INTEGRATING MULTIPLE KNOWLEDGE SOURCES FOR DETECTION AND CORRECTION OF REPAIRS IN HUMAN-COMPUTER DIALOG*

John Bear, John Dowding, Elizabeth Shriberg†

SRI International
Menlo Park, California 94025

ABSTRACT

We have analyzed 607 sentences of spontaneous human-computer speech data containing repairs, drawn from a total corpus of 10,718 sentences. We present here criteria and techniques for automatically detecting the presence of a repair, its location, and making the appropriate correction. The criteria involve integration of knowledge from several sources: pattern matching, syntactic and semantic analysis, and acoustics.

INTRODUCTION

Spontaneous spoken language often includes speech that is not intended by the speaker to be part of the content of the utterance. This speech must be detected and deleted in order to correctly identify the intended meaning. The broad class of disfluencies encompasses a number of phenomena, including word fragments, interjections, filled pauses, restarts, and repairs. We are analyzing the repairs in a large subset (over ten thousand sentences) of spontaneous speech data collected for the DARPA Spoken Language Program.[1] We have categorized these disfluencies as to type and frequency, and are investigating methods for their automatic detection and correction. Here we report promising results on detection and correction of repairs by combining pattern matching, syntactic and semantic analysis, and acoustics. This paper extends work reported in an earlier paper (Shriberg et al., 1992a).

The problem of disfluent speech for language understanding systems has been noted but has received limited attention. Hindle (1983) attempts to delimit and correct repairs in spontaneous human-human dialog, based on transcripts containing an "edit signal," or external and reliable marker at the "expunction point," or point of interruption. Carbonell and Hayes (1983) briefly describe recovery strategies for broken-off and restarted utterances in textual input. Ward (1991) addresses repairs in spontaneous speech, but does not attempt to identify or correct them. Our approach is most similar to that of Hindle. It differs, however, in that we make no assumption about the existence of an explicit edit signal. As a reliable edit signal has yet to be found, we take it as our problem to find the site of the repair automatically.

It is the case, however, that cues to repair exist over a range of syllables. Research in speech production has shown that repairs tend to be marked prosodically (Levelt and Cutler, 1983) and there is perceptual evidence from work using lowpass-filtered speech that human listeners can detect the occurrence of a repair in the absence of segmental information (Lickley, 1991).

In the sections that follow, we describe in detail our corpus of spontaneous speech data and present an analysis of the repair phenomena observed. In addition, we describe ways in which pattern matching, syntactic and semantic analysis, and acoustic analysis can be helpful in detecting and correcting these repairs. We use pattern matching to determine an initial set of possible repairs; we then apply information from syntactic, semantic, and acoustic analyses to distinguish actual repairs from false positives.

*This research was supported by the Defense Advanced Research Projects Agency under Contract ONR N00014-90-C-0085 with the Office of Naval Research. It was also supported by a Grant, NSF IRI-8905249, from the National Science Foundation. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency of the U.S. Government, or of the National Science Foundation.

†Elizabeth Shriberg is also affiliated with the Department of Psychology at the University of California at Berkeley.

[1] DARPA is the Defense Advanced Research Projects Agency of the United States Government

THE CORPUS

The data we are analyzing were collected as part of DARPA's Spoken Language Systems project. The corpus contains digitized waveforms and transcriptions of a large number of sessions in which subjects made air travel plans using a computer. In the majority of sessions, data were collected in a Wizard of Oz setting, in which subjects were led to believe they were talking to a computer, but in which a human actually interpreted and responded to queries. In a small portion of the sessions, data were collected using SRI's Spoken Language System (Shriberg et al., 1992b), in which no human intervention was involved. Relevant to the current paper is the fact that although the speech was spontaneous, it was somewhat planned (subjects pressed a button to begin speaking to the system) and the transcribers who produced lexical transcriptions of the sessions were instructed to mark words they inferred were verbally deleted by the speaker with special symbols. For further description of the corpus, see MADCOW (1992).

