PREDICTING AND MANAGING SPOKEN DISFLUENCIES DURING HUMAN-COMPUTER INTERACTION*

Sharon Oviatt

Computer Dialogue Laboratory & Artificial Intelligence Center
SRI International, 333 Ravenswood Avenue, Menlo Park, CA. 94025

ABSTRACT

This research characterizes the spontaneous spoken disfluencies typical of human-computer interaction, and presents a predictive model accounting for their occurrence. Data were collected during three empirical studies in which people spoke or wrote to a highly interactive simulated system. The studies involved within-subject factorial designs in which input modality and presentation format were varied. Spoken disfluency rates during human-computer interaction were documented to be substantially lower than rates typically observed during comparable human-human speech. Two separate factors, both associated with increased planning demands, were statistically related to increased speech disfluency rates: (1) length of utterance, and (2) lack of structure in the presentation format. Regression techniques revealed that a linear model based simply on utterance length accounts for over 77% of the variability in spoken disfluencies. Therefore, design techniques capable of channeling users' speech into briefer sentences potentially could eliminate most spoken disfluencies. In addition, the degree of structure in the presentation format was manipulated in a manner that successfully eliminated 60 to 70% of all disfluent speech. The long-term goal of this research is to provide empirical guidance for the design of robust spoken language technology.

1. INTRODUCTION

Recently, researchers interested in spoken language processing have begun searching for reliable methods to detect and correct disfluent input automatically during interactions with spoken language systems [2, 4, 9]. In general, this research has focused on identifying acoustic-prosodic cues for detecting self-repairs, either alone or in combination with syntactic, semantic, and pattern matching information. To date, however, possible avenues for simply reducing or eliminating disfluencies through manipulation of basic interface features have not been explored.

Another underdeveloped but central theme in disfluency research is the relation between spoken disfluencies and planning demands. Although it is frequently claimed that disfluencies rise with increased planning demands of different kinds [3], the nature of this relation remains poorly understood. The major factors contributing to planning have yet to be identified and defined in any comprehensive manner, or linked to disfluencies and self-repairs. From the viewpoint of designing systems, information on the dynamics of what produces disfluencies, and how to structure interfaces to minimize them, could improve the robust performance of spoken language systems.

A related research issue is the extent to which qualitatively different types of speech may differ in their disfluency rates. That is, does the rate of spoken disfluencies tend to be stable, or variable? If variable, do disfluency rates differ systematically between human-human and human-computer speech? And are disfluency rates sufficiently variable that techniques for designing spoken language interfaces might exert much leverage in reducing them? To compare disfluency rates directly across different types of human-human and human-computer interactions, research needs to be based on comparable rate-per-word measures, the same definition of disfluencies and self-repairs, and so forth, in order to obtain meaningful comparisons.

For the purpose of the present research, past studies by the author and colleagues [1, 6, 7] were reanalyzed: (1) to yield data on the rate of disfluencies for four different types of human-human speech, and (2) to conduct comparative analyses of whether human-human disfluencies differ from human-computer ones. In addition, three simulation studies of human-computer interaction were conducted, which generated data on spoken and handwritten disfluencies. Apart from comparing disfluencies in different communication modalities, two separate factors associated with planning demands were examined. First, presentation format was manipulated to investigate whether degree of structure might be associated with disfluencies. It was predicted that a relatively unconstrained format, which requires the speaker to self-structure and plan to a greater degree, would lead to a higher rate of speech disfluencies. Second, the rate of disfluencies was examined in sentences of varying length. Spoken utterances graduated in length were compared to determine whether longer sentences have an elevated rate of disfluencies per word, since they theoretically require more planning. Finally, implications are outlined for designing future interfaces capable of substantially reducing disfluent input.

2. SIMULATION EXPERIMENTS ON HUMAN-COMPUTER INTERACTION

This section outlines three experiments on human spoken and handwritten input to a simulated system, with spoken disflu-

*This research was supported by Grant No. IRI-9213472 from the National Science Foundation, contracts from USWest, AT&T/NCR, and ATR International to SRI International, and equipment donations from Apple Computer, Sun Microsystems, and Wacom Inc.

encies constituting the primary analytical focus.

