SRI International

Technical Note 522 • February 22, 1993

# A System for Labeling Self-Repairs in Speech

*Prepared by:*

John Bear
Computer Scientist

John Dowding
Computer Scientist

Artificial Intelligence Center
Computing and Engineering Sciences Division

*and*

Elizabeth Shriberg
Research Linguist

Patti Price
Senior Computer Scientist

Speech Research and Technology Program
Computing and Engineering Sciences Division



333 Ravenswood Avenue • Menlo Park, CA 94025-3493 • (415) 326-6200 • FAX: (415) 326-5512 • Telex: 334486

# Report Documentation Page

*Form Approved*
*OMB No. 0704-0188*



| 1. REPORT DATE | 2. REPORT TYPE | 3. DATES COVERED |
|---|---|---|
| **22 FEB 1993** | | **00-02-1993 to 00-02-1993** |

| 4. TITLE AND SUBTITLE | 5a. CONTRACT NUMBER |
|---|---|
| **A System for Labeling Self-Repairs in Speech** | 5b. GRANT NUMBER |
| | 5c. PROGRAM ELEMENT NUMBER |
| 6. AUTHOR(S) | 5d. PROJECT NUMBER |
| | 5e. TASK NUMBER |
| | 5f. WORK UNIT NUMBER |
| 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **SRI International,333 Ravenswood Avenue,Menlo Park,CA,94025** | 8. PERFORMING ORGANIZATION REPORT NUMBER |
| 9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) | 10. SPONSOR/MONITOR'S ACRONYM(S) |
| | 11. SPONSOR/MONITOR'S REPORT NUMBER(S) |

12. DISTRIBUTION/AVAILABILITY STATEMENT
**Approved for public release; distribution unlimited**

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | | |
| **unclassified** | **unclassified** | **unclassified** | | **9** | |

# 1. INTRODUCTION

This document outlines a system for labeling self-repairs in spontaneous speech. The system marks the location and extent of a repair, as well as relevant words in the region of the repair. Together these labels determine the relationship between the "error" and the hypothesized "correction." The system is designed to be able to capture distinctions among different repair patterns while remaining easy to learn, apply, and integrate into existing transcription formats. Although the system was originally developed to aid our research on automatic detection and correction of repairs (Shriberg, Bear, & Dowding, 1992; Bear, Dowding & Shriberg, 1992), we hope that it may also prove useful for annotation of spontaneous speech data in related fields.

By "self-repairs" we refer to cases in which one or more words (or word fragments) must be disregarded in determining a speaker's "intended" utterance. Although one can never be sure exactly what a speaker intends, listeners can often reliably make such judgments. For example, given the utterance: "Show me flights from Boston from Denver to Dallas," most listeners would agree that "from Boston" should be disregarded, and that "Show me flights from Denver to Dallas" should be taken as the speaker's intended utterance. Often such judgments can be made on the basis of a transcription alone; listening to the utterance makes available prosodic cues which can greatly facilitate these judgments.

The definition of what constitutes a repair varies in the literature (e.g., Levelt, 1989; Blackmer & Mitton, 1991; Shriberg, Bear & Dowding, 1992). The present system is designed to annotate four types of phenomena:

- repairs involving replacements (as in the example above) or insertions
- repetitions of a string of one or more words ("Show me show me the flight...")
- fresh starts ("Show me the What are the flights...")
- cases involving a word fragment ("Show me the flights from Bos- Denver").

A number of other spontaneous speech phenomena are *not* of concern to this system. For example, filled pauses ("um," "uh") or other fillers ("well," "okay") are not marked unless they occur within an actual repair. This system also does not label silent pauses, uncorrected mispronunciations, repairs involving more than one speaker, and repairs involving a single speaker but in which the correction is a considerable distance (more than one sentence away) from the error.