NOTATION

In order to classify these repairs, and to facilitate communication among the authors, it was necessary to develop a notational system that would: (1) be relatively simple, (2) capture sufficient detail, and (3) describe the vast majority of repairs observed. Table 1 shows examples of the notation used, which is described fully in Bear et al. (1992).

The basic aspects of the notation include marking the interruption point, the extent of the repair, and relevant correspondences between words in the region. To mark the site of a repair, corresponding to Hindle's "edit signal" (Hindle, 1983), we use a vertical bar (|). To express the notion that words on one side of the repair correspond to words on the other, we use a combination of a letter plus a numerical index. The letter M indicates that two words match exactly. R indicates that the second of the two words was intended by the speaker to replace the first. The two words must be similar–either of the same lexical category, or morphological variants of the same base form (including contraction pairs like "I/I'd"). Any other word within a repair is notated with X. A hyphen affixed to a symbol indicates a word fragment. In addition, certain cue words, such as "sorry" or "oops" (marked with CR) as well as filled pauses (CF) are also labeled

I	want	fl-		flights	to	boston.	
		M_1-	|	M_1			
what		what	are	the	fares		
M_1	|	M_1					
show	me	flights		daily	flights		
		M_1	|	X	M_1		
I	want	a	flight	one	way	flight	
		M_1	|	X	X	M_1	
I	want	to	leave	depart	before	...	
		R_1	|	R_1			
what	are		what	are	the	fares	
M_1	M_2	|	M_1	M_2			
... fly	to	boston		from	boston		
	R_1	M_1	|	R_1	M_1		
... fly	from	boston		from	denver		
	M_1	R_1	|	M_1	R_1		
what	are		are	there	any	flights	
X	X	|					

Table 1: Examples of Notation

if they occur immediately before the site of a repair.

DISTRIBUTION

Of the 10,000 sentences in our corpus, 607 contained repairs. We found that 10% of sentences longer than nine words contained repairs. In contrast, Levelt (1983) reports a repair rate of 34% for human-human dialog. While the rates in this corpus are lower, they are still high enough to be significant. And, as system developers move toward more closely modeling human-human interaction, the percentage is likely to rise.

Although only 607 sentences contained deletions, some sentences contained more than one, for a total of 646 deletions. Table 2 gives the breakdown of deletions by length, where length is defined as the number of consecutive deleted words or word fragments. Most of the deletions

Deletion Length	Occurrences	Percentage
1	376	59%
2	154	24%
3	52	8%
4	25	4%
5	23	4%
6+	16	3%

Table 2: Distribution of Repairs by Length

Type	Pattern	Freq.
	Length 1 Repairs	
Fragments	$M_1-, R_1-, X-$	61%
Repeats	$M_1 \mid M_1$	16%
Insertions	$M_1 \mid X_1 \ldots X_i M_1$	7%
Replacement	$R_1 \mid R_1$	9%
Other	$X \mid X$	5%
	Length 2 Repairs	
Repeats	$M_1\ M_2 \mid M_1\ M_2$	28%
Replace 2nd	$M_1\ R_1 \mid M_1\ R_1$	27%
Insertions	$M_1 M_2 \mid M_1 X_1 \ldots X_i M_2$	19%
Replace 1st	$R_1\ M_1 \mid R_1\ M_1$	10%
Other	...|...	17%

Table 3: Distribution of Repairs by Type

Match Length	Fill Length 0	1	2	3
1	.82 (39)	.74 (65)	.69 (43)	.28 (39)
2	1.0 (10)	.83 (6)	.73 (11)	.00 (1)
3	1.0 (4)	.80 (5)	1.0 (2)	—
4	1.0 (2)	1.0 (1)	—	—

— indicates no observations

Table 4: Fill Length vs. Match Length

were fairly short; deletions of one or two words accounted for 82% of the data. We categorized the length 1 and length 2 repairs according to their transcriptions. The results are summarized in Table 3. For simplicity, in this table we have counted fragments (which always occurred as the second deleted word) as whole words. The overall rate of fragments for the length 2 repairs was 34%.

