Preliminaries to a Theory of Speech Disfluencies

by

Elizabeth Ellen Shriberg

B.A. (Harvard University) 1987
M.A. (University of California at Berkeley) 1990

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Psychology

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA at BERKELEY

Committee in charge:

Professor Susan M. Ervin-Tripp, Chair
Professor John J. Ohala
Dr. Patti J. Price
Professor Herbert H. Clark

1994

Preliminaries to a Theory of Speech Disfluencies

Elizabeth Ellen Shriberg

Abstract

This thesis examines disfluencies (e.g., “um”, repeated words, and a variety of forms of self-repair) in the spontaneous speech of adult normal speakers of American English. Despite their prevalence, disfluencies have traditionally been viewed as irregular events and have received little attention. The goal of the thesis is to provide evidence that, on the contrary, disfluencies show remarkably regular trends in a number of dimensions. These regularities have consequences for models of human language production; they can also be exploited to improve performance in speech applications.

The method includes analysis of over 5000 hand-annotated disfluencies from a database (250,000 words) containing three different styles of spontaneous speech: task-oriented human-computer dialog, task-oriented human-human dialog, and human-human conversation on a prescribed topic. The approach is theory-neutral and strongly data-driven. The annotations correspond to observable characteristics (“features”) in the data, including: 1) the speech domain; 2) the speaker; 3) the sentence in which a disfluency occurs; 4) word-related characteristics of the disfluency; and 5) simple acoustic characteristics of the disfluency. A methodology is developed for representing these features in a database format, and an algorithm is provided for automatic disfluency type classification based on this representation.

Results show regular trends in disfluency rates by sentence length, by disfluency position, by presence of another disfluency in the same sentence, by disfluency type, and by combinations of these features both across and within speakers. Regularities are also found for word-related features of the disfluency, including the number of excised words, the rate of cut-off words, and the rate of editing phrases. Additional analyses describe characteristics of overlapping disfluencies and prosodic characteristics of the simplest disfluency types. Across analyses, data from the three different speech styles are compared; where relevant, simple parametric models are provided.

In sum, disfluencies show regularities in a variety of dimensions. These regularities can help guide and constrain models of spoken language production. In addition they can be modeled in applications to improve the automatic processing of spontaneous speech.

Contents

List of Figures

List of Tables

Acknowledgments

I am sincerely grateful to the many people who helped make this work possible.

First, I am indebted to the members of my committee, who provided essential advice and comments, kind encouragement, and who have also influenced me in ways that extend far beyond the thesis. Patti Price provided the opportunity for me to conduct this research at the Speech Technology and Research (STAR) laboratory at SRI International. She has consistently emphasized the importance of increased communication among psychologists, linguists and engineers; this goal is reflected in the thesis and is one I hope to continue to pursue. Herb Clark provided a second academic home at Stanford, and helped me to see speech communication from new perspectives during many lively discussions. John Ohala went to great lengths from great distances to comment on this work, and has provided continual guidance in all areas of speech science. Sue Ervin-Tripp advised me early on in graduate school to “do my own thing” and has helped me in every way to do so.

Four people deserve special mention for their extended influence on the content of the thesis. Robin Lickley, John Bear, John Dowding, and Mark Anderson contributed significantly through earlier collaborative work and in-depth discussions. I have also benefitted from correspondence with Peter Heeman, Willem Levelt, Christine Nakatani, Mari Ostendorf, Sharon Oviatt, Stefanie Shattuck-Hufnagel, and Kees van Deemter.

Many members of the STAR laboratory went out of their way to provide essential assistance on this project. In particular I would like to thank: Victor Abrash, Harry Bratt, Eric Jackson, Tom Kuhn, and Mitch Weintraub for technical assistance; Mike Cohen for discussions on speech recognition; Kate Hunicke-Smith for help in testing early versions of the labeling system; and Joani Ichiki and Louise Mason for administrative assistance.

In addition, I thank Mark Anderson for programming assistance, Robert Hasson and Randi Engle for advice on statistics, Jack Godfrey for information on the Switchboard corpus, and Gay Baldwin and Marc Swerts for valuable comments on the text.

Finally, I am grateful to the many friends who offered support throughout the process. Most importantly, I thank my family--Linda, Larry, Kathryn, and Janet--for their continued love and encouragement.

This research was supported by the Advanced Research Projects Agency under NSF Grant IRI-9314961, and by NSF Grant IRI-8905249. The views and conclusions contained in this document are those of the author and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Advanced Research Projects Agency or the National Science Foundation.

Chapter 1: Introduction

1.1 Motivation

Disfluencies such as “um”s and “uh”s, false starts, and word repetitions are prevalent in spontaneous spoken language, yet have received surprisingly little attention. By comparison, “slips of the tongue” (of which the “spoonerism” is the prototypical example) occur rarely but have been studied much more extensively. Attention to disfluencies (henceforth, “DFs”) is certain to gain importance, however, as research in speech-related fields focuses increasingly on spontaneous speech, rather than speech that is read or rehearsed. In psycholinguistics, research has expanded outward from laboratory speech toward an understanding of the complexities of everyday speech. In the rapidly-growing area of speech applications, spontaneous speech input is necessary for many tasks, minimizes user training and cognitive load, and provides a more pleasant user interface.