2.1. Method

Subjects, Tasks, and Procedure- Forty-four subjects participated in this research as paid volunteers. A "Service Transaction System" was simulated that could assist users with tasks that were either (1) verbal-temporal (e.g., conference registration or car rental exchanges, in which proper names and scheduling information predominated), or (2) computational-numeric (e.g., personal banking or scientific calculations, in which digits and symbol/sign information predominated). During the study, subjects first received a general orientation to the Service Transaction System, and then were given practice using it to complete tasks. They received instructions on how to enter information on the LCD tablet when writing, speaking, and free to use both modalities. When speaking, subjects held a stylus on the tablet as they spoke.

People also were instructed on completing tasks in two different presentation formats. In an unconstrained format, they expressed information in an open workspace, with no specific system prompts used to direct their speech or writing. People simply continued providing information while the system responded interactively with confirmations. For example, in this format they spoke digits, computational signs, and requested totals while holding their stylus on an open "scratch pad" area of their LCD screen. During other interactions, the presentation format was explicitly structured, with linguistic and graphical cues used to structure the content and order of people's input as they worked. For example, in the verbal-temporal simulations, form-based prompts were used to elicit input (e.g., **Car pickup location** [ ]), and in the computational-numeric simulation, patterned graphical layouts were used to elicit specific digits and symbols/signs.

Other than specifying the input modality and format, an effort was made not to influence the manner in which people expressed themselves. People's input was received by an informed assistant, who performed the role of interpreting and responding as a fully functional system would. Essentially, the assistant tracked the subject's written or spoken input, and clicked on predefined fields at a Sun SPARCstation to send confirmations back to the subject.

Semi-Automatic Simulation Technique- In developing this simulation, an emphasis was placed on providing automated support for streamlining the simulation to the extent needed to create facile, subject-paced interactions with clear feedback, and to have comparable specifications for the different input modalities. In the present simulation environment, response delays averaged 0.4 second, with less than a 1-second delay in all conditions. In addition, the simulation was organized to transmit analogues of human backchannel and propositional confirmations, with propositional-level confirmations embedded in a compact transaction receipt. The simulation also was designed to be sufficiently automated so that the assistant could concentrate attention on monitoring the accuracy of incoming information, and on maintaining sufficient vigilance to ensure prompt responding. This semi-automation contributed to the fast pace of the simulation, and to a low rate of technical errors. Details of the simulation technique and its capabilities have been detailed elsewhere [8].

Research Design and Data Capture- Three studies were completed in which the research design was a completely crossed factorial with repeated measures. In all studies, the main factors of interest included: (1) communication modality - speech-only, pen-only, combined pen/voice, and (2) presentation format - form-based, unconstrained. The first two studies examined disfluencies during communication of verbal-temporal content. To test the generality of certain findings, a third study was conducted that compared disfluencies in computational-numeric content.

In total, data were available from 528 tasks for analysis of spoken and written disfluencies. All human-computer interactions were videotaped. Hardcopy transcripts also were created, with the subject's handwritten input captured automatically, and spoken input transcribed onto the printouts.

Transcript Coding- To summarize briefly, spontaneously occurring disfluencies and self-corrections were totaled for each subject and condition. The total number of disfluencies per condition then was converted to a rate per 100 words, and average disfluency rates were summarized as a function of condition and utterance length. Disfluencies were classified into the following types: (1) content self-corrections— task-content errors that were spontaneously corrected as the subject spoke or wrote, (2) false starts— alterations to the grammatical structure of an utterance that occurred spontaneously as the subject spoke or wrote, (3) verbatim repetitions— retracings or repetitions of a letter, phoneme, syllable, word, or phrase that occurred spontaneously as the subject spoke or wrote, (4) filled pauses— spontaneous nonlexical sounds that fill pauses in running speech, which have no analogue in writing, (5) self-corrected spellings and abbreviations— spontaneously corrected misspelled words or further specification of abbreviations, which occur in writing but have no analogue in speech.

2.2. Results

Figure 1 summarizes the percentage of all spoken and written disfluencies representing different categories during communication of verbal-temporal content (i.e., studies 1 and 2). However, when people communicated digits (i.e., study 3), disfluencies representing the different categories were distributed differently. Filled pauses dropped from 46% to 15.5% of all observed disfluencies. In contrast, content corrections of digits increased from 25% to 34%, repetitions increased from 21% to 31.5%, and false starts increased from 8% to 19% of all disfluencies. This drop in filled pauses and increase in other types of disfluency is most likely related to the much briefer utterance lengths observed during the computational-numeric tasks. Clearly, the relative distribution of different types of disfluency fluctuates with the content and structure of the information presented.