A major repair type involved matching strings of identical words. More than half (339 out of 436) of the nontrivial repairs (more editing necessary than deleting fragments and filled pauses) in the corpus were of this type. Table 4 shows the distributions of these repairs with respect to two parameters: the length in words of the matched string, and the number of words between the two matched strings. Numbers in parentheses indicate the number of occurrences, and probabilities represent the likelihood that the phrase was actually a repair and not a false positive. Two trends emerge from these data. First, the longer the matched string, the more likely the phrase was a repair. Second, the more words there were intervening between the matched strings, the less likely the phrase was a repair.

SIMPLE PATTERN MATCHING

We analyzed a subset of 607 sentences containing repairs and concluded that certain simple pattern-matching techniques could successfully detect a number of them. The pattern-matching component reported on here looks for identical sequences of words, and simple syntactic anomalies, such as "a the" or "to from."

Of the 406 sentences containing nontrivial repairs, the program successfully found 309. Of these it successfully corrected 177. There were 97 sentences that contained repairs which it did not find. In addition, out of the 10,517 sentence corpus ($10,718 - 201$ trivial), it incorrectly hypothesized that an additional 191 contained repairs. Thus of 10,517 sentences of varying lengths, it pulled out 500 as possibly containing a repair and missed 97 sentences actually containing a repair. Of the 500 that it proposed as containing a repair, 62% actually did and 38% did not. Of the 62% that had repairs, it made the appropriate correction for 57%.

These numbers show that although pattern matching is useful in identifying possible repairs, it is less successful at making appropriate corrections. This problem stems largely from the overlap of related patterns. Many sentences contain a subsequence of words that match not one but several patterns. For example the phrase "FLIGHT <word> FLIGHT" matches three different patterns:

show the	**flight**		earliest	**flight**
	M_1	|	X	M_1

show the	**flight**	time		**flight**	date
	M_1	R_1	|	M_1	R_1

show the	delta	**flight**		united	**flight**
	R_1	M_1	|	R_1	M_1

Each of these sentences is a false positive for the other two patterns. Despite these problems of overlap, pattern matching is useful in reducing the set of candidate sentences to be processed for repairs. Rather than applying detailed and possibly time-intensive analysis techniques to 10,000 sentences, we can increase efficiency by limiting ourselves to the 500 sentences selected by the pattern matcher, which has (at least on one measure) a 75% recall rate. The repair sites hypothesized by the pattern matcher constitute useful input for further processing based on other sources of information.

NATURAL LANGUAGE CONSTRAINTS

Here we describe two sets of experiments to measure the effectiveness of a natural language processing system in distinguishing repairs from false positives. One approach is based on parsing of whole sentences; the other is based on parsing localized word sequences identified as potential repairs. Both of these experiments rely on the pattern matcher to suggest potential repairs.

The syntactic and semantic components of the Gemini natural language processing system are used for both of these experiments. Gemini is an extensive reimplementation of the Core Language Engine (Alshawi et al., 1988). It includes modular syntactic and semantic components, integrated into an efficient all-paths bottom-up parser (Moore and Dowding, 1991). Gemini was trained on a 2,200-sentence subset of the full 10,718-sentence corpus. Since this subset excluded the unanswerable sentences, Gemini's coverage on the full corpus is only an estimated 70% for syntax, and 50% for semantics.[2]

Global Syntax and Semantics

In the first experiment, based on parsing complete sentences, Gemini was tested on a subset of the data that the pattern matcher returned as likely to contain a repair. We excluded all sentences that contained fragments, resulting in a dataset of 335 sentences, of which 179 contained repairs and 176 contained false positives. The approach was as follows: for each sentence, parsing was attempted. If parsing succeeded, the sentence was marked as a false positive. If parsing did not succeed, then pattern matching was used to detect possible repairs, and the edits associated with the repairs were made. Parsing was then reattempted. If parsing succeeded at this point, the sentence was marked as a repair. Otherwise, it was marked as **no opinion**.