Although DFs have commonly been viewed as noisy or irregular events, results to be presented in this thesis indicate that, on the contrary, DFs show remarkable regularities in a number of dimensions. These regularities have consequences for cognitive models of speech communication. They can also be exploited to improve performance in speech applications.

1.2 Scope

This thesis is concerned with same-turn DFs in the spontaneous speech of adult normal speakers of American English. The DFs considered are cases in which a contiguous stretch of linguistic material must be deleted to arrive at the sequence the speaker “intended,” likely the one that would be uttered upon a request for repetition. Such cases include a wide class of phenomena commonly referred to as filled pauses, repetitions, false starts, repairs, and a variety of other terms. For example, the following utterance contains three DFs; in each case, a line is drawn through the contiguous material to be deleted:

~~they~~ they basically reviewed ~~oregon's plan or~~ the oregon plan toward ~~uh~~ nationalizing health care

The thesis does not consider DFs that occur across turn boundaries (whether by the same or a different speaker) or that span more than one sentence (as explained further in Chapter 4). Nor does it consider phenomena that do not involve deletion of linguistic material; therefore, unfilled pauses and uncorrected prosodic errors are beyond the scope of the present work. These phenomena, while important, proved difficult to identify reliably in the spontaneous speech styles examined.

Also playing no role in this work are non-linguistic intrusions such as laughter or coughing. Filled pauses, as will be argued in this work, *are* linguistic elements and therefore are included. By filled pause, only “um” and “uh” are implied. Other elements that have been grouped with filled pauses as “fillers” in some accounts--in particular discourse markers (“well”, “like”)--do not fall under the category of “disfluency” in the present work because they are arguably part of the speaker's intended utterance. References to DFs in research areas related to but outside the scope of the present work are provided in Chapter 2.

1.3 Goal

The goal of this thesis is to illuminate regularities in DF production. This thesis does *not* attempt to propose an all-encompassing theory of DFs. While such a theory is an ultimate goal, we must recognize that we are at an early stage in our understanding of disfluent phenomena. Although studies of DFs in a variety of disciplines (as reviewed in Chapter 2) have provided considerable information about specific aspects of DFs, we currently lack the knowledge necessary for an integrated theory.

To draw an analogy (albeit on a much grander scale): scientists could not attempt a Grand Unified Theory of physics without a century of work explaining subdivisions of physical phenomena (gravity, electromagnetism, relativity). Similarly, in the case of DFs, we cannot realistically attempt a unified theory of disfluency production without first having a strong understanding of appropriate subareas for disfluency modeling. This goal of this thesis is to illuminate relevant subareas for DF modeling by identifying regular trends in DF production.

1.4 Approach

Because this work predates a unified theory of DFs, the approach adopted is strongly data-driven. In all cases the effort has been to identify systematic variation *close* to the data; i.e. the observations are not phrased in terms of theoretical entities but rather in terms of

observable features of the data. Relevant subareas for independent modeling are then illuminated solely on the basis of regular trends in the data.

In keeping with the data-driven approach, the main focus is on summarizing the raw observable data, with less space devoted to interpretation and discussion. Where appropriate, reference is made to relevant previous work, and implications for psychological models and speech applications are suggested. However, these are kept general and brief, with the goal only of stimulating future discussion.

Ideally, the analyses should have long-lasting validity over a large number of domains and questions. To this end, the approach has been to use: 1) a large amount of speech data; and 2) theory-neutral *representations* of the data. The set of DFs studied is much larger in size than those of previous studies, and includes speech from many different speakers in three different speech settings; this quantity and variety tests coverage of the representational system and provides validity in statistical analyses. A theory-neutral system is developed to represent the DFs; this is not only appropriate given the current lack of a unified theory, it also allows for breadth of application to alternative or evolving theoretical perspectives.

1.5 Features Analyzed

DFs in this thesis are analyzed with respect to "features," where features are defined as observable characteristics in the data. A distinction is drawn between "features" and "variables." Variables are defined as underlying explanatory factors to which an observable trend can be attributed. Features are the surface manifestations of these variables. While the data-driven approach adopted dictates that the analyses herein are described in terms of features, the long-term goal is of course to discover the variables underlying these features, since it is in terms of variables rather than features that a unified theory must be phrased.

We expect that the study of DFs should cut across areas relevant to the study of fluent speech. Therefore, a variety of different features are examined. The complete set of features analyzed is discussed in Chapter 4.

As an arbitrary convenience for organizing the large number of features, as well as for relating the present studies to previous work, features are organized into a smaller number of logically orthogonal "dimensions." Note that while this organization of features is useful for discussion, no *theoretical* significance is attached to these groupings. Features are

epiphenomena of underlying explanatory variables, and the relationships among these explanatory variables are inherently unknown. Therefore, feature dimensions have no reality at the level of *interpretation*.

1.5.1 Domain features

Domain features include observable or manipulated aspects of the speech setting in which a DF is produced, such as the purpose of the discourse, or the communication mode. In this work, a single domain feature is examined, corresponding to the corpus of speech data from which the DF was drawn.

1.5.2 Speaker features

Speaker features are aspects associated with the individual who produces a DF. Speaker features examined include "speaker identity" in analyses of individual differences, speech rate, and speaker gender.

