

# *Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender\**

HEATHER BORTFELD,<sup>1</sup> SILVIA D. LEON,<sup>2</sup>  
JONATHAN E. BLOOM,<sup>2</sup> MICHAEL F. SCHOBER,<sup>2</sup>  
SUSAN E. BRENNAN<sup>3</sup>

<sup>1</sup> *Brown University*

<sup>2</sup> *New School for Social Research*

<sup>3</sup> *State University of New York at Stony Brook*

## KEY WORDS

*communication*

*conversation*

*disfluency*

*speech planning*

*spontaneous  
speech*

## ABSTRACT

After reviewing situational and demographic factors that have been argued to affect speakers' disfluency rates, we examined disfluency rates in a corpus of task-oriented conversations (Schober & Carstensen, 2001) with variables that might affect fluency rates. These factors included: speakers' ages (young, middle-aged, and older), task roles (director vs. matcher in a referential communication task), difficulty of topic domain (abstract geometric figures vs. photographs of children), relationships between speakers (married vs. strangers), and gender (each pair consisted of a man and a woman). Older speakers produced only slightly higher disfluency rates than young and middle-aged speakers. Overall, disfluency rates were higher both when speakers acted as directors and when they discussed abstract figures, confirming that disfluencies are associated with an increase in planning difficulty. However, fillers (such as *uh*) were distributed somewhat differently than repeats or restarts, supporting the idea that fillers may be a resource for or a consequence of interpersonal coordination.

## 1. INTRODUCTION

Spontaneous conversation is notoriously disfluent.<sup>1</sup> Consider this snippet of conversational speech reproduced from a corpus collected during a laboratory study of task-oriented

---

\* *Acknowledgements:* This material is based upon work supported by the National Science Foundation under Grants No. IRI-9402167, IRI-9711974 and SBR-9730140. We thank Maurice Williams for assistance with coding and Ellen Bard, Robin Lickley, and an anonymous reviewer for many good suggestions.

*Address for correspondence:* Susan E. Brennan, Department of Psychology, State University of New York, Stony Brook, NY, 11794–2500; e-mail <susan.brennan@sunysb.edu>.

<sup>1</sup> Fox Tree (1995) as well as Kasl and Mahl (1965) estimate about 6% of words in spontaneous speech are disfluent (this estimate does not include silent pauses, which can be difficult to identify consistently). While estimates of disfluency rates vary, they are high enough to have a significant impact on speech processing.

conversation (see Brennan & Clark, 1996). Two people, Matt and Anne, are discussing a series of photos. Matt is helping Anne arrange them in a particular order, and he tries to direct her attention to a particular one:

Matt: then we have the dog... ah the larger dog the larger picture of a dog

Anne: okay

It appears that Anne has no trouble understanding what Matt means with his simple utterance (in fact, she immediately picked up the picture that he was describing). But note what follows his initial sentence *then we have the dog*. After a brief pause, Matt emits a sound, *ah*, and then repeats his last noun phrase twice, modifying it each time. The resulting utterance is, strictly speaking, not fluent, if fluent speech is defined as an uninterrupted sequence of words that follows the rules of English syntax. But Anne, in comprehending what Matt said, may not even have noticed anything amiss.

Although disfluencies such as these may not thwart speech comprehension, they are interesting for several reasons. First, they pose a problem for most theories of parsing, which are designed to handle only grammatical or “well-formed” utterances (see Fox Tree, 1995, for a review). Second, by demonstrating how speech planning and articulation break down, departures from fluent and grammatical speech provide useful data about the architecture of the speech production system and the constraints upon it (Dell, 1986; Fromkin, 1971, 1973, 1980; Garrett, 1975; Levelt, 1989; MacKay, 1970, 1972, 1973; Shattuck-Hufnagel, 1979, 1982; Shattuck-Hufnagel & Klatt, 1979). Third, in certain circumstances, disfluencies can display metalinguistic information to listeners about a speaker’s confidence (Brennan & Williams, 1995), inform listeners about a speaker’s planning difficulties (Brennan & Schober, 2001; Schachter, Christenfeld, Ravina, & Bilous, 1991), or, possibly, serve as devices for coordinating conversational interaction (Brennan & Kipp, 1996; Maclay & Osgood, 1959; Shriberg, 1996; Wilkes-Gibbs, 1986). Last but not least, spontaneous human speech contains disfluencies that pose problems for speech recognition systems (Butzberger, Murveit, Shriberg, & Price, 1992; Hindle, 1983; Nakatani & Hirschberg, 1994; Oviatt, 1995; Shriberg, 1996; Shriberg, Bear, & Dowding, 1992; Shriberg, Wade, & Price, 1992).

In this paper we investigate various situational and demographic factors that have been argued to affect speakers’ disfluency rates. Rather than comparing disfluency rates across different corpora, where differences in rates might reflect differences in the circumstances of data collection or in coding criteria, we examine disfluency rates within one large corpus of conversations (Schober & Carstensen, 2001). In this corpus, the conversations were collected under controlled conditions with systematic manipulation (fortuitously) of various factors that might affect disfluency—speakers’ age, gender, familiarity with the partner, conversational role, and topic. In order to avoid the possibility of attributing intentions to speakers inaccurately, disfluencies are coded only according to surface forms (in contrast to Carletta, Caley, & Isard, 1993). Finally, in contrast to studies that have aggregated different types of disfluencies (e. g., Oviatt, 1995), we consider the idea that different types of disfluencies may be associated with different processes, such as cognitive planning load and coordination of communication.

## 1.1 What makes speech disfluent?

### 1.1.1 Processing load

Since the 1950s, speech errors and disfluencies produced by normal speakers have been studied as a window into the intermediate linguistic products and cognitive processes of speech planning (Bock, 1986; Bock & Levelt, 1994; Dell, 1986; Fromkin, 1971, 1973, 1980; Garrett, 1975; Goldman-Eisler, 1958; Levelt, 1989; MacKay, 1970, 1972, 1973; Maclay & Osgood, 1959; Nooteboom, 1969; Shattuck-Hufnagel, 1979, 1982; Shattuck-Hufnagel & Klatt, 1979); these studies provide systematic evidence of how conceptual, syntactic, and articulatory processes cope with increased processing load. Recently, additional evidence associating disfluencies with increased processing load has turned up in descriptive studies of speech corpora. In Oviatt's (1995) study of disfluencies in six types of task-oriented conversations, long utterances had higher disfluency rates than short ones. This finding is supported by Shriberg's (1996) study of disfluencies in three different task-oriented conversational corpora, in which she found that the longer the sentence, the less likely it is to be fluent. The association of disfluencies with planning load is consistent with findings that disfluencies are more likely near the beginnings of turns or sentences, where planning effort is presumably higher (where Boomer, 1965, found more fillers and silent pauses, and where Shriberg found more disfluencies as well).

The topic or domain of a conversation is another characteristic that may cause the planning load of utterances to vary. In one study, social science lectures contained more disfluencies of one sort—fillers—than hard science lectures, and humanities lectures contained the most of all (Schachter, Christenfeld, Ravina, & Bilous, 1991). These findings were not due to individual differences between speakers; rates for individuals did not differ when they all addressed the same topic. Schachter and colleagues (Schachter et al., 1991; Schachter, Rauscher, Christenfeld, & Crone, 1994) suggested that speakers use more fillers when they must choose from a larger range of expressive options.

### 1.1.2 Coordination functions

Another possibility for why disfluencies are more common at the beginnings of utterances is that they might be used to coordinate interaction (Shriberg, 1996). In particular, Shriberg found that filler rates were not correlated with sentence length, and so may not be related to planning load in the way that other disfluencies are. If disfluencies such as fillers serve a communicative function, they may provide information that enables two people in conversation to better coordinate interaction, manage turn-taking, or align their mental states (Brennan & Schober, 2001; Brennan & Williams, 1995; Clark, 1994; Clark & Wasow, 1996; Fox Tree, 1995; Fox Tree & Clark, 1997; Levelt, 1989; Shriberg, 1996; Smith & Clark, 1993). For instance, time is a resource that people manage jointly in conversations, and managing resources involves making tradeoffs. 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 a delay in producing a word or phrase by uttering a filler such as *um*, *uh* (or British *er*), and *uh* (Clark, 1994; Smith & Clark, 1993). There is evidence that fillers can perform this sort of function: Speakers answering general knowledge

questions display accurate information about their mental search processes (Brennan & Williams, 1995; Smith & Clark, 1993); that is, they pause longer and use more fillers before producing an answer that they lack confidence in (and that is more likely to be incorrect) than before one that they have a strong feeling of knowing (and that is more likely to be correct). And speakers pause longer and use more fillers before a *non*-answer (e.g., *I don't know*), when they actually *do* know the answer but are just unable to retrieve it. This metacognitive display can be used by listeners to judge how likely the speaker is to know the correct answer (Brennan & Williams, 1995), as well as by survey interviewers to judge when respondents are likely to be misinterpreting a question (Bloom & Schober, 1999).