Syntax Only		
	Marked as Repair	Marked as False Positive
Repairs	68 (96%)	56 (30%)
False Positives	3 (4%)	131 (70%)

Syntax and Semantics		
	Marked as Repair	Marked as False Positive
Repairs	64 (85%)	23 (20%)
False Positives	11 (15%)	90 (80%)

Table 5: Syntax and Semantics Results

Table 5 shows the results of these experiments. We ran them two ways: once using syntactic constraints alone and again using both syntactic and semantic constraints. As can be seen, Gemini is quite accurate at detecting a repair, although somewhat less accurate at detecting a false positive. Furthermore, in cases where Gemini detected a repair, it produced the intended correction in 62 out of 68 cases for syntax alone, and in 60 out of 64 cases using combined syntax and semantics. In both cases, a large number of sentences (29% for syntax, 50% for semantics) received a **no opinion** evaluation. The **no opinion** cases were evenly split between repairs and false positives in both tests.

The main points to be noted from Table 5 are that with syntax alone, the system is quite accurate in detecting repairs, and with syntax and semantics working together, it is accurate at detecting false positives. However, since the coverage of syntax and semantics will always be lower than

[2] Gemini's syntactic coverage of the 2,200-sentence dataset it was trained on (the set of annotated and answerable MADCOW queries) is approximately 91%, while its semantic coverage is approximately 77%. On a recent fair test, Gemini's syntactic coverage was 87% and semantic coverage was 71%.

the coverage of syntax alone, we cannot compare these rates directly.

Since multiple repairs and false positives can occur in the same sentence, the pattern matching process is constrained to prefer fewer repairs to more repairs, and shorter repairs to longer repairs. This is done to favor an analysis that deletes the fewest words from a sentence. It is often the case that more drastic repairs would result in a syntactically and semantically well-formed sentence, but not the sentence that the speaker intended. For instance, the sentence "show me <flights> daily flights to boston" could be repaired by deleting the words "flights daily," and would then yield a grammatical sentence, but in this case the speaker intended to delete only "flights."

Local Syntax and Semantics

In the second experiment we attempted to improve robustness by applying the parser to small substrings of the sentence. When analyzing long word strings, the parser is more likely to fail due to factors unrelated to the repair. For this experiment, the parser was using both syntax and semantics.

The phrases used for this experiment were the phrases found by the pattern matcher to contain matching strings of length one, with up to three intervening words. This set was selected because, as can be seen from Table 4, it constitutes a large subset of the data (186 such phrases). Furthermore, pattern matching alone contains insufficient information for reliably correcting these sentences.

The relevant substring is taken to be the phrase constituting the matched string plus intervening material plus the immediately preceding word. So far we have used only phrases where the grammatical category of the matched word was either noun or name (proper noun). For this test we specified a list of possible phrase types (NP, VP, PP, N, Name) that count as a successful parse. We intend to run other tests with other grammatical categories, but expect that these other categories could need a different heuristic for deciding which substring to parse, as well as a different set of acceptable phrase types.

Four candidate strings were derived from the original by making the three different possible edits, and also including the original string unchanged. Each of these strings was analyzed by the parser. When the original sequence did not parse, but one of edits resulted in a sequence that parsed, the original sequence was very unlikely to be a false positive (right for 34 of 35 cases). Furthermore, the edit that parsed was chosen to be the repaired string. When more than one of the edited strings parsed, the edit was chosen by preferring them in the following order: (1) $M_1|XM_1$, (2) $R_1M_1|R_1M_1$, (3) $M_1R_1|M_1R_1$. Of the 37 cases of repairs, the correct edit was found in 27 cases, while in 7 more an incorrect edit was found; in 3 cases **no opinion** was registered. While these numbers are quite promising, they may improve even more when information from syntax and semantics is combined with that from acoustics.