The overall baseline rate of spontaneous disfluencies and self-corrections was 1.33 per 100 words in the verbal-temporal simulations, or a total of 1.51 disfluencies per task set. The rate per condition ranged from an average of 0.78 per 100

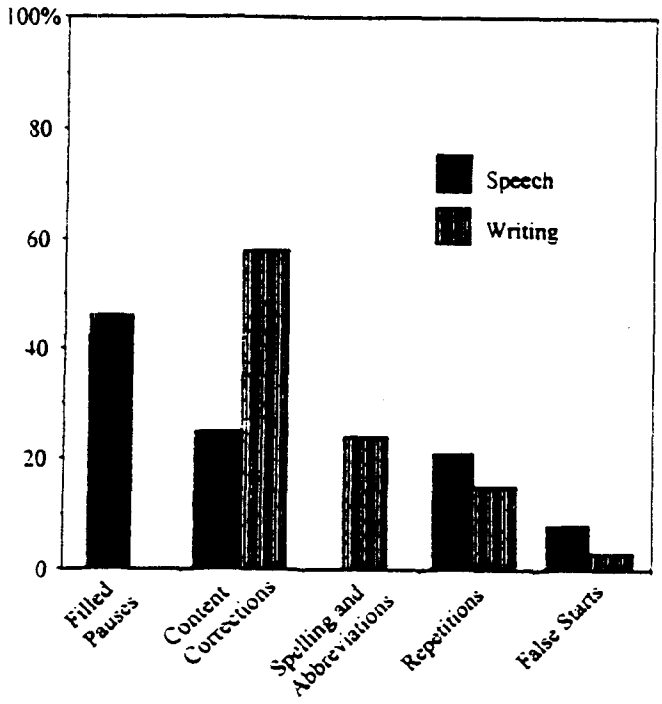

Figure 1. Percentage of all spoken and written disfluencies in different categories.

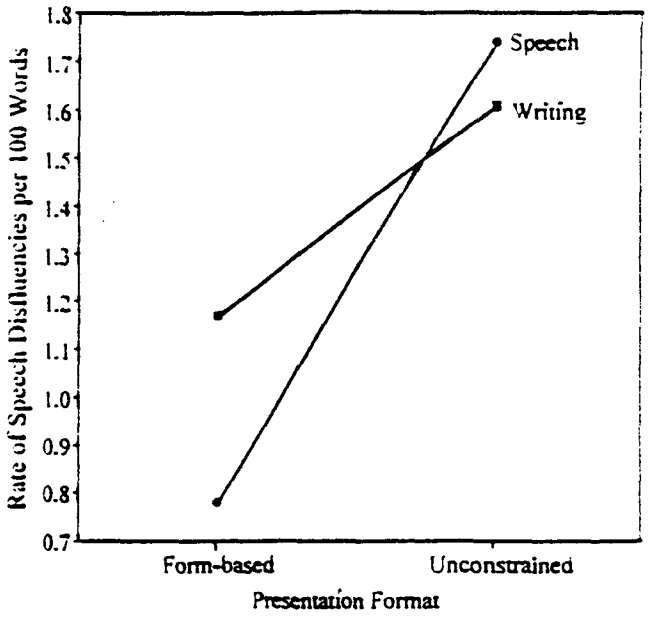

Figure 2. Increasing rate of spoken disfluencies per 100 words as a function of structure in presentation format.

words when speaking to a form, 1.17 when writing to a form, 1.61 during unconstrained writing, and a high of 1.74 during unconstrained speech. Figure 2 illustrates this rate of disfluencies as a function of mode and format.

Wilcoxon Signed Ranks tests revealed no significant modality difference in the rate of disfluent input, which averaged 1.26 per 100 words for speech and 1.39 for writing, T+ = 75 (N = 17), $z < 1$. However, the rate of disfluencies was 1.68 per 100 words in the unconstrained format, in comparison with a reduced .98 per 100 words during form-based interactions. Followup analyses revealed no significant difference in the disfluency rate between formats when people wrote, T+ = 64.5 (N = 14), $p > .20$. However, significantly increased disfluencies were evident in the unconstrained format compared to the form-based one when people spoke, T+ = 88 (N = 14), $p < .015$, one-tailed. This significant elevation was replicated for unconstrained speech that occurred during the free choice condition, T+ = 87 (N = 14), $p < .015$, one-tailed, which simulated a multimodal spoken exchange rather than a unimodal one.