1.5.3 Sentence features

Sentence features make reference to the sentence in which a DF occurs, without making reference to the internal arrangement of words in the sentence. This is a somewhat contrived dimension, but useful for distinguishing the types of features examined in this thesis from syntactic features, which do not play a direct role in this work. Sentence features examined include the length of the sentence in which a DF occurs, the sentence position of a DF (sentence-initial or sentence-medial), and the presence of other DFs in the same sentence.

1.5.4 Acoustic features

Acoustic features are concrete measurements of properties of the speech signal, including fundamental frequency and duration. These features are studied for only the two simplest classes of DF phenomena, filled pauses and repetitions.

1.5.5 "Pattern" features

"Pattern" features are aspects relevant only to disfluent speech. Pattern features analyzed include the number of words deleted in a DF, the presence of editing phrases such as "oops," and the presence of cut-off words. In addition, a derived pattern feature,

disfluency “type,” is analyzed. This is considered a pattern feature because the method developed for type classification is based solely on other pattern features representing correspondences between replaced and replacing material in a DF, as explained in Chapter 4.

1.6 Chapter Summary

The goal of this thesis is to provide evidence that DFs show regular trends in a variety of dimensions. Because this work predates a unified theory of DFs, the approach is strongly data-driven, with a large set of DFs examined in terms of directly observable features. The analyses aim to illuminate observable features and feature combinations that show systematic trends in the data. Each regular trend delineates an autonomous area of inquiry that can be separately modeled. Modeling these trends is important for devising and testing theories of disfluency production, as well as for improving the automatic processing of spontaneous speech.

Chapter 2: Background

2.1 Chapter Overview

Although DFs have received relatively little attention with respect to *regularities*, they have nevertheless garnered interest from a diverse set of disciplines. This chapter reviews previous work, which serves as an important starting point for the thesis.

Section 2.2 provides references to studies of disfluent phenomena that are beyond the scope of the present work. Section 2.3 reviews previous classification systems, with a focus on methodological concerns. Section 2.4 reviews a sizable literature on DF production. As an organizational convenience, this section is subdivided according to the feature dimensions introduced in Chapter 1. Section 2.5 reviews studies of DF processing, by both humans and machines. Section 2.6 summarizes the main points from the review, and outlines four areas that remain to be addressed to enable cross-study, data-driven research on DFs; these areas guide the methodology and analyses of the thesis.

2.2 Related Work

A large body of work in related areas is beyond the scope of the present investigation and consequently not represented in this review. As set forth in Chapter 1, this thesis concerns: 1) same-turn DFs in the spontaneous speech of adult normal speakers of American English; and 2) only those disfluent phenomena in which the speaker's hypothesized intended utterance is arrived at by deleting a contiguous stretch of linguistic material.

Related areas falling outside the scope specified in point (1) include: cross-speaker or cross-turn DFs (Schegloff, 1979; Selting, 1988; Goodwin, 1986; Couper-Kuhlen, 1992); DFs or “dysfluencies” such as stuttering, in disordered speech (Wingate, 1984; Postma & Kolk, 1992); DFs in the speech of children (Hawkins, 1971; Kowal, O'Connell, & Sabin, 1975; Levin & Silverman, 1965); DFs in language acquisition (Peters & Menn, 1993); and DFs in a second language (Voss, 1979). While cross-linguistic work is not a focus, work in languages other than English is well represented to the extent that the studies appear to have general relevance across languages.

Topics falling outside the scope specified in point (2) above include: a very large literature on unfilled pauses (Goldman-Eisler, 1968; Boomer, 1965; O'Connell & Kowal, 1983; Butcher, 1981); grammatically incoherent utterances (Brown, 1980); “tails” or “right dislocations” (Geluykens, 1987), grammatical additions spoken as an afterthought (Carletta, Caley, & Isard, 1993) and discourse markers (Schiffrin, 1987; Redeker, 1991). Phenomena which are included as DFs in the thesis if corrected, but which are not described in particular detail, are speech errors (Fromkin, 1980; Nooteboom, 1980; Shattuck-Hufnagel, 1986; Baars, 1992); and prosodic errors (Cutler, 1983).

2.3 Classification Systems

Classification systems are discussed in three sections. Section 2.3.1 describes the structure of DFs and defines a standard set of terms for relevant DF regions. Section 2.3.2 briefly describes common aspects of previous systems. The focus is on Section 2.3.3, which discusses problematic issues concerning the use of previous systems in cross-study comparisons, as well as in data-driven, cross-corpus research. These concerns set the stage for the DF representation and classification systems developed in Chapter 4.

2.3.1 DF structure and terminology

Studies in linguistics (Hockett, 1967), conversation analysis (Schegloff, 1987; Goodwin, 1986), psycholinguistics (Levelt, 1983) and computational linguistics (Hindle, 1983) have independently noted that the surface form of the majority of same-turn DFs shows first the material that will ultimately be replaced, then optionally one or more editing phrases (such as “i'm sorry”), followed by the replacing material. These regions are contiguous, and removal of a contiguous stretch of material containing the error yields the hypothesized intended utterance.

Despite the agreement on structure, however, there has been a wide variety of terms used in referring to the relevant regions. Therefore, for this review, as well as for the remainder of the thesis, terms are standardized to the set shown in Figure 1. Terms are adapted from Levelt (1983), with some modifications.