A filler may also help a listener realize that a speaker has just misspoken. In one series of studies, listeners were faster and more accurate in comprehending repair words such as *orange* in utterances like *Move to the yel- uh, orange square* when the interrupted word was followed by *uh* than when it was not (Brennan & Schober, 2001). This disfluency advantage appears to be due to the additional time that elapses while the filler is being uttered. Comprehension was also faster with the disfluency (e.g., *yel- uh*) than when it was replaced with an unaccounted-for silent pause of equal length. In these cases, disfluencies may serve as displays by speakers that warn listeners of or account for delays and problems in speaking (Brennan & Schober, 2001; see also Clark, 1994).

There is at least one other way in which fillers may be useful in conversation; fillers may help people manage turn-taking. The hypothesis that fillers act as turn-keeping cues (blocking listeners from interrupting the speaker with a new speaking turn) was proposed by Maclay and Osgood (1959) and has been entertained by others (Ball, 1975; Carletta, Caley, & Isard, 1995; Rochester, 1973), although the evidence in support of this hypothesis has been questioned (Cook & Lalljee, 1970; Lalljee & Cook, 1969). The story of how fillers affect turn-taking may be more complicated than one where an *uh* simply helps a speaker keep an addressee from interrupting. In Wilkes-Gibbs' corpus of conversational completions (where one speaker spontaneously completed another's utterance), it appeared that fillers were sometimes interpreted as displays of trouble and requests for help, as in this example (see Clark & Wilkes-Gibbs, 1986; Wilkes-Gibbs, 1986):

A: and number 12 is, uh, ...

B: chair.

A: with the chair, right.

Here B may have taken *uh* to be a request by A for help in producing the right word; if this is so, then the disfluency was used as a collaborative tool. If fillers warn addressees that the speaker is still working on the utterance, then this may result in the addressee chiming in (if he can help with the speaker's problem), and otherwise waiting for the speaker to continue (if he cannot).

The idea that fillers may serve (at least in part) as a resource for interpersonal coordination is not incompatible with Schachter et al.'s (1991) finding of higher filler rates in domains with more indeterminacy. That is, when choosing words is more difficult, a speaker's need to account to her audience for any delays is presumably greater. This idea is also consistent with Kasl and Mahl's (unexpected) finding of a 41% increase in fillers (but not other kinds of disfluencies) in audio-only conversations between people in different rooms, compared to conversations in the same room with visual contact (Kasl & Mahl,

1965).<sup>2</sup> Consistent with this difference across communication media, Oviatt (1995) found that people talking on the telephone produced more disfluencies than those talking face-to-face, 8.83 to 5.50 disfluencies per 100 words (although she did not present filler rates separately from total disfluency rates, which also included corrections, false starts, and repeats). Differences in disfluency rates in conversations conducted over different media may, then, be influenced by the resources these media offer for coordination. That is, when eye contact and other visual cues are available, there may be redundant ways of signaling such things as the intention to continue speaking, difficulty with an utterance in progress, or other metacognitive information, leading to lower rates of fillers. This possibility is supported by the finding that filler rates are lower in speech produced while gesturing than in speech produced while not gesturing (Christenfeld, Schachter, & Bilous, 1991).

Additional evidence that certain disfluencies are associated with coordination between speakers and listeners is gleaned from their relative distributions in speech with and without interactive partners. For instance, conversational speech is more disfluent than monolog speech; in Oviatt's (1995) study there were more disfluencies in dialogs (5.50–8.83 disfluencies per 100 words) than in monologs (3.60 per 100 words). These findings are consistent with an older experiment in which dialogs contained higher disfluency rates than monologs of various types (Broen & Siegel, 1972); since people in that study rated being fluent as more important in their monologs than in their conversations (they were led to believe their conversations were not part of the experiment), Broen and Siegel (1972) proposed that the heightened fluency in monologs resulted from more careful monitoring.

It is particularly interesting that disfluency rates also differ in speech to human versus machine partners. In Oviatt's (1995) corpora, disfluency rates in speech intended for a human partner were much higher (5.50–8.83 per 100 words) than in speech intended for a machine speech recognizer (only .78–1.87 per 100 words). The disfluencies counted were fillers, corrections to content or grammar, and verbatim repetitions of words or phrases (separate distributions were not provided for the different types of disfluencies). Likewise, in her studies of speech to human and machine partners, Shriberg (1996) found that rates of repeated and deleted words, and especially, fillers, were higher in speech to human partners than to speech recognizers. She suggested that these types of disfluencies may help people coordinate conversational exchanges (Shriberg, 1996).

### 1.1.3 *Familiar versus unfamiliar conversational partners*

The Schober and Carstensen (2001) corpus allows us to examine whether people are more or less disfluent when talking to strangers than when talking with their spouses. The predictions to be made are unclear. On one hand, we might expect people to be more disfluent with strangers than with intimates, because they might be more anxious with unfamiliar partners; higher disfluency rates have been associated with anxiety (at least for disfluencies other than fillers, Mahl, 1987). On the other hand, to the extent that disfluencies are coordinating devices, we might expect people to be more disfluent with intimates;

---

<sup>2</sup> Kasl and Mahl (1965) speculated that in the absence of visual cues about a listener's understanding, speakers experienced more uncertainty, and fillers enabled them to gain more planning time (p. 432).

intimates might be more likely to display their planning problems to each other and rely on each other for help, whereas strangers might plan what they say to each other more carefully. The existing evidence is scant, although studies comparing speech with computers to speech with humans may be relevant (computer partners may be more like strangers). Shriberg's and Oviatt's studies have shown that there are *fewer* disfluencies with computers than with people; if this can be extended to strangers, we should expect fewer disfluencies with strangers than with intimates.

#### 1.1.4 Age

Age-related changes in cognitive, motor, and perceptual functioning may affect speech in several relevant ways. Older speakers have more difficulty retrieving words than do younger speakers (Burke, MacKay, Worthley, & Wade, 1991; Rastle & Burke, 1996), although the ability to define words remains intact and may even improve with age (for reviews, see Obler & Albert, 1984; Sandson, Obler, & Albert, 1987). In a study using a picture description task, speakers in their fifties, sixties, and seventies were less efficient than speakers in their forties, in that the older speakers' messages contained less content per unit time (Shewan & Henderson, 1988). And speakers over 50 appear to use more elaborate syntactic forms than younger speakers (Obler & Albert, 1984). Such age-related changes seem likely to make conversation more effortful and to generate higher disfluency rates. Indeed, some studies (Albert, 1980; Schow, Christensen, Hutchinson, & Nerbonne, 1978) have found higher disfluency rates (including repetitions, restarts, and fillers) among older speakers. On the other hand, Shewan and Henderson (1988) found no reliable age differences in disfluency rates (but the only disfluencies they tallied were repetitions). The Schober and Carstensen corpus allows us to examine how age affects disfluency rates over a fairly wide range of ages.

#### 1.1.5 Gender

In Shriberg's (1996) study, men produced more fillers than women did, but the sexes were equal with respect to other types of disfluency rates. Shriberg cautiously suggested that using more fillers may be a way for men to try to hold on to the conversational floor, but pointed out that in her corpora, gender was confounded with occupation and education level. In the Schober and Carstensen corpus, socioeconomic status was balanced across gender, and so we can discover whether Shriberg's observation is corroborated.

#### 1.1.6 Effects of these variables upon disfluencies

It is likely that the mapping of factors like cognitive load, addressee characteristics or relationship, communication medium, or speaker characteristics (such as state of arousal, age, or gender) onto disfluency rates is not a simple one. Although these factors may operate independently to produce disfluent speech, they may also work in concert. For instance, Shriberg found that repetitions and deletions increase with sentence length much more sharply in conversations with human partners than in conversations with machine partners (Shriberg, 1996). This could mean that cognitive load is of less consequence with machine partners than with other people; or else it could be due to disfluency rates with machines being quite low to begin with. As Oviatt (1995) points out, human-machine interaction is often characterized by a more structured (and less cognitively demanding) interactive format (such as when the computer takes the initiative and poses question after question to a human

user); in this situation, utterances tend to be shorter. In Oviatt's study, the lower disfluency rates with computers were not due solely to shorter utterances; structured formats led to lower disfluency rates even when utterance length was controlled for. An additional factor is that many spoken language interfaces require a button press to initiate spoken input, which may encourage speakers to plan their utterances off-line.