A very similar pattern of disfluency rates per condition emerged when people communicated digits. In study 3, the baseline rate of spontaneous disfluencies averaged 1.37 per 100 words, with 0.87 when speaking to a form, 1.10 when writing to a form, 1.42 during unconstrained writing, and a high of 1.87 during unconstrained speech. Likewise, Wilcoxon Signed Ranks tests revealed no significant difference in the disfluency rate between formats when people wrote, T+ = 36.5 (N = 11), $p > .20$, although significantly increased disfluencies again were apparent in the unconstrained format compared to the form-based one when people spoke, T+ = 77 (N = 13), $p < .015$, one-tailed.

For studies 1 and 2, disfluency rates were examined further for specific utterances that were graduated in length from 1 to 18 words.[1] First, these analyses indicated that the average rate of disfluencies per 100 words increased as a function of utterance length for spoken disfluencies, although not for written ones. When the rate of spoken disfluencies was compared for short (1-6 words), medium (7-12 words), and long utterances (13-18 words), it increased from 0.66, to 2.14, to 3.80 disfluencies per 100 words, respectively. Statistical comparisons confirmed that these rates represented significant increases from short to medium sentences, t = 3.09 (df = 10), $p < .006$, one-tailed, and also from medium to long ones, t = 2.06 (df = 8), $p < .04$, one-tailed.

A regression analysis indicated that the strength of predictive association between utterance length and disfluency rate was $\rho^2_{XY} = .77$ (N = 16). That is, 77% of the variance in the rate of spoken disfluencies was predictable simply by knowing an utterance's specific length. The following simple linear model, illustrated in the scatterplot in Figure 3, summarizes this relation: $Y_{ij} = \mu Y + \beta_{Y \bullet X}(X_j - \mu X) + e_{ij}$, with a Y-axis constant coefficient of -0.32, and an X-axis beta coefficient representing utterance length of +0.26. These data indicate that the demands associated with planning and generating longer constructions lead to substantial elevations in the rate of disfluent speech.

To assess whether presentation format had an additional influence on spoken disfluency rates beyond that of utterance length, comparisons were made of disfluency rates occur-

[1]The average utterance length in study 3, in which people conveyed digits during scientific calculations and personal banking tasks, was too brief to permit a parallel analysis.

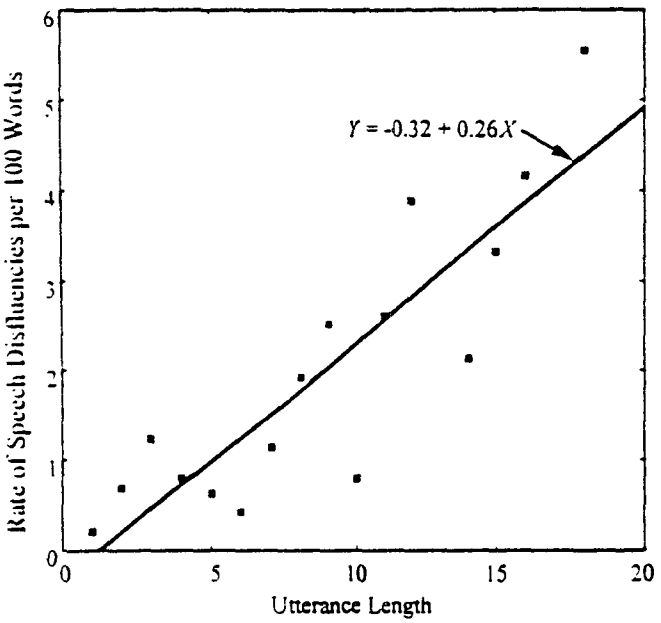

Figure 3. Linear regression model summarizing increasing rate of spoken disfluencies per 100 words as a function of utterance length.

ring in unconstrained and form-based utterances that were matched for length. These analyses revealed that the rate of spoken disfluencies also was significantly higher in the unconstrained format than in form-based speech, even with utterance length controlled, t (paired) = 2.42 (df = 5), $p < .03$, one-tailed. That is, independent of utterance length, lack of structure in the presentation format also was associated with elevated disfluency rates.