In Sections 2 through 5, we describe our conventions for marking the site of a repair, and for marking words that distinguish among different repair patterns that we have found useful in our own research. All of the examples included actually occurred in our corpus (our data consisted of human-computer dialog in the air travel planning domain, see MADCOW, 1992). In Section 6, we provide a suggestion for how these labels may be integrated into existing transcription systems.

## 2. REPAIR SITE

We have adopted a vertical bar (|) notation for marking the site of the repair. The bar marks the resumption of fluent speech; it appears where Hindle (1983) puts his double-dash representing what he calls an "edit signal." In the examples that follow, we place labels on the line below the text.

Example:

List these in increasing in order of increasing fare

|

In the example just cited, the material following the bar ("in order of increasing fare") is a continuation of some of the material that preceded the bar ("List these"). In some repairs, however, the material after the bar constitutes the beginning of a new sentence. These repairs are often referred to as "fresh starts" (e.g., Levelt, 1989).

We mark fresh starts with a special kind of bar notation, so that they can be distinguished from other types of repairs. For fresh starts we use either a period-bar (.|) or a double-bar (||). The .| notation is used for cases in which there is a semantic relationship between the words preceding and following the bar; using this notation commits the labeler to labeling relationships between individual words on either side of the bar (as explained in Section 3). For instance, in the example below, "what is the cheapest" appears on both sides of the bar, and "fare" can be thought of as replacing the word fragment "fl-."

Example:

What is the cheapest fl- what is the cheapest fare

.|

For fresh starts in which a new idea is initiated, we use a double-bar (||) to mark the repair site. Use of the double bar means that the labeler is not committed to marking the relationships between words preceding and following the repair site. In the next example, there is a change in the semantics of the utterance, and although there are matching words on either side of the double-bar (i.e. "does this flight") it would be more difficult to annotate this utterance at the word level because of the presence of many unmatched words.

Example:

What time does this flight arrive where does this flight make a stop

||

Use of the .| versus || notation for repairs that constitute fresh starts is therefore a decision on the part of the labeler that is made by considering both the semantic relatedness of the material preceding and following the repair site, and the degree to which there are word-

by-word correspondences between these two portions of the utterance. A rule of thumb is to use the double-bar for any cases in which it would be difficult to determine word-by-word correspondences.

## 3. WORD-LEVEL LABELS

Individual words in the region of a repair are annotated with one of four possible labels: *M* (for "matching"), *R* (for "replacement"), *X* (for "insertion" or "deletion") or *C* (for "cue word").

### 3.1 Matching Words

Repairs often include repetitions of words or phrases. We note these words with the letter *M* (for match) plus a numerical index. Two occurrences of $M_i$ indicate a repetition of the same word.

Examples:

I want to go to | to Boston

$M_1$ | $M_1$

I'd like | I'd like to stop in Washington

$M_1$ $M_2$ | $M_1$ $M_2$

### 3.2 Replacements

In many cases we want to express the notion of one word replacing another. This we indicate with an *R* and a numerical index.

Examples:

to the city at Atlanta | in Atlanta using ground transportation

$R_1$ $M_1$ | $R_1$ $M_1$

What are the cheap | cheapest one way flights

$R_1$ | $R_1$

In the first example, "in" replaces "at." In both examples the relationship between the two elements constituting the replacement is one of shared grammatical category. In the second example, not only do the two words have the same grammatical category, they are also different morphological forms of the same word.

Finally, in the case of similar but different contractions as illustrated below, we have elected to use both *M* and *R* where appropriate, though clearly there are other reasonable alternatives. To represent the contracted forms, we use a caret (^) to link the associated labels.

Examples:

| All right | I'll | | I'm | interested in flight five eleven |
|---|---|---|---|---|
| | $M_1$^$R_1$ | \| | $M_1$^$R_1$ | |

| I'd | like | | I | would | like | breakfast served |
|---|---|---|---|---|---|---|
| $M_1$^$R_1$ | $M_2$ | .\| | $M_1$ | $R_1$ | $M_2$ | |

Note that these examples of contractions differ from the example in Section 3.1. Where the entire contraction is repeated, as in Section 3.1, we simply treat the word as a single unit and annotate it with $M_i$. When only part of the contraction is repeated, we break the contraction down and annotate each of the parts individually.