The term “reparandum” (abbreviated “RM”) will refer to the entire stretch of speech to be deleted as opposed to only the altered element (“boston”) as originally used by Levelt. This is

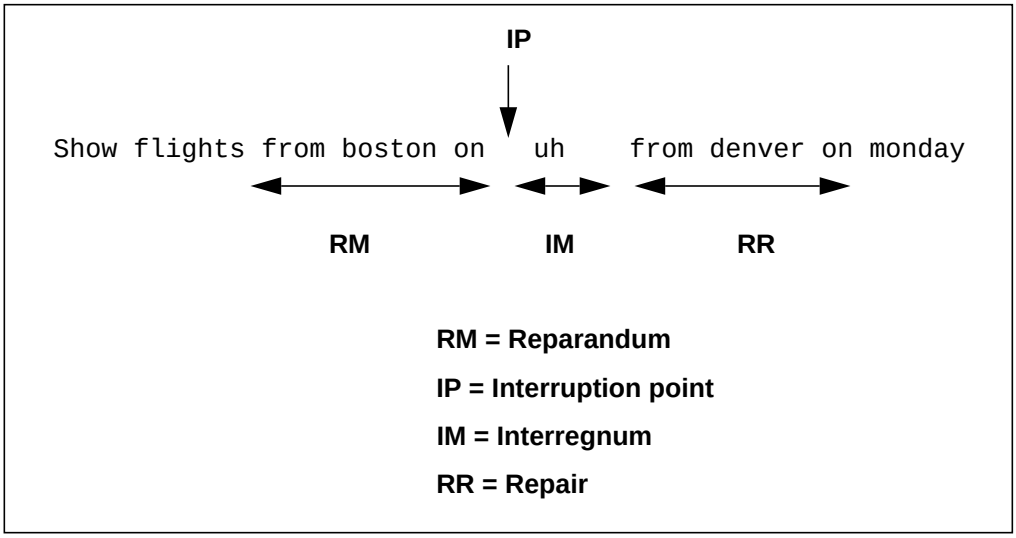


Figure 1. Terminology for DF Regions

because it is useful to have some term to refer to the entire deleted region, and because "reparandum" has come to be used this way by various authors.

The "interruption point" ("IP") is equivalent to the "moment of interruption" of Levelt and to the "cutoff" of Blackmer and Mitton (1991). Note that the location of the IP is based on the *surface* DF; detection of trouble by the speaker may have actually occurred earlier than this point in time.

The term "interregnum"[1] ("IM") is new and refers to the "editing phase" of Levelt, the "cut-off to repair" region of Blackmer and Mitton, and the "disfluency interval" of Nakatani and Hirschberg (1994). "Interregnum" is a more neutral term than "editing phase" (Levelt, 1983); it can be used to specify the temporal region from the end of the reparandum to the onset of the repair even if this region contains no editing term, and it does not imply an editing function for the speaker (e.g., an unfilled pause at the IM may be used for replanning without implying editing). Both the RM and any filled pauses or editing expressions in the IM are deleted to arrive at the intended utterance.

---

1. This term is intended to convey simply the break or gap between the RM and RR; it is not intended to invoke the political sense of "interregnum."

The “repair” (“RR”) region consists of the stretch of speech that “corresponds to” the material in the reparandum. The notion of correspondence is defined further in Chapter 4.

As pointed out by a number of authors (e.g., Maclay & Osgood, 1959; Schegloff, Jefferson, & Sacks, 1977; Levelt, 1983) DFs are not limited only to correction of *error*. In this work, DFs that do not show overt correction are analyzed as having the same surface structure as those that do. For example, a DF may contain only repeated words in the RM and RR:

```
Show me flights from boston  on         on   monday
                           |-RM-|-IM-|-RR-|
```

Or a DF may contain material only in the IM:

```
Show me flights from boston on        um       monday
                                |-RM-|-IM-|-RR-|
```

Both of these cases fit under the definition of DFs considered in this thesis, since both contain material in at least one of the two deleted regions (RM and IM). Further details are provided in Chapter 4.

2.3.2 Systems in general

The literature on classification systems shows great cross-study variation in systems. This is ignoring, for the moment, differences in terminology (see Section 2.3.3.1). Examples of different systems include: Mahl (1956), Maclay and Osgood (1959), Blankenship and Kay (1964), Deese (1980), Levelt (1983), Blackmer and Mitton (1991), Bear, Dowding, Shriberg, and Price (1993), Carletta, Caley, and Isard (1993), and Heeman and Allen (1994). It is probably not an exaggeration to say that there are as many different classification systems as there are studies involving classification. While some of this variation may be attributed to differences in the goals and perspectives of different studies, much of it does not seem to be necessitated by these factors.

Description of individual systems is beyond the scope of this review; for further reading see Lickley (1994), as well as Fox Tree (1993). It will suffice here to point out what the systems tend to have in common. Generally speaking, classification systems have been based on the nature of *correspondences in wording between material in the RM and material in the RR*.[2] Note that this aspect of DFs belongs in the "pattern features" dimension as defined in this thesis. Although the number of basic classes ranges from two to over a dozen classes, the majority of systems seem to minimally distinguish four classes: 1) correspondence, but no change from RM to RR (i.e. exact repetitions); 2) correspondence, with modification from RM to RR (e.g., replacements and insertions; 3) abandonment of material with no corresponding RR (often called "fresh starts"); and 4) no material in RM (e.g., filled pauses). However, many systems combine these pattern features with features grouped under other (logically independent) dimensions in this thesis, such as semantic or syntactic features.