ACOUSTICS

A third source of information that can be helpful in detecting repairs is acoustics. In this section we describe first how prosodic information can help in distinguishing repairs from false positives for patterns involving matched words. Second, we report promising results from a preliminary study of cue words such as "no" and "well." And third, we discuss how acoustic information can aid in the detection of word fragments, which occur frequently and which pose difficulty for automatic speech recognition systems.

Acoustic features reported in the following analyses were obtained by listening to the sound files associated with each transcription, and by inspecting waveforms, pitch tracks, and spectrograms produced by the Entropic Waves software package.

Simple Patterns

While acoustics alone cannot tackle the problem of locating repairs, since any prosodic patterns found in repairs are likely to be found in fluent speech, acoustic information can be quite effective when combined with other sources of information, in particular with pattern matching.

In studying the ways in which acoustics might help distinguish repairs from false positives, we began by examining two patterns conducive to acoustic measurement and comparison. First, we focused on patterns in which there was only one matched word, and in which the two occurrences of that word were either adjacent or separated by only one word. Matched words allow for comparison of word duration; proximity helps avoid variability due to global intonation contours not associated with the patterns themselves. We present

here analyses for the $M_1|M_1$ ("flights for <one> one person") and $M_1|XM_1$ ("<flight> earliest flight") repairs, and their associated false positives ("u s air five one one," "a flight on flight number five one one," respectively).

In examining the $M_1|M_1$ repair pattern, we found that the strongest distinguishing cue between the repairs ($N = 20$) and the false positives ($N = 20$) was the interval between the offset of the first word and the onset of the second. False positives had a mean gap of 42 msec ($s.d. = 55.8$) as opposed to 380 msec ($s.d. = 200.4$) for repairs. A second difference found between the two groups was that, in the case of repairs, there was a statistically reliable reduction in duration for the second occurrence of M_1, with a mean difference of 53.4 msec. However because false positives showed no reliable difference for word duration, this was a much less useful predictor than gap duration. F0 of the matched words was not helpful in separating repairs from false positives; both groups showed a highly significant correlation for, and no significant difference between, the mean F0 of the matched words.

A different set of features was found to be useful in distinguishing repairs from false positives for the $M_1|XM_1$ pattern. A set of 12 repairs and 24 false positives was examined; the set of false positives for this analysis included only fluent cases (i.e., it did not include other types of repairs matching the pattern). Despite the small data set, some suggestive trends emerge. For example, for cases in which there was a pause (200 msec or greater) on only one side of the inserted word, the pause was never after the insertion (X) for the repairs, and rarely before the X in the false positives. A second distinguishing characteristic was the peak F0 value of X. For repairs, the inserted word was nearly always higher in F0 than the preceding M_1; for false positives, this increase in F0 was rarely observed. Table 6 shows the results of combining the acoustic constraints just described. As can be seen, such features in combination can be quite helpful in distinguishing repairs from false positives of this pattern. Future work will investigate the use of prosody in distinguishing the $M_1|XM_1$ repair not only from false positives, but also from other possible repairs having this pattern, i.e., $M_1R_1|M_1R_1$ and $R_1M_1|R_1M_1$.

	Pauses after X (only) and F0 of X less than F0 of 1st M_1	Pauses before X (only) and F0 of X greater than F0 of 1st M_1
Repairs	.00	.92
False Positives	.58	.00

Table 6: Combining Acoustic Characteristics of $M_1|XM_1$ Repairs

Cue Words

A second way in which acoustics can be helpful given the output of a pattern matcher is in determining whether or not potential cue words such as "no" are used as an editing expression (Hockett, 1967) as in "...flights <between> <boston> <and> <dallas> <no> between oakland and boston." False positives for these cases are instances in which the cue word functions in some other sense ("I want to leave boston no later than one p m."). Hirshberg and Litman (1987) have shown that cue words that function differently can be distinguished perceptually by listeners on the basis of prosody. Thus, we sought to determine whether acoustic analysis could help in deciding, when such words were present, whether or not they marked the interruption point of a repair.