From a pragmatic viewpoint, it also is informative to compare the total number of disfluencies that would require processing during an application. Different design alternatives can be compared with respect to effective reduction of total disfluencies, which then would require neither processing nor repair. In studies 1 and 2, a comparison of the total number of spoken disfluencies revealed that people averaged 3.33 per task set when using the unconstrained format, which reduced to an average of 1.00 per task set when speaking to a form. That is, 70% of all disfluencies were eliminated by using a more structured form. Likewise, in study 3, the average number of disfluencies per subject per task set dropped from 1.75 in the unconstrained format to 0.72 in the structured one. In this simulation, a more structured presentation format successfully eliminated 59% of people's disfluencies as they spoke digits, in comparison with the same people completing the same tasks via an unconstrained format.

During post-experimental interviews, people reported their preference to interact with the two different presentation formats. Results indicated that approximately two-thirds of the subjects preferred using the more structured format. This 2-to-1 preference for the structured format replicated across both the verbal and numeric simulations.

3. EXPERIMENTS ON HUMAN-HUMAN SPEECH

This section reports on data that were analyzed to explore the degree of variability in disfluency rates among different types of human-human and human-computer spoken interaction, and to determine whether these two classes differ systematically.

3.1. Method

Data originally collected by the author and colleagues during two previous studies were reanalyzed to provide comparative information on human-human disfluency rates for the present research [1, 6, 7]. One study focused on telephone speech, providing data on both: (1) two-person telephone conversations, and (2) three-person interpreted telephone conversations, with a professional telephone interpreter intermediating. Methodological details of this study are provided elsewhere [7]. Essentially, within-subject data were collected from 12 native speakers while they participated in task-oriented dialogues about conference registration and travel arrangements. In the second study, also outlined elsewhere [1, 6], speech data were collected on task-oriented dialogues conducted in each of five different communication modalities. For the present comparison, data from two of these modalities were reanalyzed: (1) two-party face-to-face dialogues, and (2) single-party monologues into an audiotape machine. A between-subject design was used, in which 10 subjects described how to assemble a water pump. All four types of speech were reanalyzed from tape-recordings for the same categories of disfluency and self-correction as those coded during the simulation studies, and a rate of spoken disfluencies per 100 words was calculated.

3.2. Comparative Results

Table 1 summarizes the average speech disfluency rates for the four types of human-human and two types of human-computer interaction that were studied. Disfluency rates for each of the two types of human-computer speech are listed in Table 1 for verbal-temporal and computational-numeric content, respectively, and are corrected for number of syllables per word. All samples of human-human speech reflected substantially higher disfluency rates than human-computer speech, with the average rates for these categories confirmed to be significantly different, t = 5.59 (df = 38), $p < .0001$, one-tailed. Comparison of the average disfluency rate for human-computer speech with human monologues, the least discrepant of the human-human categories, also replicated this difference, t = 2.65 (df = 21), $p < .008$, one-tailed. The magnitude of this disparity ranged from 2-to-11-times higher disfluency rates for human-human as opposed to human-computer speech, depending on the categories compared. Further analyses indicated that the average disfluency rate was significantly higher during telephone speech than the other categories of human-human speech, t = 2.12 (df = 20), $p < .05$, two-tailed.

4. DISCUSSION

Spoken disfluencies are strikingly sensitive to the increased planning demands of generating progressively longer utter-

Type of Spoken Interaction	Disfluency Rate
Human-human speech:	
Two-person telephone call	**8.83**
Three-person interpreted telephone call	**6.25**
Two-person face-to-face dialogue	**5.50**
One-person noninteractive monologue	**3.60**
Human-computer speech:	
Unconstrained computer interaction	**1.74 / 1.87**
Structured computer interaction	**0.78 / 0.87**

Table 1: Spoken disfluency rates per 100 words for different types of human-human and simulated human-computer interaction.

ances. Of all the variance in spoken disfluencies in the first two studies, 77% was predictable simply by knowing an utterance's specific length. A linear model was provided, $Y = -0.32 + 0.26X$, to summarize the predicted rate of spoken disfluencies as a function of utterance length. Knowledge of utterance length alone, therefore, is a powerful predictor of speech disfluencies in human-computer interaction.

Spoken disfluencies also are influenced substantially by the presentation format used during human-computer interaction. An unconstrained format, which required the speaker to self-structure and plan to a greater degree, led speakers to produce over twice the rate of disfluencies as a more structured interaction. Furthermore, this format effect was replicated across unimodal and multimodal spoken input, and across qualitatively very different spoken content. Since the observed difference between formats occurred in samples matched for length, it is clear that presentation format and utterance length each exert an independent influence on spoken disfluency levels.