2.3.3 Methodological concerns

This section discusses issues in DF classification from the perspective of the goal of the thesis. In working towards a unified theory of DFs, it will be crucial to be able to compare results across different studies, as well as to use a classification system appropriate to data-driven, pretheoretical analyses. In reviewing the literature on classification systems, a number of difficulties were encountered that made it clear a new system would need to be developed for the present work. Main issues are discussed in the subsections below.

2.3.3.1 Differences in terminology

A rather bewildering number of different terms have been used to refer to classes of DFs, including: abridged repair, aposiopesis, appropriateness repair, anacolouthon, correction, different repair, error repair, false start, filler, fresh start, filled hesitation, filled pause, full sentence restart, insertion, lexical repair, modification repair, production repair, repeat, repetition, reformulation, restart, sentence correction, sentence incompletion, sentence restart, stutter (an unfortunate term since it may be unrelated to stuttering in disordered speech), unfilled

---

[2.] There are, however, exceptions. For example, Bock's study of production problems (Bock, 1991) concerned levels of processing; accordingly, errors were classified according to the size of the production unit (e.g., sentence, phrase, word, morpheme, syllable, phoneme).

pause, word change and word omission. The proliferation in terms cannot be accounted for by differences in class membership; in many cases different terms have been used for what appear to be the same phenomena.

In addition, different cover terms (or terms used to refer to the set of DF types encountered in a study) have been used (e.g., “DFs,” “(self)repairs,” “(self)corrections,” “reformulations,” “restarts,” “edits,” and “hesitations”). To add confusion to matters, a cover term in one study is often equivalent to a term for a subclass in another study. For example, some authors use “hesitation” as a cover term (Maclay & Osgood, 1959; Blankenship & Kay, 1964) while others use it to refer only to DFs containing no changed material (e.g., Carletta, et al., 1993). Similarly, O'Shaughnessy (1994) uses “restart” as a cover term, but for Erman (1987), “restart” is one of “repetition,” “restart,” “insertion,” and “correction”. “Disfluency” is used as a cover term by Lickley (1994), Fox Tree (1993), and in this work, but is one of “speech error,” “self-repair,” and “disfluency” in Postma, Kolk, and Povel (1990).

2.3.3.2 Systems using hierarchical groupings

Differences across systems in hierarchical grouping of classes provide an additional source of variation, and are particularly problematic for cross-study comparisons when results for subordinate classes are not recoverable. Two common differences among systems concern superordinate categories for two frequently-occurring phenomena: 1) filled pauses (“um” and “uh”); and 2) repetitions (e.g., “the the”).

In a number of studies, the filled pauses “um” and “uh” have been treated like nonlinguistic elements, grouped with unfilled pauses and sometimes with nonlinguistic events such as grunts, laughter, and coughing. In other accounts, filled pauses have been grouped as “fillers” with a wide variety of linguistic elements, including coordinating conjunctions (Blankenship & Kay, 1964), discourse markers such as “you know” and “well” (Adams, 1982; Lutz & Mallard, 1986; Kowal, O’Connell & Sabin, 1975; James, 1972), and a variety of “waffling” behaviors (Carletta et al., 1993).

However, departures from these intuitive groupings also occur. For example, Heeman and Allen (1994) grouped filled pauses together with DFs containing a single word fragment in a class called “abridged” in their work on work on automatic processing of DFs. The abridged

class contains what would seem from a speech production point of view to be disparate categories, since filled pauses suggest mere planning pauses, whereas cut-off words are associated with repairs of error. However, from the automatic processing point of view, these cases share the characteristic that they are always deleted when found.

Repetitions (such as “the the” or “to the to the”) are another commonly-occurring class that has been grouped different ways in hierarchical classification systems. Some studies distinguish “covert” DFs, or DFs that do not involve a change of wording (hypothesized to reflect prearticulatory editing),[3] from “overt” DFs, or those which show a surface correction (e.g, Levelt, 1983). The same distinction appears to be conveyed by the “simple” and “complex” restarts classified by O’Shaughnessy (1994); in this case the terms correspond to the degree of difficulty the DFs pose for automatic detection algorithms. This approach groups repetitions with events like filled and even unfilled pauses. An alternative approach (e.g., Deese, 1980; Fox Tree, 1993) is to distinguish DFs having material in the RM from those having no material in the RM. This perspective groups repetitions with DFs involving changes of wording, such as replacement repairs.

It is worth noting that from the perspective of this work, the problem of hierarchical grouping is not limited to difficulties in cross-study comparisons. It is also a fundamental concern for pretheoretical research. Hierarchical systems based on surface intuitive groupings obscure discovery of regularities in pretheoretical analyses, because the data will be analyzed in only one way. In terms of underlying variables, the phenomena may actually be grouped in more than one way, depending on the question at hand.