### 3.3 Insertions and Deletions

Words which figure in a repair (typically those which occur between the repair site and a word marked with M or R) and which are not themselves marked with an *M* or *R* are marked with an *X*. *X*s which occur to the left of a vertical bar indicate deletions; those that occur to the right indicate insertions.

Example:

| List | the | aircraft | | list | types | of | aircraft ... |
|---|---|---|---|---|---|---|---|
| $M_1$ | $X$ | $M_2$ | .\| | $M_1$ | $X$ | $X$ | $M_2$ |

This example illustrates a potential difficulty in deciding whether to use X or R. The best we can say here is that there is no obvious syntactic or semantic relationship between "the" and "types of." If we had the same grammatical category repeated, or nouns describing the same semantic class, such as "aircraft/airplanes," then we would use R instead of X.

Since we do not annotate a construction as as a repair unless some of the words were intended to be deleted, we never have an annotation such as " \| X " where nothing to the left of the bar is annotated. We have also never encountered a sentence which we felt ought to be labeled " X \| X ".

### 3.4 Cues

We label cue words and phrases (such as "I'm sorry") that occur immediately before the repair site with *C*. For cue phrases, each individual word is marked with a C.

Examples:

| from | Atlanta | back | to | Pittsburgh | I'm | sorry | | back | to | Denver |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $M_1$ | $M_2$ | $R_1$ | $C$ | $C$ | \| | $M_1$ | $M_2$ | $R_2$ |

| to | Atlanta | I | mean | sorry | Dallas | Fort | Worth | to | Atlanta |
|---|---|---|---|---|---|---|---|---|---|
| $M_1$ | $C$ | $C$ | $C$ | \| | $X$ | $X$ | $X$ | $X$ | $M_1$ |

## 4. LABELING NONWORDS

### 4.1 Filled Pauses

We differ from some researchers (e.g. Levelt, 1989; Blackmer & Mitton, 1991) in that we do not label any cases as repairs if simply a filled pause (typically "uh" or "um") is present. We do, however, label filled pauses that occur within a longer repair. These filled pauses are marked with *FP*.

Examples:

| Show | me | just | the | economy | class | fares | uh | | flights |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | $R_1$ | $FP$ | \| | $R_1$ |

| How | long | is | the | layover | in | Denver | uh | | in | Dallas |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | $M_1$ | $R_1$ | $FP$ | \| | $M_1$ | $R_1$ |

### 4.2 Word Fragments

Word fragments occur frequently immediately before a repair site. We indicate fragments by attaching a hyphen to the appropriate label. For example, if we want to indicate that a word is a replacement for a previously uttered word fragment, we add a hyphen to the $R_i$, as in the following example.

Example:

| on | July | fif- | | on | July | twentieth |
|---|---|---|---|---|---|---|
| $M_1$ | $M_2$ | $R_1$- | \| | $M_1$ | $M_2$ | $R_1$ |

In this example, the labeler's judgment is that "twentieth" is meant to replace the fragment "fif-" which was likely to have been the start of the word "fifteenth."

Previously we have used $M_i$ to indicate repetition of identical words and $R_i$ to indicate two words that are similar but not identical. In cases in which a word fragment like "phila-" is followed by a similar word like "Philadelphia"—that is, in which a labeler feels it is likely that the fragment was the beginning of what would have been a matched word—the label $M_i$- should be used.