In these three studies, a substantial 60 to 70% of all spoken disfluencies were eliminated simply by using a more structured format. That is, selection of presentation format was remarkably effective at channeling a speaker's language to be less disfluent. In part, this was accomplished by reducing sentential planning demands during use of the structured formats – i.e., reducing the need for people to plan the content and order of information delivered (see Oviatt, forthcoming [5]). It also was accomplished in part by the relative brevity of people's sentences in the structured formats. The percentage of moderate to long sentences increased from 5% of all sentences during structured interactions to 20% during unconstrained speech— a 4-fold or 300% increase. In addition, whereas the average disfluency rate was only 0.66 for short sentences, this rate increased to 2.81 for sentences categorized as moderate or lengthy— a 326% increase. The structured format not only was effective at reducing disfluencies, it also was preferred by a factor of 2-to-1.

Wide variability can be expected in the disfluency rates typical of qualitatively different types of spoken language. Based on the six categories compared here, rates were found to vary by a magnitude of 2-to-11-fold between individual categories, with the highest rates occurring in telephone speech, and the lowest in human-computer interaction. This variability suggests that further categories of spoken language should be studied individually to evaluate how prone they may be to disfluencies, rather than assuming that the phenomenon is stable throughout spoken language. Future work exploring disfluency patterns during more complex multimodal exchanges will be of special interest.

Finally, future work needs to investigate other major human-computer interface features that may serve to decrease planning load on users, and to estimate how much impact they have on reducing disfluencies. Such information would permit proactive system design aimed at supporting more robust spoken language processing. For future applications in which an unconstrained format is preferred, or disfluencies and self-repairs otherwise are unavoidable, methods for correctly detecting and processing the ones that occur also will be required. To the extent that promising work on this topic can incorporate probabilistic information on the relative likelihood of a disfluency for a particular utterance (e.g., of length N), based on either the present or future predictive models, correct detection and judicious repair of actual disfluencies may become feasible.

5. ACKNOWLEDGMENTS

Sincere thanks to the generous people who volunteered to participate in this research as subjects. Thanks also to Michael Frank, Martin Fong, and John Dowding for programming the simulation environment, to Martin Fong and Dan Wilk for playing the role of the simulation assistant during testing, to Jeremy Gaston, Zak Zaidman, and Aaron Hallmark for careful preparation of transcripts, and to Jeremy Gaston, Zak Zaidman, Michelle Wang, and Erik Olsen for assistance with data analysis. Finally, thanks to Gary Dell and Phil Cohen for helpful manuscript comments.

References

1. P. R. Cohen. The pragmatics of referring and the modality of communication. *Computational Linguistics*, 1984, 10(2):97–146.
2. D. Hindle. Deterministic parsing of syntactic nonfluencies. In *Proceedings of the 21st. Annual Meeting of the ACL*, 1983, Cambridge, Mass. 123–128.
3. W. J. M. Levelt. *Speaking: From Intention to Articulation.* ACL/M.I.T. Press, Cambridge, Mass., 1989.
4. C. Nakatani and J. Hirschberg. A corpus-based study of repair cues in spontaneous speech. In *Journal of the Acoustical Society of America*, in press.
5. S. L. Oviatt. Predicting spoken disfluencies during human-computer interaction. Journal manuscript, in submission.
6. S. L. Oviatt and P. R. Cohen. Discourse structure and performance efficiency in interactive and noninteractive spoken modalities. *Computer Speech and Language*, 1991, 5(4):297–326.

7. S. L. Oviatt and P. R. Cohen. Spoken language in interpreted telephone dialogues. *Computer Speech and Language*, 1992, 6:277–302.

8. S. L. Oviatt, P. R. Cohen, M. W. Fong, and M. P. Frank. A rapid semi-automatic simulation technique for investigating interactive speech and handwriting. In *Proceedings of the 1992 ICSLP*, 1992, ed. by J. Ohala et al., University of Alberta, vol. 2, 1351-1354.

9. E. Shriberg, J. Bear, and J. Dowding. Automatic detection and correction of repairs in human-computer dialog. In *Proceedings of the DARPA Speech and Natural Language Workshop*, 1992, Morgan Kaufmann, Inc., San Mateo, CA, 23–26.