2.3.3.3 Systems using semantic distinctions

Levelt (1983), and Levelt and Cutler (1983) analyzed descriptions of colored patterns, and drew a distinction between corrections of outright error (“error repairs”) such as:

```
go straight on to red -- er to yellow
```

[3] The implication that DFs showing no surface change in wording reflect preverbal editing of trouble needs to be reconciled, however, with a view of filled pauses and repetitions as hesitations reflecting incomplete plans (Clark & Clark, 1977). These would seem to be logically distinct situations, since detection of trouble implies that minimally something has been planned.

and elaborations or changes that do not contradict what was previously said (“appropriateness repairs”) such as:

```
draw a line up -- straight up to red
```

This semantic distinction appears to reflect important underlying differences in DFs. For example, the distinction was found to correlate with features such as the presence of mid-word cut-offs and prosodic marking (these features are discussed in more detail later in this chapter).

Methodologically, however, the distinction is often less than straightforward, because it would seem that the same surface DF may constitute either an error repair or an appropriateness repair depending on the situation.[4] For example, a repair such as:

```
move the block -- the green block
```

is a mere appropriateness repair if there is only one block available to be moved. If, however, there are many differently-colored blocks present, and the speaker intends that only the green block be moved, this repair seems better classified as an error repair--i.e. a correction of error in which too little information was supplied for the listener to determine the intended referent. Classification of such cases therefore requires sophisticated interpretation of contextual information.

Furthermore, classification is necessarily subjective when the discourse history of referents must be considered. For example, in a situation in which there are multiple, differently-colored blocks present, but the repair above occurs at a point in the discourse at which “the green block” is the most recently-mentioned block (i.e. the “default” or “active” block in the current state of the discourse) it is indeterminate whether the insertion of “green” in the repair above constitutes an error or an appropriateness repair. This classification hinges on the *speaker's* model of the listener's discourse model, i.e. whether or not the speaker believes the listener construed “green block” or simply “block.” (Note that the listener's *actual* model is

4. The author is grateful to Kees van Deemter and Willem Levelt for providing helpful insights on this issue.

irrelevant). The classification is necessarily subjective because in most cases it is not possible to determine the speaker's model of the listener (at least not from the discourse alone).

A related concern, particularly for cross-domain comparisons, is that the error/ appropriateness distinction seems to be sensitive to differences in mood (e.g., interrogative or imperative mood versus declarative mood) as well as by some dimension conveying the degree of importance attached to a correction. For example, in the context of a query for information from a large database, the utterance:

```
show the flights -- the early morning flights to boston
```

may constitute an error repair if the unrepaired query would have returned an objectionable amount of excess information that adversely affected the user’s ability to perform the task at hand.

Thus, although the semantic distinction between error and appropriateness repairs is undoubtedly an important one, it is (at least at present) a difficult distinction to apply objectively both within and across speech domains. Therefore, semantic features are not among the features annotated in the present work.

2.3.3.4 Systems producing nonorthogonal classes

Many systems have distinguished a class of DFs in which the speaker backs up to the beginning of the sentence and starts afresh. For example:

```
a) show me the -- which morning flights go to boston
```

This class has been referred to by a variety of terms, including “false start”, “sentence incompletion” (both referring to the RM); and “fresh start” or “sentence correction” (referring to the RR).

The difficulty with this classification from the perspective of the current work is that backing up to a sentence boundary appears to be a feature independent of the type of change effected from RM to RR. For example, in cases (b)-(d) below, the speaker has retraced to the

beginning of the sentence, but these cases also correspond to different *types* of changes (a repetition, an insertion, and a substitution, respectively; see Chapter 4).

```
b) show the -- show the morning flights to boston

c) show the flights -- show the morning flights to boston

d) show the morning -- show the evening flights to boston
```

If retracing to a sentence boundary is the criterion for the fresh start class, then the examples above will be classified as fresh starts, while any counterparts differing only in amount of retracing will fall into classes based on the change from RM to RR. For example, cases (e) and (f) below will fall into a separate class from case (c) above.

```
e) show the flights -- the morning flights to boston

f) show the flights -- morning flights to boston
```

Although the fresh start class has typically been described on the basis of the sentence-initial RR feature, it may be intended to capture those cases in which there is no correspondence (i.e. no repeated words, replaced material or substituted material) between the deleted material and the new sentence, as in example (a). However, such cases of noncorrespondence can also be observed within-sentence, as in example (g):

```
g) show the flights that -- to boston
```

Here, “that” is abandoned, but it is not followed by the beginning of a new sentence. In the classification system developed in Chapter 4, examples (a) and (g) are both treated simply as “deletion” DFs, or cases in which the material in the RM bears no correspondence to following material. Retracing to sentence beginning is treated as a classification-independent feature.

Previous classification systems have included other classes that present essentially the same problem. For example, Blankenship and Kay (1964) included a separate class for any DFs containing a word fragment (i.e. cases in which speech is cut off mid-word; see Section 2.4.6.1 below, and Chapter 3). Like sentence position, word fragmentation is a feature that can be observed across DFs differing in correspondence relationships (see Chapter 4), and thus should be treated independently from correspondence in a classification system.

2.3.3.5 Systems allowing incomplete coverage

Under the topic of coverage, three issues are worthy of mention: 1) handling of ambiguous DFs; 2) handling of overlapping DFs; and 3) presence of a "garbage" category.

Ambiguous DFs are those for which class membership is not clear because the information needed to discriminate among alternative analyses is not available in the surface DF. Further definition requires reference to the particular classification system used. The issue of ambiguity has received little attention in the literature. It is not clear whether such cases have been classified according to a "first guess," or placed in multiple classes, or put in an "unclassifiable" class (see below).