This study is motivated by the expectation that cognitive, social, and situational factors may interact to affect speech production. Disfluencies may arise from quite different processes or within quite different situations. As we proposed earlier, perhaps some disfluencies serve an interpersonal coordination function, such as displaying a speaker's intentional or metacognitive state to a partner, while others simply represent casualties of an overworked production system. One thing that is certain is that disfluency rates vary across different speech corpora (Oviatt, 1995; Shriberg, 1996). But determining the factors that influence disfluencies is difficult when the corpora being compared have been collected under very different conditions, such as with different tasks and different samples of speakers. Another problem with comparing the results of multiple studies is that there are many different ways in which disfluencies can be counted or categorized. Finally, the evidence that fillers may arise from different processes than other types of disfluencies warrants looking at them separately.

Here, we look at distributions of disfluencies using a corpus of spontaneous task-oriented, two-person conversations all collected under the same controlled conditions and balanced for variables that may influence disfluency rates for the reasons we have discussed. The corpus was audiotaped during a laboratory study of referential communication at Stanford University (Schober & Carstensen, 2001). Speaker age, relationship to addressee, and topic under discussion were all systematically varied. The sex of the speaker was also recorded. The design of this study enables us to make direct comparisons of disfluency rates across conditions with an eye towards causality and with greater validity than if we tried to study these variables by comparing across corpora.

## 2. METHOD

### 2.1 Corpus and Design

#### 2.1.1 Speakers

The corpus contained approximately 192,000 words uttered by 48 pairs of people in conversation. Among these pairs, 16 pairs were young (mean age, 28;10), 16 pairs were middle-aged (mean age, 47;11), and 16 pairs were older (mean age, 67;2). The 48 pairs of speakers comprised 24 pairs of male and female strangers and 24 married couples, divided equally by age (8 pairs each of young, middle-aged, and older speakers). They were recruited through the Stanford Alumni Association to participate for pay in a referential communication study (reported in Schober & Carstensen, 2001). Each participant was married and college educated, none had significant hearing loss, and the three age groups were no different in years of postsecondary education. The young pairs had been married for a mean of approximately 3 years, 9 months, the middle-aged pairs, 25 years, 3 months, and the older pairs, 42 years, 6 months.

### 2.1.2 Referential communication task

Each pair discussed objects from a familiar domain (photographs of children) and an unfamiliar domain (black and white abstract geometric forms known as tangrams). There were two sets of picture cards; one depicted 12 children and the other, 12 tangrams. Matching a set of picture cards constituted one trial. Each pair completed four trials, matching each of the two sets of 12 pictures a total of two times. With each trial, members alternated in the roles of director and matcher, so that each speaker served as director in a tangram picture trial once and in a child picture trial once, and as matcher for the other two trials. During a trial, each member of a pair had an identical copy of the set of picture cards. The task was for the matcher to get all 12 picture cards lined up in the same order as the director's cards. Since directors in a referential communication task know the correct target positions of the cards, they usually take more initiative in the task than do matchers. Members of a pair were visually separated but could communicate freely. Half of the time, they matched pictures of children first, and half of the time, tangrams. Half of the time, females performed as the first director, and half of the time, males did.

### 2.1.3 Design

In sum, the experiment had a mixed factorial design, with three between-subjects variables: the relationship between the speakers (married vs. strangers), speaker's gender, and age (young vs. middle-aged vs. older). The within-subjects variables were domain familiarity (children vs. tangrams) and role (matcher vs. director). This counterbalanced design is summarized in Table 1. For one pair (speakers #37 and #38) data from one out of the four trials were lost due to tape recorder malfunction, so the degrees of freedom in the ANOVAS represent the contributions of 94 speakers.

## 2.2 Transcription and coding

### 2.2.1 Transcription

All conversations had been audiotaped in stereo. Each conversation was first transcribed by one of two conversation researchers to a level of detail that captured all words and word fragments audible to the ear, as well as overlapping speech. Onsets and offsets of overlapping speech were marked to the level of which words overlapped. Also transcribed were nonlexical fillers (such as *uh*) and other vocalizations (such as laughter). Transcriptions were made using a foot-pedal-controlled tape player that afforded automatic rewinding for repeated playing at slow and normal speeds. Since our goal was to study lexical repetitions, syntactic restarts, and fillers, each transcriber made multiple passes over the tapes in order to transcribe utterances to this level of detail.<sup>3</sup> After each conversation was transcribed, the other conversation researcher checked the transcript while listening to the tape. Areas of overlap and disfluency were audited multiple times as necessary. Any stretches of speech with discrepant

---

<sup>3</sup> Note that disfluencies can also be considered at a finer grain, with cues that may include aberrations in pronunciation, errors in stress, and variations in intonation (see, e.g., Lickley, 1996; Lickley & Bard, 1996, 1998; Shriberg & Lickley, 1993). These levels of analysis are outside the scope of our project.



TABLE 1

Experimental design for the corpus of conversations (see Schober &amp; Carstensen, 2000)

	<i>Domain</i>			
	<i>Tangrams</i>		<i>Children</i>	
	<i>Director</i>	<i>Matcher</i>	<i>Director</i>	<i>Matcher</i>
<b>Young</b>				
Strangers (AB pairs)	A	B	A	B
	B	A	B	A
Married (CD pairs)	C	D	C	D
	D	C	D	C
<b>Middle-aged</b>				
Strangers	E	F	E	F
	F	E	F	E
Married	G	H	G	H
	H	G	H	G
<b>Older</b>				
Strangers	I	J	I	J
	J	I	J	I
Married	K	L	K	L
	L	K	L	K

*Note:* A, C, E, G, I, and K each represent eight different women and B, D, F, H, J, and L each represent eight different men. CD, GH, and KL pairs are married to each other. AB, EF, and IJ pairs are strangers to each other, but each of these participants is married to a member of a different pair. Strangers attended the experiment as quartets; that is, while a woman in cell A was paired with a man from cell B, her husband (also in B), was paired in another room with her partner's wife (also in A).

tape and transcript were reaudited and discussed by both researchers until agreement was reached and transcripts achieved an appropriate level of accuracy.

### 2.2.2 Coding and categorization of disfluencies

The text transcripts were then imported into *Sequence*, a Macintosh application for segmenting, coding, and analyzing types, numbers, and sequences of behavioral events (Dijkstra, 1999). The transcribed speech was coded as disfluent if it contained any of the following categories of events: Repeats (*just on the left left side*), restarts (e.g., *imme- just below the left side*), fillers (e.g., *uh, ah, um, er*), or editing expressions (e.g., *I mean, rather, that is, sorry, oops*). Although editing expressions (distinct from fillers) were coded, they turned out to be too rare to include in the disfluency counts or analyses. Where one disfluency occurred right after another (such as a restart after a filler), these were coded as separate disfluencies. If there were several disfluencies of the same type in a row (e.g., several repeated tokens of the same word), these were coded individually as well. Each repeated stretch of speech or restart was coded as one disfluency, even if the repeated or repaired phrase consisted of more than one word. Material was counted as a repeat only when it was repeated by the same speaker (as opposed to being echoed by the other speaker). When no repeats, restarts, fillers, or editing expressions were present, a turn was coded as

fluent. In those rare cases when it was unclear whether a turn or part of a turn contained any disfluency (e.g., when speech was unintelligible), it was coded as unknown.

### 2.2.3 Reliability

The text transcript of the corpus was divided into halves, and a different team of two coders coded each half. In addition, each team coded an additional six trials from the other team's half; these comprised Trials 2 and 4 from one pair randomly selected from each of the six between-subjects cells of the experimental design (2 levels of Relationship  $\times$  3 of Age). So 12.5% of the trials were double-coded. The coders were blind to which cells of the experimental design the speakers were in. Interrater reliability was excellent; there was 92.8% agreement, with a Cohen's Kappa of .91.

### 2.2.4 Coding for location of fillers

Next, another coder<sup>4</sup> judged each filler in the entire corpus as to whether it appeared within a turn that also involved any overlapping speech (either before or after the filler). Given that one alternative is that fillers may be used by a speaker to try to hold on to a turn, we coded the following: distance (in words) between the filler and any overlapping speech that followed, whether the filler appeared within an overlap, and whether the filler began or ended a turn. A filler was coded as beginning a turn if it was the first word in the turn or if it appeared just after a nonpropositional element that started the turn (including, e.g., discourse markers, e.g., *well*, or acknowledgments, e.g., *okay*). Consistent with this criterion, a filler ended a turn if no propositional element followed it in the turn. If the entire turn consisted of only a filler, it was coded as fitting none of these categories, but as appearing alone in the turn.

The coder also categorized each filler as to whether or not it appeared in between repeated stretches of speech, between the reparandum and the repair of a restart, or between syntactic constituents<sup>5</sup> as opposed to internal to (interrupting) a syntactic phrase. Phrase-internal fillers were those that appeared between the head of the phrase and some complement of the phrase; in addition, these were also categorized as to the most local phrase they interrupted (NP, VP, PP, degree phrase, or conjoined phrase, or else whether they appeared within a complement phrase just before a subordinated clause). The goals of this part of the coding scheme were to capture whether fillers tended to follow the (more or less) discrete installments that tend to make up utterances of conversational speech or whether they tended to *interrupt* such units, as well as to determine what sorts of syntactic constituents were most likely to be implicated. We were ultimately interested in knowing how fillers might be distributed either as displays of word-finding problems or covert repairs (likely functions for phrase-internal fillers) or as coordination signals (a likely function for between-phrase fillers, as well as for those that begin or end turns). The coding of filler locations is summarized in the *Appendix*.