Example:

| Also | list | fi- | flights | from | Atlanta | to | Boston... |
|---|---|---|---|---|---|---|---|
| | | $M_1$- | \| $M_1$ | | | | |

Fragments that seem to be neither matched nor replaced by a word to the right of the repair site are labeled with $X$- .

| Show | me | the | s- | flights | that | are | nonstop |
|---|---|---|---|---|---|---|---|
| | | | $X$- \| | | | | |

## 5. REPAIR EXTENT: HOW MUCH TO ANNOTATE

We have been tacitly following some important conventions about how far to the left and right of the repair site words should be labeled. Repairs whose repair site is marked by | or .| follow these conventions: To the left of the vertical bar, we always annotate all of the words to be "deleted" and only those. An $X$ under a word to the left of the bar means it was intended to be "deleted," hence we do not put an $X$ under a word to the left of the bar unless we think it is part of the error. The words to the right of the bar are only labelled if we believe they are part of the "correction." Typically the last word labeled in a correction will be labeled with either an $M_i$ or an $R_i$, and we do not label the rest of the words in the utterance after that with $X$.

Example:

| | I'd | like | I'd | like | to | stop | in | Washington |
|---|---|---|---|---|---|---|---|---|
| Correct: | $M_1$ | $M_2$ | .\| $M_1$ | $M_2$ | | | | |
| Incorrect: | $M_1$ | $M_2$ | .\| $M_1$ | $M_2$ | $X$ | $X$ | $X$ | $X$ |

| | What | is | the | earliest | flight | leaving | leaving | Boston |
|---|---|---|---|---|---|---|---|---|
| Correct: | | | | | | $M_1$ \| | $M_1$ | |
| Incorrect: | $X$ | $X$ | $X$ | $X$ | $X$ | $M_1$ \| | $M_1$ | |

For fresh starts whose repair site is labeled with ||, we label all words leftward from the repair site to the beginning of the sentence (they should always be either Xs, Cs, or FPs), but do not label any words to the right of the repair site.

Example:

Now could you What is the ground transportation available

*X X X* ||

## 6. LABELS IN TRANSCRIPTIONS

For purposes of exposition, we have in this document associated labels with transcriptions simply by placing the labels directly under the words they refer to. In practice, this can be awkward if the utterance is long and/or contains more than one repair, and in general it adds clutter to transcriptions. A simple convention that avoids these problems is to associate an identification number with each repair, and to indicate this number at the repair site in a transcript. The particular sequence of labels associated with the repair can then be listed in a separate file, under the identification number. Because no words are "skipped" when labeling leftward and rightward of the repair site, and since the location of the identification number in the transcript corresponds to the bar in the label sequence, the linking of labels to words in the transcript is completely determined.

Example:

I'd like to f- #001 go at nine #002 ten

001. $R_1$- | $R_1$

002. $R_1$ | $R_1$

Corrected sentence: I'd like to go at ten.

In the example above, we have used a pound sign (#) followed by a number as an identifier. The format and characters used in identifiers is arbitrary, however; identifiers should be determined individually by researchers to avoid any potential confusion with symbols they use in their own transcription system.

## 7. ACKNOWLEDGMENTS

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 Sceince 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.

## 8. BIBLIOGRAPHY

Bear, J., Dowding, J. & Shriberg, E. (1992). Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog. *Proceedings of the Association for Computational Linguistics*, pp. 56-63.

Blackmer, E. & Mitton, J. (1991). Theories of monitoring and the timing of repairs in spontaneous speech. *Cognition*, 39, pp. 173-194.

Hindle, D. (1983). Deterministic parsing of syntactic non-fluencies. *Proceedings of the Association for Computational Linguistics*, pp. 123-128.

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

MADCOW (1992). Multi-site data collection for a spoken language corpus. *Proceedings of the DARPA Speech and Natural Language Workshop*, February 23-26, 1992.

Shriberg, E. E., Bear, J. & Dowding, J. (1992). Automatic detection and correction of repairs in human-computer dialog. *Proceedings of the DARPA Speech and Natural Language Workshop*, February 23-26, 1992.