Overlapping DFs involve multiple interruption points in close proximity, such that some words play a role in more than one DF. Further definition is provided in Chapter 4, where overlapping (or "complex") DFs are described within the framework of the PLS. Previous systems have not explicitly discussed the classification of such cases.

Finally, many systems have included a "garbage" class for DFs that do not fit into any existing class in the system. From the perspective of a pretheoretical classification system it is undesirable to allow such a class, because full coverage is the best test for a good fit of the system to the observed phenomena.

2.4 Studies of Disfluency Production

The focus turns now to descriptive studies of DFs. This section reviews a large number of studies on rates and characteristics of DFs. For convenience, the studies are organized by feature dimension, as introduced in Chapter 1.

2.4.1 Domain features

Although a variety of speech domains have been examined in previous studies, cross-study comparisons are difficult to interpret due to differences in the classification of phenomena, as well as differences in methods for measuring DF rates. Therefore, for attribution of rates of DFs to specific features of a domain, it is necessary to refer to controlled studies. These studies are reviewed in two sections: 1) rate of DFs overall; and 2) rate of filled pauses.

2.4.1.1 Rate of DFs overall

The rate of DFs has been found to be affected by cognitive variables, such as task complexity, amount of preplanning or rehearsal, and degree of structure imposed on the discourse. In classic work, Goldman-Eisler (1968) found that subjects produced fewer DFs when describing a cartoon than interpreting it. Goldman-Eisler also found that DFs were reduced with preplanning or rehearsal. Additional early studies of the effect of similar factors on DF rates are reviewed in Rochester (1973). More recently, in a study of human-computer dialog, Oviatt (1994) found a lower rate of DFs when the task involved a constrained format (prompting from the computer for specific information) than when it involved an unconstrained format (no computer prompting).

The ability to preplan speech is a factor to consider in explaining the much lower rate of DFs in human-computer dialog than in human-human dialog across studies (e.g., Levelt, 1983; Blackmer & Mitton, 1991; Shriberg, Bear, & Dowding, 1992; Bear, Dowding, & Shriberg, 1992; Lickley, 1994). Many factors distinguish any particular human-computer dialog task from any particular human-human dialog task; much work remains to be done to determine which factors are relevant to DF production.

One clear difference between certain speech applications and human-human discourse, however, is the ability to completely control the onset of a turn. The presence of a “push-to-talk” device in a speech application allows speakers to preplan their utterance, which as mentioned above, reduces DF production. Two recent studies, however, suggest that the ability to control the onset of a turn cannot completely account for the lower rate of DFs in human-computer dialog.

In one study (Moore & Browning, 1992), a push-to-talk mechanism was not available in either of two conditions (human or computer listener). British English speakers called a route-planning service, and in all cases spoke with a human operator. However, in the human-computer condition, the human's voice was altered to make it sound robotic. Speakers made DFs in nearly 40% of turns to the unaltered voice, as opposed to under 20% to the altered voice. A second study took the opposite approach. Suhm et al. (1994) found a higher rate of DFs in a human-human scheduling task than in human-computer air travel planning dialog when a push-to-talk mechanism was used in both conditions. Thus, although the ability to control the onset of a turn is likely to play a role related to preplanning, additional factors must be considered in explaining differences in DF rates between human-human and human-computer dialog.

DF rates have also been found to be related to a perceived need to speak carefully. Broen and Siegel (1972) found that subjects were least disfluent in situations they ranked most highly for the need to speak carefully. Subjects made fewer DFs while speaking in front of an imagined audience (ranked highest for the need to speak carefully) than when simply speaking in front of a television set, and made the most DFs in casual conversation (ranked lowest for need to speak carefully). Similarly, Deese (1980) found that “type II” DFs (DFs other than unfilled or filled pauses) were less frequent when speaking before a general audience than before an audience familiar with the discussion topic.

A perceived need to speak clearly has also been found to affect DF rates in human-computer dialog. Shriberg, Wade, and Price (1992) and Wade, Shriberg, and Price (1992) found that subjects making air travel plans by talking to a computer sometimes used a hyperarticulate, or over-enunciated speech style. The hyperarticulate style was associated with a lower rate of DFs. The tendency to hyperarticulate did not appear to be a characteristic of speakers, since it could be manipulated by instructions; subjects who were told they should speak naturally were less likely to hyperarticulate, and tended to produce more DFs than subjects who were given no instructions regarding speech style.

Consistent with these results are studies showing that subjects can suppress their production of DFs if instructed or incentivized to do so (Siegel, Lenske, & Broen, 1969). It should be pointed out, however, that it is not clear from these studies what effect the suppression of DFs has on the speech produced (e.g., in terms of complexity, efficiency, or rate of speech).

2.4.1.2 Rate of filled pauses

Predictors of filled-pause rates remain somewhat mysterious. Unlike the case for other types of DFs, filled-pause rates have been found to be remarkably stable over increased task complexity, as well as over induced anxiety (Goldman-Eisler, 1961; Kasl & Mahl, 1965; and additional studies reviewed in Rochester, 1973).