In a preliminary study of the cue words "no" and "well," we compared 9 examples of these words at the site of a repair to 15 examples of the same words occurring in fluent speech. We found that these groups were quite distinguishable on the basis of simple prosodic features. Table 7 shows the percentage of repairs versus false positives characterized by a clear rise or fall in F0

	F0 rise	F0 fall	Lexical stress	Cont. speech
Repairs	.00	1.00	.00	.00
False Positives	.87	.00	.87	.73

Table 7: Acoustic Characteristics of Cue Words

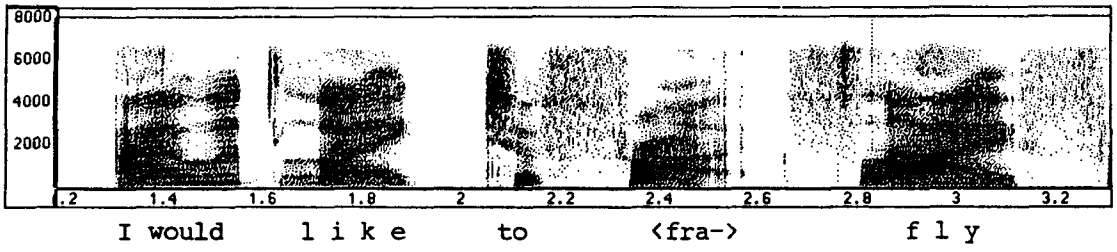

Figure 1: A glottalized fragment

(greater than 15 Hz), lexical stress (determined perceptually), and continuity of the speech immediately preceding and following the editing expression ("continuous" means there was no silent pause on either side of the cue word). As can be seen, at least for this limited data set, cue words marking repairs were quite distinguishable from those same words found in fluent strings on the basis of simple prosodic features.

Fragments

A third way in which acoustic knowledge can assist in detecting and correcting repairs is in the recognition of word fragments. As shown earlier, fragments are exceedingly common; they occurred in 366 of our 607 repairs. Fragments pose difficulty for state-of-the-art recognition systems because most recognizers are constrained to produce strings of actual words, rather than allowing partial words as output. Because so many repairs involve fragments, if fragments are not represented in the recognizer output, then information relevant to the processing of repairs is lost.

We found that often when a fragment had sufficient acoustic energy, one of two recognition errors occurred. Either the fragment was misrecognized as a complete word, or it caused a recognition error on a neighboring word. Therefore if recognizers were able to flag potential word fragments, this information could aid subsequent processing by indicating the higher likelihood that words in the region might require deletion. Fragments can also be useful in the detection of repairs requiring deletion of more than just the fragment. In approximately 40% of the sentences containing fragments in our data, the fragment occurred at the right edge of a longer repair. In a portion of these cases, for example,

"leaving at <seven> <fif-> eight thirty,"

the presence of the fragment is an especially important cue because there is nothing (e.g., no matched words) to cause the pattern matcher to hypothesize the presence of a repair.

We studied 50 fragments drawn at random from our total corpus of 366. The most reliable acoustic cue over the set was the presence of a silence following the fragment. In 49 out of 50 cases, there was a silence of greater than 60 msec; the average silence was 282 msec. Of the 50 fragments, 25 ended in a vowel, 13 contained a vowel and ended in a consonant, and 12 contained no vocalic portion.

It is likely that recognition of fragments of the first type, in which there is abrupt cessation of speech during a vowel, can be aided by looking for heavy glottalization at the end of the fragment. We coded fragments as glottalized if they showed irregular pitch pulses in their associated waveform, spectrogram, and pitch tracks. We found glottalization in 24 of the 25 vowel-final fragments in our data. An example of a glottalized fragment is shown in Figure 1.

Although it is true that glottalization occurs in fluent speech as well, it normally appears on unstressed, low F0 portions of a signal. The 24 glottalized fragments we examined however, were not at the bottom of the speaker's range, and most had considerable energy. Thus when combined with the feature of a following silence of at least 60 msec, glottalization on syllables with sufficient energy and not at the bottom of the speaker's

range, may prove a useful feature in recognizing fragments.