---

<sup>4</sup> This coder was a 4th year Ph.D. student in linguistics who was not among the coauthors and who was naive to the purposes of our study.

<sup>5</sup> Fillers were coded as appearing between syntactic constituents if they appeared between any two of the following: NP, coordinated NP, topicalized element, VP, prepositional phrase, relative clause.

### 3. RESULTS

We began by examining word counts for the different types of speakers and conversations in this corpus; word counts in referential communication are assumed to be related to cognitive effort or task difficulty (Bortfeld & Brennan, 1997; Clark & Wilkes-Gibbs, 1986; Schober, 1995; Schober & Clark, 1989). In Schober and Carstensen's original study, matching tangrams was clearly more taxing than matching pictures of children: Director-matcher pairs made more errors on tangrams than they did on pictures of children, and tangrams were rated as the more difficult domain after the study (Schober & Carstensen, 2001). Married pairs and strangers performed equally well on the task, as did pairs from all three age groups.

#### 3.1 Word counts

All words, including fillers, word fragments, and other words implicated in repeats or restarts, were included in the word counts. Directors produced over twice as many words as did matchers,  $F(1,82)=98.14$ ,  $p<.001$ , in turns that were, on average, twice as long as matchers' turns, 11.4 words to 5.7 words. Directors' utterances ranged from 1–168 words, and matchers' ranged from 1–122. This matches our expectations for the division of labor between the two roles; directors typically took more of the initiative for verbally establishing the identity and location of the target object by describing and comparing it to other objects, while matchers spent much of their time searching for the target and giving acknowledgments, as here (overlapping speech appears between asterisks):

D: um, the next one is the kid with the Mickey Mouse t-shirt that says estimate uh, established \*nineteen twenty\*

M: \*gotcha\*

D: eight at the bottom?

M: okay

The domain of discussion mattered as well; overall, speakers used over two and a half times as many words to discuss tangrams as to discuss pictures of children,  $F(1,82)=246.21$ ,  $p<.001$ . This domain difference was greater for matchers than for directors, interaction,  $F(1,82)=14.46$ ,  $p<.001$ . In other words, with tangrams, the more taxing domain, matchers appeared to distribute more of their effort to the verbal part of the task, uttering 34% of the words in conversations about tangrams as opposed to only 27% in conversations about pictures of children. When establishing a referent was difficult, matchers often took the initiative and proposed alternative descriptions of target objects, as in this example:

D: uh hoo the next one is um let's see. okay it's similar to the one I said had the periscope

M: yes

D: but it would be scrunched down more and the second diamond is out in front like for feet... does that make any sense... the head...

M: uh

D: it's a smaller figure

M: there isn't there is no is there let me tell \*you\*

D: \*mm hm\*

**TABLE 2**

Word counts and disfluency rates per 100 words for young, middle-aged, and older married pairs and strangers and for familiar and unfamiliar domains (pictures of children & tangrams)

	<i>Fillers</i>	<i>Repeats</i>	<i>Restarts</i>	<i>Total</i>	<i>Word Counts</i>
Tangrams	2.38	1.77	2.22	6.37	717
Young	2.22	1.51	2.13	5.87	552
Middle-aged	2.04	1.86	2.00	5.89	773
Older	2.87	1.96	2.51	7.36	825
Children	2.74	1.17	1.65	5.55	283
Young	2.76	0.90	1.58	5.24	245
Middle-aged	2.56	1.38	1.54	5.47	303
Older	2.88	1.22	1.84	5.94	306
Married	2.56	1.36	2.00	5.91	488
Young	2.38	1.01	2.03	5.41	404
Middle-aged	2.25	1.53	1.64	5.42	506
Older	3.03	1.54	2.31	6.87	553
Strangers	2.56	1.59	1.88	6.02	516
Young	2.60	1.41	1.68	5.69	393
Middle-aged	2.34	1.71	1.90	5.95	576
Older	2.72	1.64	2.07	6.43	578
<b>Means, overall</b>	2.56	1.47	1.94	5.97	502

M: is the highest thing on there a diamond

D: yes

M: head

D: yes

M: and then right under the diamond head to the right is a triangle

D: uh huh

M: \*okay\*

D: \*uh huh\*

M: okay

There was no difference in the number of words uttered by men versus women, nor for married couples versus strangers. However, there were reliable word count differences by age: more words per round were uttered by older speakers (566) than by middle-aged speakers (541) and by middle-aged speakers than by younger speakers (399), linear trend,  $t(82)=3.53, p=.001$ . And there was an age-by-domain interaction,  $F(1,82)=5.50, p=.006$ : that is, word counts increased substantially with speaker's age in conversations about tangrams while they did so only slightly in conversations about children (see Table 2).

These differences in word counts are consistent with the macroplanning or cognitive load that speakers bear under different conditions; that is, in a referential communication task, directors need to plan longer utterances than matchers, and unusual objects are more

difficult to describe than common objects. Next we examined the disfluency rates and how they differed by domain, speaker's gender, marital status, role, length of utterance, and speaker's age. We also examined whether married pairs, who may be able to rely on an ability to coordinate built up over years of interacting with one another, produce more overlapping stretches of speech than strangers.

### 3.2 Disfluency rates

All the disfluency rates we report are per 100 words and consist of repeated words or phrases, restarts, and fillers, unless otherwise specified. Speakers produced, on average, 5.97 disfluencies every 100 words. This matches rates found by previous studies for these types of disfluencies.

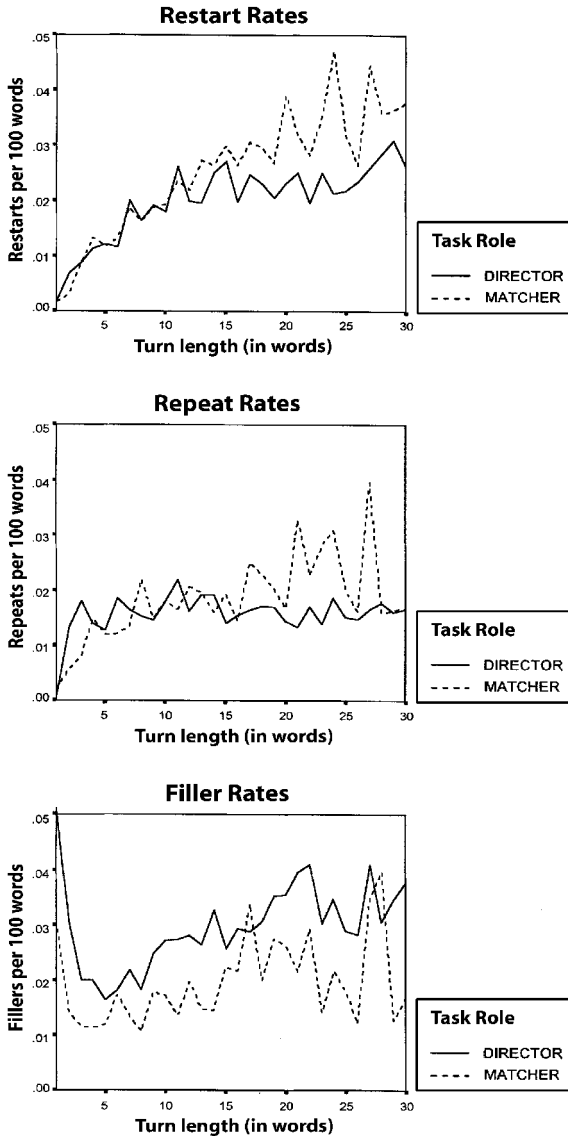
As predicted, disfluency rates increased when speakers were faced with heavier planning demands. This difference emerged in three ways: first, for unfamiliar versus familiar domains; second, for the task role of director versus matcher, and third, for longer turns versus shorter turns.

When speakers discussed tangrams, they produced greater rates of disfluencies than when they discussed children, 6.37 versus 5.55,  $F(1,82)=14.67, p<.001$ . This effect was due mainly to repeats (1.77 vs. 1.17,  $F(1,82)=41.64, p=.005$ ) and restarts (2.22 vs. 1.65,  $F(1,82)=37.82, p<.001$ ). For fillers, however, the difference was in the *opposite* direction: speakers produced slightly but reliably higher filler rates while describing children than tangrams, 2.74 versus 2.38,  $F(1,82)=8.00, p=.006$ . The finding that fillers are distributed differently than repeats and restarts supports the idea that they arise from different processes; we will return to this idea in the *General Discussion*.

