedges, to invite input. Dialogue models should track the forms of referring expressions used in a discourse so far, enabling agents to use the same terms consistently to refer to the same things.

Because communication media shape conversations and their emergent corpora, minor differences in features of a dialogue interface can have major impact on the form of the language that is generated, as well as on coordination costs that language users pay.

Finally, dialogue models should keep a structured record of jointly achieved contributions that is updated and revised incrementally. No agent is omniscient; a dialogue model represents only one agent's estimate of the common ground so far (see Cahn & Brennan, 1999). There are many open and interesting questions about how to best structure the contributions from interacting partners into a dialogue model, as well as how such a model can be used to support incremental processes of generation, interpretation, and repair.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grants No. IRI9402167, IRI9711974, and IRI9980013. I thank Michael Schober for helpful comments.

References

Bock, J. K. (1986). Meaning, sound, and syntax: Lexical priming in sentence production. *J. of Experimental Psychology: Learning, Memory, & Cognition, 12*, 575-586.

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

Bortfeld, H., & Brennan, S. E. (1997). Use and acquisition of idiomatic expressions in referring by native and non-native speakers. *Discourse Processes, 23*, 119-147.

Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., & Brennan, S. E. (2000). Disfluency rates in spontaneous speech: Effects of age, relationship, topic, role, and gender. Manuscript under review.

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

Brennan, S. E. (1996). Lexical entrainment in spontaneous dialog. *Proc. 1996 International*

Symposium on Spoken Dialogue (ISSD-96) (pp. 41-44). Acoustical Society of Japan: Phila., PA.

Brennan, S. E. (to appear). The vocabulary problem in spoken dialog systems. In S. Luperfoy (Ed.), *Automated Spoken Dialog Systems*, Cambridge, MA: MIT Press.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *J. of Experimental Psychology: Learning, Memory, & Cognition, 22*, 1482-1493.

Brennan, S. E., & Ohaeri, J. O. (1999). Why do electronic conversations seem less polite? The costs and benefits of hedging. *Proc. Int. Joint Conference on Work Activities, Coordination, and Collaboration (WACC '99)* (pp. 227-235). San Francisco, CA: ACM.

Brennan, S. E., & Schober, M. F. (in press). How listeners compensate for disfluencies in spontaneous speech. J. of Memory & Language.

Brennan, S. E., & Williams, M. (1995). The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *J. of Memory & Language, 34*, 383-398.

Cahn, J. E., & Brennan, S. E. (1999). A psychological model of grounding and repair in dialog. *Proc. AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems* (pp. 25-33). North Falmouth, MA: AAAI.

Clark, H.H. (1994). Managing problems in speaking. *Speech Communication, 15*, 243-250.

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

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127-149).

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

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

Core, M. G., & Schubert, L. K. (1999). A model of speech repairs and other disruptions. *Proc. AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems.* North Falmouth, MA: AAAI.

Fox Tree, J.E. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *J. of Memory & Language, 34*, 709-738.

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

Hindle, D. (1983). Deterministic parsing of syntactic non-fluencies. In *Proc. of the 21st Annual Meeting, Association for Computational Linguistics*, Cambridge, MA, pp. 123-128.

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

Levelt, W. (1989). *Speaking: From intention to articulation.* Cambridge, MA: MIT Press.

Lickley, R., & Bard, E. (1996). On not recognizing disfluencies in dialog. *Proc. International Conference on Spoken Language Processing (ICSLIP '96)*, Philadelphia, 1876-1879.

Nakatani, C. H., & Hirschberg, J. (1994). A corpus-based study of repair cues in spontaneous speech. *J of the Acoustical Society of America, 95*, 1603-1616.

Nooteboom, S. G. (1980). Speaking and unspeaking: Detection and correction of phonological and lexical errors in spontaneous speech. In V. A. Fromkin (Ed.), *Errors in linguistic performance.* New York: Academic Press.

Ohaeri, J. O. (1995). Lexical convergence with human and computer partners: Same cognitive process? *Unpub. Master's thesis*. SUNY, Stony Brook, NY.

Ohaeri, J. O. (1998). Group processes and the collaborative remembering of stories. *Unpublished doctoral dissertation*. SUNY, Stony Brook, NY.

Oviatt, S. (1995). Predicting spoken disfluencies during human-computer interaction. *Computer Speech and Language, 9*, 19-35.

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

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

Shriberg, E. (1996). Disfluencies in Switchboard. *Proceedings, International Conference on Spoken Language Processing*, Vol. Addendum, 11-14. Philadelphia, PA, 3-6 October.

Shriberg, E., Bear, J., & Dowding, J. (1992). Automatic detection and correction of repairs in human-computer dialog. In M. Marcus (Ed.), *Proc DARPA Speech and Natural Language Workshop* (pp. 419-424). Morgan Kaufmann.

Shriberg, E.E. & Lickley, R.J. (1993). Intonation of clause-internal filled pauses. *Phonetica*, 50, 172-179.

Smith, V., & Clark, H. H. (1993). On the course of answering questions. *J. of Memory and Language, 32*, 25-38.

Stellmann, P., & Brennan, S. E. (1993). Flexible perspective-setting in conversation. *Abstracts of the Psychonomic Society, 34th Annual Meeting* (p. 20), Washington, DC.

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

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