CONCLUSION

In summary, disfluencies occur at high enough rates in human-computer dialog to merit consideration. In contrast to earlier approaches, we have made it our goal to detect and correct repairs automatically, without assuming an explicit edit signal. Without such an edit signal, however, repairs are easily confused both with false positives and with other repairs. Preliminary results show that pattern matching is effective at detecting repairs without excessive overgeneration. Our syntactic/semantic approaches are quite accurate at detecting repairs and correcting them. Acoustics is a third source of information that can be tapped to provide evidence about the existence of a repair.

While none of these knowledge sources by itself is sufficient, we propose that by combining them, and possibly others, we can greatly enhance our ability to detect and correct repairs. As a next step, we intend to explore additional aspects of the syntax and semantics of repairs, analyze further acoustic patterns, and pursue the question of how best to integrate information from these multiple knowledge sources.

ACKNOWLEDGMENTS

We would like to thank Patti Price for her helpful comments on earlier drafts, as well as for her participation in the development of the notational system used. We would also like to thank Robin Lickley for his feedback on the acoustics section, Elizabeth Wade for assistance with the statistics, and Mark Gawron for work on the Gemini grammar.

REFERENCES

1. Alshawi, H, Carter, D., van Eijck, J., Moore, R. C., Moran, D. B., Pereira, F., Pulman, S., and A. Smith (1988) *Research Programme In Natural Language Processing: July 1988 Annual Report*, SRI International Tech Note, Cambridge, England.
2. Bear, J., Dowding, J., Price, P., and E. E. Shriberg (1992) "Labeling Conventions for Notating Grammatical Repairs in Speech," unpublished manuscript, to appear as an SRI Tech Note.
3. Hirschberg, J. and D. Litman (1987) "Now Let's Talk About Now: Identifying Cue Phrases Intonationally," *Proceedings of the ACL*, pp. 163-171.
4. Carbonell, J. and P. Hayes, P., (1983) "Recovery Strategies for Parsing Extragrammatical Language," *American Journal of Computational Linguistics*, Vol. 9, Numbers 3-4, pp. 123-146.
5. Hindle, D. (1983) "Deterministic Parsing of Syntactic Non-fluencies," *Proceedings of the ACL*, pp. 123-128.
6. Hockett, C. (1967) "Where the Tongue Slips, There Slip I," in *To Honor Roman Jakobson: Vol. 2*, The Hague: Mouton.
7. Levelt, W. (1983) "Monitoring and self-repair in speech," *Cognition*, Vol. 14, pp. 41-104.
8. Levelt, W., and A. Cutler (1983) "Prosodic Marking in Speech Repair," *Journal of Semantics*, Vol. 2, pp. 205-217.
9. Lickley, R., R. Shillcock, and E. Bard (1991) "Processing Disfluent Speech: How and when are disfluencies found?" *Proceedings of the Second European Conference on Speech Communication and Technology*, Vol. 3, pp. 1499-1502.
10. MADCOW (1992) "Multi-site Data Collection for a Spoken Language Corpus," *Proceedings of the DARPA Speech and Natural Language Workshop*, February 23-26, 1992.
11. Moore, R. and J. Dowding (1991) "Efficient Bottom-up Parsing," *Proceedings of the DARPA Speech and Natural Language Workshop*, February 19-22, 1991, pp. 200-203.
12. Shriberg, E., Bear, J., and Dowding, J. (1992 a) "Automatic Detection and Correction of Repairs in Human-Computer Dialog" *Proceedings of the DARPA Speech and Natural Language Workshop*, February 23-26, 1992.
13. Shriberg, E., Wade, E., and P. Price (1992 b) "Human-Machine Problem Solving Using Spoken Language Systems (SLS): Factors Affecting Performance and User Satisfaction," *Proceedings of the DARPA Speech and Natural Language Workshop*, February 23-26, 1992.
14. Ward, W. (1991) "Evaluation of the CMU ATIS System," *Proceedings of the DARPA Speech and Natural Language Workshop*, February 19-22, 1991, pp. 101-105.