In the role of director, speakers produced 7.00 disfluencies per 100 words versus 4.93 in the role of matcher,  $F(1,82)=96.99, p<.001$ . To break these disfluencies down further: Directors produced more fillers than matchers, 3.30 versus 1.81,  $F(1,82)=101.59, p<.001$ , more restarts, 2.17 versus 1.70,  $F(1,82)=21.10, p<.001$ , and slightly but not reliably higher rates of repeats, 1.53 versus 1.42,  $F(1,82)=1.34, ns$ .

Recall that Oviatt (1995) and Shriberg (1996) found that longer utterances are associated with higher disfluency rates than shorter utterances. Since our directors tended to produce longer utterances than their matcher partners, their higher disfluency rates might be entirely due to the difficulty of planning longer utterances rather than to any other aspects of the director roles. To tease apart role from utterance length, we compared director-matcher disfluency rates for turns<sup>6</sup> of a given length to see if the role differences persisted. This we did by removing unusually long turns (>30 words long, or 3 SDs from the mean), then computing disfluency rates for each remaining turn in the corpus, and then collapsing these to compute the average disfluency rates for each turn length. Filler rates were correlated with turn lengths .452 and .365 for directors and matchers respectively ( $ps<.05$ ), restart rates were correlated with turn lengths .817 and .904 for directors and matchers

<sup>6</sup> Note that we considered utterances to consist of turns. Strictly speaking, a turn may contain more than one utterance (when the addressee of an utterance fails to take a turn, a speaker may continue with another utterance—see Sacks et al., 1974). The problem is that determining where an utterance within a multiutterance turn ends can be quite subjective.

**Figure 1**

Rates of restarts, repeats, and fillers by turn lengths

respectively ( $ps < .01$ ), and repeats were correlated with turn lengths for matchers, .613 ( $p < .01$ ), but not for directors, .261 ( $ns$ ).

A more detailed look is provided by the graphs in Figure 1, which show the distribution of disfluency rates for directors and matchers for turns from 1–30 words long. For restart and repeat rates, the distributions for directors and matchers are indistinguishable for turns up to about 11 and 16 words, respectively. For turns longer than that, the distributions diverge, with matchers actually having higher restart and repeat rates than directors.

TABLE 3

Locations of fillers within turns

	<i>Subtotals</i>	<i>Fillers</i>	<i>% of Total Fillers</i>
<b>Turn-Initiating:</b>			
Beginning of turn		997	19.22
Alone (forming an entire turn)		196	3.78
<b>After an Installment:</b>			
End of turn		396	7.63
Mid-turn, before a fluent phrase		1186	22.86
<b>Within an Installment:</b>			
Within a disfluency		929	17.91
Restart ( <i>the happy uh the smiling child</i> )	646		
Repeat ( <i>the uh the happy child</i> )	283		
Mid-turn, Interrupting a Phrase		1483	28.59
Within a noun phrase	509		
Within a verb phrase	512		
Within a prepositional phrase	248		
Within a conjunction phrase	173		
Other (degree or complement phrase)	41		
<b>TOTAL</b>		5187	100.00

This pattern supports the explanation that for restart and repeat rates, increased utterance length is more of a driving factor than the role of director per se, at least for turns of average length. Since directors' turns were longer on average than matchers', most turn lengths in Figure 1 represent quite different numbers of turns contributed by directors versus by matchers, with the exception of turn lengths ranging from 6–10 words (directors contributed 2212 turns in this range, while matchers contributed a nearly equal number, 2228). For turns in this range, the effects of director-matcher role for restart rates,  $F(1,4) = .164$ , *ns*, and repeat rates,  $F(1,4) = .05$ , *ns*, disappear entirely.

However, there is quite a different pattern for fillers. With turns controlled for length, directors *still* produced consistently higher filler rates than matchers. And for turns in the 6–10 word range, turn lengths that were produced equally as often by directors as matchers, there was still a strong effect of role, with directors producing higher rates of fillers than matchers,  $F(1,4) = 19.42$ ,  $p = .01$ .

Next we examined filler rates and locations for evidence about their possible functions. Speakers used fillers to begin a turn, to end it, alone (forming the entire turn), between intact phrases, interrupting phrases that would otherwise have been fluent, and in the midst of restarts and repeats (see Table 3). The evidence is rather complex, as most locations of a filler within an utterance may have a variety of explanations. For instance, a turn-initial filler may reflect the effort of planning the entire utterance, or just finding an immediately upcoming word; it may be not only a symptom of difficulty but an actual display to the listener; it may serve as a signal by which a speaker establishes a turn at speaking (thereby taking over the conversational floor; see Sacks, Schegloff, & Jefferson [1974]) or by

which the speaker tries to keep another speaker at bay (Maclay & Osgood, 1959). A filler at the end of a turn may display a request for help, as proposed by Wilkes-Gibbs (1986) in her study of how speakers complete each other's utterances. And turns consisting of a filler alone may implicate any of these functions (see, for instance, the matcher's second turn in the previous tangram example).

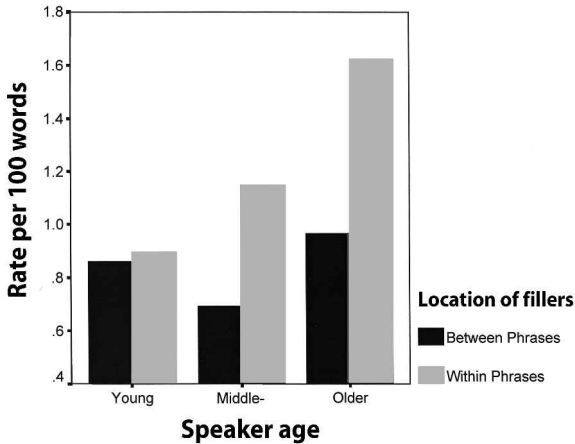
The picture is further complicated because conversational turns emerge from the coordinated action of more than one individual and are under no individual's control; an utterance-initial filler may have been a solicitation for help that never came, or a filler at the end of an utterance may be a failed attempt to prevent a partner from speaking. We proceed with these caveats in mind.

Because of the likelihood that fillers are related to coordination between speakers, we collapsed the filler data by pairs, including a measure, for each pair, of the rate at which they produced overlapping stretches of speech (normalized for the total number of words the pair produced). We then considered the possibility that pairs of speakers with higher rates of overlaps might produce higher rates of fillers. One set of mediating assumptions would be that such pairs competed more for the conversational floor, and that using fillers would be an attempt to manage this competition. However, filler rates were not correlated with overlap rates ( $r = .195, ns$ ). Note that overlaps need not show competition between speakers for a limited resource, but could well be used as an interactive resource for collaborative ends. In our corpus, matchers routinely overlapped directors to indicate they understood, a very efficient strategy. Jefferson (1973) has argued that overlapping speech can be the result of fine coordination between speakers (see also Leon, 1996).

Speakers often deliver utterances in short installments, and it has been proposed that addressees project the ends of syntactic or intonational phrases as cues for when to take over the speaking turn (Sacks et al., 1974). So if fillers signal mainly that speakers plan to continue speaking, then we would expect to see more fillers *between* than *within* phrases. This did not turn out to be the case; the categories in Table 3 show a total of 1582 fillers between phrases or at the ends of turns (note that these were coded without appeal to intonation) versus 2412 fillers within otherwise fluent phrases or disfluencies (none of which would have qualified as "transition relevance places" in Sacks et al.'s terminology). We compared the rates of between- against within-constituent fillers for the 48 pairs, with the pairs' overlap rates as a covariate. The only effects had to do with age: older speakers used more fillers of both sorts,  $t(41) = 3.12, p = .004$ , and speaker age interacted with the location of fillers between or within constituents. Figure 2 shows that with increasing age, pairs produced higher rates of fillers within-constituents but similar rates between constituents. This is consistent with findings that older people have more trouble retrieving words than younger people (Obler & Albert, 1984; Sandson, Obler, & Albert, 1987). Note that a filler may simultaneously have cognitive and interpersonal explanations; it is possible for a filler to be not only a symptom of a word-finding problem but also a display that solicits help with the problem.

Overall, older speakers produced higher disfluency rates (6.65, with repeats, restarts, and fillers combined) than middle-aged (5.69) and younger (5.55) speakers, linear trend,  $t(82) = 1.94, p < .05$ . The important distinction was whether the speaker was in the older group (which ranged from 63 to 72 years of age); there was no difference between the



**Figure 2**

Use of fillers by speakers of different ages

younger and middle-aged groups. This was true despite the fact that the middle-aged pairs uttered more words than the young pairs and similar amounts as the older pairs. See Table 2, which summarizes disfluency rates for young, middle-aged, and older married couples and strangers.

As for married pairs, they were no more fluent in their conversations than were strangers; there were no differences by relationship in rates of restarts, repeats, or fillers. This is contrary to what would be expected if experience or comfort with a partner were to increase fluency, or if anxiety evoked by conversing with an unfamiliar partner were to increase disfluency. It is also contrary to what would be expected if strangers planned their speech more carefully than intimates, or if certain disfluencies were coordination devices that *only* intimates could use to elicit help from their partners.

The only reliable difference between married pairs and strangers was in the number of times in which one partner's speech overlapped the other's. When speech overlaps, it is sometimes assumed that one speaker is interrupting or taking the conversational floor away from another. On this view, we might expect fewer interruptions between strangers than married partners, since politeness or formality seems more likely to characterize the speech of people who do not know one another. On the other hand, if speakers who know each other well are better at coordinating their interaction, we would expect more overlaps between strangers than married pairs. This is what we found: pairs of strangers overlapped parts of their utterances an average of 4.9 times per 100 words, in contrast to 3.9 times per 100 words for pairs married to each other,  $F(1,42)=5.70, p=.02$ .

Recall that Shriberg (1996) found that men produced more fillers than women did. To see if our data followed this pattern, we included speaker's gender in our comparisons. We found that while men produced no more words than women did, they had a higher rate of disfluencies overall, 6.80 to 5.12 per 100 words,  $F(1,82)=15.05, p=.001$ . Why should men be more disfluent than women? When we broke this difference down further, it turned out to be due mainly to higher rates of fillers, 3.04 to 2.07,  $F(1,82)=12.67, p=.001$  and repeats, 1.74 to 1.21,  $F(1,82)=9.39, p=.003$ . It is worth noting that clinical disfluency, or stuttering, is more common in males than females (Guyette & Baumgartner, 1988; Yairi & Ambrose, 1992). Although none of our speakers would have qualified for a clinical

diagnosis of stutterer (and although speech by stutterers is considered by most to be qualitatively different than speech by nonstutterers), we believe the higher incidence of repeats by men is worth noting. Men produced slightly but not significantly higher rates of restarts than women, 2.03 to 1.85.

Acting in the role that involved taking more initiative made more of a difference in men's disfluency rates than in women's. Men produced 2.57 more disfluencies per 100 words when they were directors than when they were matchers, while women produced only 1.58 more, interaction,  $F(1,82)=5.61, p=.02$ . The main contributor to this finding was the filler rate; men produced 1.99 more fillers per 100 words as directors than as matchers, and women produced only 1.00 more, interaction,  $F(1,82)=10.39, p=.002$  (this pattern was not present for restart or repeat rates). Not only were there higher filler rates in conversations about children than about tangrams, but the gender by role interaction for filler rates was more pronounced with children than with tangrams, three-way interaction,  $F(1,82)=6.47, p=.01$ . Again, this was due to men, who produced moderately higher filler rates as directors than as matchers when discussing tangrams but much higher filler rates as directors than as matchers (more than double, in fact) when discussing children (see Table 4A). Women produced moderately higher filler rates as directors than as matchers, and this happened regardless of what they were discussing. Perhaps male directors describing children produced more fillers because they found this domain taxing (a cognitive load explanation). This explanation is not particularly satisfying by itself, since men produced so many more words describing tangrams than they did describing children. Perhaps, on the other hand, they used more fillers to let their female partners know when they were having trouble producing appropriate descriptions of the children or to solicit help from them (implicating a coordination function for fillers); consider the following example:

D: number ten has uh something written there on uh the right side of his shirt above the pocket. It says "sky" something or uh

M: mm hm

D: little boy with uh. grin \*nice grin\*

M: \*mm hm\* white shirt, striped pants

D: striped pants, yeah, that's it

M: okay

Since these are post hoc explanations — we did not predict gender differences on any theoretical basis — we offer them with caution and because of their descriptive interest. We would like to see these gender differences replicated with other corpora before drawing conclusions about them.

#### 4. GENERAL DISCUSSION

We are aware that corpus analysis has serious limitations because of the need for categorizing diverse conversational events together; the events in a category may resemble one another on some dimension but perform quite different functions or have effects over quite different contexts. It is certainly possible for one event (such as an *uh*) to have multiple functions simultaneously.

**TABLE 4A**

Filler rates per 100 words by gender, role, and domain (word counts in parentheses)

	<i>Tangrams</i>	<i>Children</i>	<i>Both Domains</i>
<b>Male Speakers</b>	2.73 (730)	3.36 (279)	3.04 (506)
As Directors	3.49 (943)	4.60 (402)	4.04 (675)
As Matchers	1.97 (517)	2.13 (159)	2.05 (338)
<b>Female Speakers</b>	2.03 (703)	2.11 (287)	2.07 (495)
As Directors	2.57 (943)	2.57 (424)	2.57 (684)
As Matchers	1.49 (464)	1.65 (149)	1.57 (306)
<b>All Speakers</b>	2.38 (717)	2.74 (283)	2.56 (500)

**TABLE 4B**

Repeat rates per 100 words by gender, role, and domain (restart rates in parentheses)

	<i>Tangrams</i>	<i>Children</i>	<i>Both Domains</i>
<b>Male Speakers</b>	2.01 (2.32)	1.46 (1.72)	1.74 (2.03)
As Directors	2.23 (2.60)	1.27 (1.98)	1.76 (2.30)
As Matchers	1.79 (2.05)	1.64 (1.47)	1.72 (1.76)
<b>Female Speakers</b>	1.54 (2.12)	0.87 (1.58)	1.21 (1.85)
As Directors	1.63 (2.39)	0.96 (1.72)	1.30 (2.05)
As Matchers	1.45 (1.85)	0.78 (1.44)	1.12 (1.65)
<b>All Speakers</b>	1.77 (2.22)	1.17 (1.65)	1.47 (1.94)

With this caveat in mind, we selected a corpus that varied task role (director vs. matcher), difficulty of domain (abstract geometric figures vs. photographs of children), relationship between speakers (married to each other vs. strangers), and gender (each pair of speakers consisted of a man and a woman). This design enabled us to make direct comparisons of disfluency rates across conditions, unlike studies that have made comparisons across corpora (e.g., Oviatt, 1995). We coded disfluencies by their surface forms and not by any attributions about what the speakers' motivations may have been (e.g., Carletta et al. (1993) coded a wider variety of expressions as fillers when it appeared that the speaker was stalling), since we are interested in examining possible causes and functions of disfluencies. We avoided coding silent pauses since their identification is extremely subjective. Our corpus was relatively large, and the coding highly reliable. The fillers we coded were limited to *um*, *uh*, *er*, and *ah*, unlike some other studies (e.g., Broen & Siegel, 1972, included *well*, *you know*, and *I mean* as fillers despite the possibility that these discourse markers have quite distinct discourse functions; see Schiffrin, 1987).

From our data, we advance several main conclusions. First, the distributions of disfluencies in this corpus support the idea that some but not all disfluency rates increase as heavier demands are placed on the speech planning system. Tangrams are clearly harder than pictures of children for people to describe and match; this assumption was corrobora-

rated by the elevated word counts and error rates for tangrams over children. Descriptions of tangrams contained reliably higher repeat rates compared to descriptions of children. Another way in which planning demands are heavier is across task roles; those who took the initiative (directors) produced higher restart and filler rates (but similar repeat rates) compared to those who did not (matchers). For restart rates, the effect of task role is due to the fact that directors must plan longer utterances on average than matchers.

Second, fillers were distributed somewhat differently than repeats and restarts, suggesting that they may be related to processes of interpersonal coordination. With a more difficult task, speakers are more likely to have trouble *and* to display that trouble to an addressee, so the effects of cognitive load will not be independent of effects of interpersonal coordination (if, indeed, the latter are at work). If fillers help speakers coordinate with their addressees (e.g., by displaying delays in producing utterances and perhaps by soliciting help or responses from partners), then we should expect directors, who take most of the initiative in a matching task, to produce more fillers than matchers. We found this to be the case even when turn length was controlled, so directors' elevated rates of fillers are probably not due to cognitive load alone. This effect of role was also consistent across both domains and for both sexes.

Furthermore, when it came to the topic domain of conversations, more fillers were produced while discussing pictures of children than tangrams, which should not have been the case if fillers were due mainly to planning difficulty. As Table 4A shows, the elevated filler rate for pictures of children is due entirely to male directors. This suggests an interesting (although post hoc) explanation: Perhaps male speakers perceived an imbalance between themselves and their female partners in expertise about the face pictures and so were more likely to display their trouble and appeal to their partners for help. We do not have independent confirmation for this notion, however, and so we offer this possibility not as a conclusion, but as a hypothesis for future study.

Third, while our study does not directly examine speech to machines, the data do hold some relevance for spoken-language systems. Previous studies (Oviatt, 1995; Shriberg, 1996) have not fully accounted for the intriguing finding that people are more fluent in speech to machines than to other people (in comparisons across corpora). Oviatt (1995) found that some of the difference appears to be due to the fact that utterances to machines are typically shorter and take place within a more structured format in which the machine has the initiative (and so people have to do less planning), but that even when utterance length was controlled for, disfluency rates to machines were still lower. Low disfluency rates to machines also seem to fly in the face of early studies that managed to associate disfluencies (other than fillers) with anxiety (see Mahl, 1987 for a review)—if anxiety is at work, then we should expect people to be more disfluent with machines, not less<sup>7</sup>. In

---

<sup>7</sup> This expectation has intuitive appeal, although Oviatt's and Shriberg's data show *lower* disfluency rates with machines. Other researchers have shown that speakers who are not used to speaking to machines are more likely to be misunderstood by them (Kubala, Barry, Bates, Bobrow, Fung, Ingria, Makhoul, Nguyen, Schwartz, & Stallard, 1992). But errors in speech recognition by machines may have causes other than disfluencies; for instance, if speakers know the vocabulary that a system accepts, they can produce utterances that are more distinguishable (with lower perplexity) than if they do not know the vocabulary.

our corpus, people were as disfluent with strangers as they were with their spouses, and so even if they were more anxious with strangers (note that there was no independent measure of anxiety collected from the speakers in our corpus), such hypothetical anxiety apparently did not affect disfluency rates. That anxiety by itself should not automatically lead to disfluency is consistent with Oviatt's and Shriberg's lower disfluency rates to machines (if we think of a computer partner as one that is even stranger than a stranger). In fact, speakers who try to take more care with their speech may succeed in producing more fluent utterances.

We are left with three (nonexclusive) possibilities, then, for explaining the lower disfluency rates with machines: (1) perhaps people are simply more careful when speaking with machines, (2) perhaps disfluencies (particularly, fillers) are related to coordination processes that are different with machine partners than with human partners (if speakers do not display their difficulties to machine partners), and (3) perhaps, as Oviatt suggests, the corpora involving speech to machines offered a more structured format than those involving speech to people. To understand this last possibility would require a within-experiment comparison of speech to human and computer partners in which both the user and the system were able to take the initiative.

Finally, a more direct way in which our study is relevant to machine-directed speech is in the demographic characteristics of the speakers and the speakers' knowledge of the task domain at hand. Consistent with aging research that finds certain tasks becoming more difficult with age, in our study older speakers had marginally higher disfluency rates (about one more disfluency per 100 words) than middle-aged and younger ones. This was true even though middle-aged speakers produced as many words as older ones to do the same task equally well (making these two groups less efficient than younger speakers). And men produced about 1.5 more fillers and repeats combined per 100 words than women did. As for planning load, speakers produced about two more disfluencies per 100 words when performing as directors than they did as matchers, and about one more repeat and restart (combined) per 100 words when discussing an unfamiliar versus a familiar domain. These rates may ultimately be useful in constructing a speaker model that predicts disfluency rates based on who the speakers are, what they are talking about, and whether they are taking the initiative.

---

*Received: July 15, 1998; revised manuscript received: June 12, 2000;  
accepted: November 13, 2000*

---

## REFERENCES

- ALBERT, M. (1980). Language in normal and dementing elderly. In L. K. Obler & M. L. Albert (Eds.), *Language and communication in the elderly*. Lexington, MA: DC Heath and Co.
- BALL, P. (1975). Listeners' responses to filled pauses in relation to floor apportionment. *British Journal of Social and Clinical Psychology*, **14**, 423–424.
- BLOOM, J. E., & SCHOBBER, M. F. (1999). Respondent cues that survey questions are in danger of being misunderstood. *Proceedings of the American Statistical Association, Section on Survey Research Methods* (pp.992–997). Alexandria, VA: American Statistical Association.
- BOCK, J. K. (1986). Meaning, sound, and syntax: Lexical priming in sentence production. *Journal of Experimental Psychology: Learning, Memory, and 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.
- BOOMER, D. S. (1965). Hesitation and grammatical encoding. *Language and Speech*, **8**, 148–158.
- 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.
- BRENNAN, S. E., & CLARK, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **22**, 1482–1493.
- BRENNAN, S. E., & KIPP, E. G. (1996). An addressee's knowledge affects a speaker's use of fillers in question-answering. *Abstracts of the Psychonomic Society, 37th Annual Meeting* (p. 24), Chicago, IL.
- BRENNAN, S. E., & SCHOBER, M. F. (2001). How listeners compensate for disfluencies in spontaneous speech. *Journal of Memory and Language*, **44**, 274–296.
- 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. *Journal of Memory and Language*, **34**, 383–398.
- BROEN, P. A., & SIEGEL, G. M. (1972). Variations in normal speech disfluencies. *Language and Speech*, **15**, 219–231.
- BURKE, D. M., MacKAY, D. G., WORTHLEY, J. S., & WADE, E. (1991). On the tip of the tongue: What causes word finding failures in young and older adults? *Journal of Memory and Language*, **30**, 542–579.
- BUTZBERGER, J. W., MURVEIT, H., SHRIBERG, E., & PRICE, P. (1992). Spontaneous speech effects in large vocabulary speech recognition applications. In M. Marcus (Ed.), *Proceedings, DARPA Speech and Natural Language Workshop* (pp. 339–343). Morgan Kaufmann.
- CARLETTA, J., CALEY, R., & ISARD, S. (1993). *A collection of self-repairs from the map task corpus* (Technical Report HCRC/TR-47). Edinburgh, Scotland: Human Communication Research Center.
- CARLETTA, J., CALEY, R., & ISARD, S. (1995). Simulating time-constrained language production. *Language and Cognitive Processes*, **10**, 357–361.
- CHRISTENFELD, N., SCHACHTER, S., & BILOUS, F. (1991). Filled pauses and gestures: It's not coincidence. *Journal of Psycholinguistic Research*, **20**, 1–10.
- CLARK, H. H. (1994). Managing problems in speaking. *Speech Communication*, **15**, 243–250.
- 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). Washington, DC: APA. Reprinted in R. M. Baecker (Ed.), *Groupware and computer-supported cooperative work: Assisting human-human collaboration* (pp. 222–233). San Mateo, CA: Morgan Kaufman Publishers, Inc.
- CLARK, H. H., & WASOW, T. (1998). Repeating words in spontaneous speech. *Cognitive Psychology*, **37**, 201–242.
- CLARK, H. H., & WILKES-GIBBS, D. (1986). Referring as a collaborative process. *Cognition*, **22**, 1–39.
- COOK, M., & LALLJEE, M. (1970). The interpretation of pauses by the listener. *British Journal of Social and Clinical Psychology*, **9**, 375–376.
- DELL, G. S. (1986). A spreading activation theory of retrieval in sentence production. *Psychological Review*, **93**, 283–321.
- DIJKSTRA, W. (1999). Sequence 5, Version 5.0. Copyright, Wil Dijkstra and Vrije Universiteit, Department of Social Research Methodology, S. C. W. V. U. Amsterdam, NL.
- FOX TREE, J. E. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language*, **34**, 709–738.
- FOX TREE, J. E., & CLARK, H. H. (1997). Pronouncing “the” as “thee” to signal problems in speaking. *Cognition*, **62**, 151–167.
- FROMKIN, V. A. (1971). The nonanomalous nature of anomalous utterances. *Language*, **47**, 27–52.
- FROMKIN, V. A. (Ed.) (1973). *Speech errors as linguistic evidence*. The Hague: Mouton Publishers.

- FROMKIN, V. A. (1980). *Errors in linguistic performance: Slips of the tongue, ear, pen and hand*. New York: Academic Press.
- GARRETT, M. F. (1975). The analysis of sentence production. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 9) (pp. 133–177). New York: Academic Press.
- GOLDMAN-EISLER, F. (1958). Speech production and the predictability of words in context. *Quarterly Journal of Experimental Psychology*, **10**, 96.
- GUYETTE, T., & BAUMGARTNER, J. (1988). Stuttering in the adult. In N. Lass, L. McReynolds, J. Northern, & D. Yoder (Eds.), *Handbook of speech-language pathology and audiology* (pp. 640–654). Philadelphia, PA: B. C. Decker Inc.
- HINDLE, D. (1983). Deterministic parsing of syntactic nonfluencies. In *Proceedings of the 21st Annual Meeting, Association for Computational Linguistics*, Cambridge, MA, 123–128.
- JEFFERSON, G. (1973). A case of precision timing in ordinary conversation: Overlapped tag-positioned address terms in closing sequences. *Semiotica*, **9**, 47–96.
- KASL, S. V., & MAHL, G. F. (1965). The relationship of disturbances and hesitations in spontaneous speech to anxiety. *Journal of Personality and Social Psychology*, **1**, 425–433.
- KUBALA, F., BARRY, C., BATES, M., BOBROW, R., FUNG, P., INGRIA, R., MAKHOUL, J., NGUYEN, L., SCHWARTZ, R., & STALLARD, D. (1992). BBN BYBLOS and HARC February, 1992 ATIS Benchmark results. In M. Marcus (Ed.), *Fifth DARPA Speech and Natural Language Workshop* (pp. 72–77). San Mateo, CA: Morgan Kaufmann.
- LALLJEE, M., & COOK, M. (1969). An experimental investigation of the function of filled pauses in speech. *Language and Speech*, **12**, 24–28.
- LEON, S. D. (1996). *Experiencing interruptions in family conversations*. Unpublished MA thesis, New School for Social Research, New York.
- LEVELT, W. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.
- LICKLEY, R. J. (1996). Juncture cues to disfluency. *Proceedings, International Conference on Spoken Language Processing (ICSLP '96)*, Philadelphia, 2478–2481.
- LICKLEY, R. J., & BARD, E. G. (1996). On not recognizing disfluencies in dialog. *Proceedings, International Conference on Spoken Language Processing (ICSLP '96)*, Philadelphia, 1876–1879.
- LICKLEY, R. J., & BARD, E. G. (1998). When can listeners detect disfluency in spontaneous speech? *Language and Speech*, **41**, 203–226.
- MACKEY, D. G. (1970). Spoonerisms: The structure of errors in the serial order of speech. *Neuropsychologia*, **8**, 323–350.
- MACKEY, D. G. (1972). The structure of words and syllables: Evidence from errors in speech. *Cognitive Psychology*, **3**, 210–227.
- MACKEY, D. G. (1973). Complexity in output systems: Evidence from behavioral hybrids. *American Journal of Psychology*, **86**, 785–806.
- MACLAY, H., & OSGOOD, C. E. (1959). Hesitation phenomena in spontaneous English speech. *Word*, **15**, 19–44.
- MAHL, G. F. (1987). *Explorations in nonverbal and vocal behavior*. Hillsdale, NJ: Erlbaum.
- NAKATANI, C. H., & HIRSCHBERG, J. (1994). A corpus-based study of repair cues in spontaneous speech. *Journal of the Acoustical Society of America*, **95**, 1603–1616.
- NOOTEBOOM, S. G. (1969). The tongue slips into patterns. In A. G. Sciarone, A. J. van Essen, & A. A. van Raad (Eds.), *Nomen (Society): Leyden studies in linguistics and phonetics* (pp. 114–132). The Hague: Mouton Publishers.
- OBLER, L., & ALBERT, M. L. (1984). Language in aging. In M. L. Albert (Ed.), *Clinical neurology of aging* (pp. 245–253). New York, NY: Oxford University Press.
- OVIATT, S. (1995). Predicting spoken disfluencies during human-computer interaction. *Computer Speech and Language*, **9**, 19–35.
- RASTLE, K. G., & BURKE, D. M. (1996). Priming the tip of the tongue: Effects of prior processing on word retrieval in young and older adults. *Journal of Memory and Language*, **35**, 586–605.
- ROCHESTER, S. R. (1973). The significance of pauses in spontaneous speech. *Journal of Psycholinguistic Research*, **2**, 51–81.

- SACKS, H., SCHEGLOFF, E., & JEFFERSON, G. (1974). A simplest systematics for the organization of turn-taking in conversation. *Language*, **50**, 696–735.
- SANDSON, J., OBLER, L. K., & ALBERT, M. L. (1987). Language changes in healthy aging and dementia. In Rosenberg, S. (Ed.), *Advances in applied psycholinguistics* (pp. 264–292). New York, NY: Oxford University Press.
- SCHACHTER, S., RAUSCHER, F., CHRISTENFELD, N., & CRONE, K. T. (1994). The vocabularies of academia. *Psychological Science*, **5**, 37–41.
- SCHACHTER, S., CHRISTENFELD, N., RAVINA, B., & BILOUS, F. (1991). Speech disfluency and the structure of knowledge. *Journal of Personality and Social Psychology*, **60**, 362–367.
- SCHIFFRIN, D. (1987). *Discourse markers*. Cambridge, U.K.: Cambridge University Press.
- SCHOBER, M. F. (1995). Speakers, addressees, and frames of reference: Whose effort is minimized in conversations about locations? *Discourse Processes*, **20**, 219–247.
- SCHOBER, M. F., & CARSTENSEN, R. (2001). *Do age and long-term relationship matter in conversations about unfamiliar things?* Manuscript under revision.
- SCHOBER, M. F., & CLARK, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology*, **21**, 211–232.
- SCHOW, R., CHRISTENSEN, J., HUTCHINSON, J., & NERBONNE, M. (1978). *Communication disorders of the aged: A guide for health professionals*. Baltimore, MD: University Park Press.
- SHATTUCK-HUFNAGEL, S. (1979). Speech errors as evidence for a serial order mechanism in sentence production. In W. E. Cooper & E. C. T. Walker (Eds.), *Sentence processing: Psycholinguistic studies presented to Merrill Garrett* (pp. 295–342). Hillsdale, NJ: Lawrence Erlbaum.
- SHATTUCK-HUFNAGEL, S. (1982). Three kinds of speech error evidence for the role of grammatical elements in processing. In L. K. Obler & L. Menn (Eds.), *Exceptional language and linguistics* (pp. 133–142). New York: Academic Press.
- SHATTUCK-HUFNAGEL, S., & KLATT, D. (1979). The limited use of distinctive features and markedness in speech production: Evidence from speech error data. *Journal of Verbal Learning and Verbal Behavior*, **18**, 41–55.
- SHEWAN, C. M., & HENDERSON, V. L. (1988). Analysis of spontaneous language in the older normal population. *Journal of Communication Disorders*, **21**, 139–154.
- SHRIBERG, E. (1996). Disfluencies in Switchboard. *Proceedings, International Conference on Spoken Language Processing (ICSLP '96)*, 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.), *Proceedings, DARPA Speech and Natural Language Workshop* (pp. 419–424). Morgan Kaufmann.
- SHRIBERG, E., & LICKLEY, R. J. (1993). Intonation of clause-internal filled pauses. *Phonetica*, **50**, 172–179.
- SHRIBERG, E., WADE, E., & PRICE, P. (1992). Human-machine problem solving using spoken language systems (SLS): Factors affecting performance and user satisfaction. In M. Marcus (Ed.), *Proceedings, DARPA Speech and Natural Language Workshop* (pp. 49–54). Morgan Kaufmann.
- SMITH, V., & CLARK, H. H. (1993). On the course of answering questions. *Journal of Memory and Language*, **32**, 25–38.
- WILKES-GIBBS, D. (1986). *Collaborative processes of language use in conversation*. Unpublished doctoral dissertation, Stanford University, Stanford, CA.
- YAIRI, E., & AMBROSE, N. (1992). Onset of stuttering in preschool children: Selected factors. *Journal of Speech and Hearing Research*, **35**, 782–788.
-



## APPENDIX

### A. Types of disfluencies or editing expressions (mutually exclusive)

Category:	Example:
Repeat	e.g., <i>just on the left left side</i>
Restart	e.g., <i>imme- just below the left side</i>
Filler	e.g., <i>uh, ah, um, er</i>
Editing Expression	e.g., <i>I mean, rather, that is, sorry, oops</i>

### B. Locations of fillers (mutually exclusive)\*

*If a disfluency is a filler, then code whether it appears:*

Just before a restart (after the reparandum and before the repair)	e.g., <i>no it's more of uh it's flat on the bottom</i>
Just before a repeat	e.g., <i>with his um his mouth twisted</i>
Alone, such that the filler is the whole turn	e.g., <i>uh</i>
At the very start of a turn (before any prepositional content)	e.g., <i>uh, the next one is...</i>
At the very end of a turn	e.g., <i>nope he has on a belt um</i>
Between phrases	e.g., <i>a triangle um that is hanging down</i>
Within (internal to) a phrase	e.g., <i>and a white um shirt</i>

### C. Syntactic locations of phrase-internal fillers (mutually exclusive)\*\*

*If the filler is phrase-internal, then code whether it appears as a:*

Noun phrase	e.g., <i>the uh small figure with the square on top</i>
Verb phrase	e.g., <i>see um the circle</i>
Prepositional phrase	e.g., <i>the small figure with um the square on top</i>
Conjunction phrase	e.g., <i>the triangle and um the circle</i>
Complement phrase	e.g., <i>the figure that uh has the squiggly on the side</i>

\* The mutually exclusive codes within B were assigned precedence in the order in which they are listed (e.g., for B codes, if a filler appeared just before a restart, it was not coded as to whether it appeared between phrases or internal to a phrases).

\*\* Concerning the syntactic locations in C, we coded fillers as internal to the most local phrase, using only these five types of constituents. For constituents that we did not code (such as adverbs), we simply coded any filler as internal to the next highest constituent (e.g., 'moves very um quickly' would have been coded as internal to the verb phrase, and 'the very um tall man' would have been coded as internal to the noun phrase).