Maclay and Osgood (1959) suggest that filled pauses reflect the combined effects of a need to pause and a socially-determined need to "hold the floor." Under this view, filled-pause rates should vary depending on the degree of pressure to maintain the floor. Unfortunately, an experiment by Lalljee and Cook (1969) found no effect on filled pause rates when a silent versus interrupting confederate was present with the subject. However, the very low rate of interruption (three interrupts in two minutes) is a concern for the interpretation of these findings. Support for a floor-holding function of filled pauses that occur at turn beginning comes from work in Italian: d'Urso and Zammuner (1990) had subjects ask a confederate questions and found that subjects tolerated a longer silence from the confederate before jumping in if the confederate had uttered "mm" one second after hearing the question.

Filled pauses have been noted to increase when gestural cues are absent. Christenfeld, Schacter, and Bilous (1991) found that filled pauses and gestures tended not to cooccur, suggesting they may serve similar functions.

Additional studies show that filled pauses can be employed as a conversation resource to convey uncertainty (Good & Butterworth, 1980; Smith & Clark, 1993). They may also be used to secure a listener's attention at turn beginning (Goodwin, 1981; Schegloff, 1987; Sadanobu & Takubo, 1993).

2.4.2 Speaker features

A number of authors have noted that different speakers produce different rates of DFs overall, as well as different rates of specific types of DFs (Maclay & Osgood, 1959; Deese, 1980; Allwood, Nivre, & Ahlsen, 1989; Lickley, 1994.)

2.4.2.1 Speech rate

Maclay and Osgood (1959) reported on individual differences for five male speakers recorded at a conference. A significant negative correlation was found between mean speech rate (measured as total words per minute, including pauses and words deleted in DFs) and rate of DFs (DFs per 100 words). The authors concluded that *faster* speakers are "better" speakers in that they make fewer overall DFs. However, this measure of DFs included rates of unfilled pauses.

The same study also examined individual differences in rates of specific DF types. DFs were classified into four classes: unfilled pause, filled pause, repeat, and "false start" (all others). Two of the five subjects showed large negative correlations between the rate of false starts and the rate of unfilled pauses, whereas a third subject showed a positive correlation between these types. However, all five subjects showed fairly stable distributions of types over different utterances.

2.4.2.2 Gender

Results from a recent study suggest that DF rates may be correlated with gender. Lickley (1994) collected a large amount of conversational speech for each of six British English speakers (three male, three female), and found that the female speakers were less disfluent overall (as measured by total DFs per word) than the male speakers.

2.4.3 Sentence features

2.4.3.1 Sentence length

Oviatt (1994) found that the rate of DFs (DFs per 100 words) in human-computer dialog increased roughly linearly with the number of words in an utterance, and suggests that one way to reduce DFs in speech applications would be to find ways to reduce the average utterance length.

2.4.4 Syntactic features

Although this thesis does not examine syntax, it is nevertheless useful to provide a brief summary of past work in this area. In particular, the "sentence position" feature used in the

present work (see Chapter 4) captures some degree of phrasing information, since onsets of sentences are typically also onsets of phrases.

2.4.4.1 Distribution of DFs with respect to grammatical units

Studies of the distribution of DFs with respect to syntactic units were prevalent from the late 1950s to early 1970s, and represent the earliest literature on normal DFs in spontaneous American English speech. The details of this large literature, which includes studies of unfilled pausing, are beyond the scope of this review, but are summarized in reviews by Rochester (1973), and Butcher (1981).

To summarize the main points: DFs gained attention as a means by which to test the Markovian model of speech production of the early 1950s. Lounsbury (1954) proposed that hesitations (loosely construed as any DF) should occur at points of high uncertainty or low "habit strength" between words. Experiments by Goldman-Eisler (1968) and others showed that words following hesitations in original speech were more difficult for subjects to guess. Thus, in the framework of this simple model, the word was posited as the "unit of encoding" and hesitations were viewed as reflecting the process of lexical decision-making for the following word.

Results from analyses of DFs in spontaneous speech corpora, however, revealed that the lexical decision hypothesis could not completely account for the pattern of distribution of DFs. Maclay and Osgood (1959) and Blankenship and Kay (1964) found that over 40% of DFs occurred before function words; this could not be accounted for by an uncertainty model, since function words have relatively high transition probabilities. The conclusion of these and similar studies was that DFs must also reflect planning of units larger than the word. In particular, DFs were found to occur early in grammatical phrases and decrease in frequency toward the end of such units, consistent with the hypothesis that most planning must be done early in the unit and less decision-making is required as more of the content of a unit is fixed (Maclay & Osgood 1959, Beattie, 1979).

2.4.4.2 Correlated semantic and discourse variables

It should be kept in mind that onsets of grammatical units are correlated with onsets of semantic and discourse units, both of which have been associated with DF production. For example, Taylor (1969) found greater latency in time to produce sentences using topic words varying in difficulty. The longer latency corresponded to more difficult or abstract words, not to number of constituents in the utterance, suggesting the latency was associated with semantic, not grammatical planning.

In addition, DFs may serve a discourse function. They may be used at turn beginnings to secure the attention of a listener before speaking (e.g., Goodwin, 1981). A discourse function has also been suggested for the somewhat opposite situation, in which a speaker must begin but has not yet formulated an utterance: DFs may be used in such situations as a provisional start that can be repaired or incrementally elaborated upon later if necessary (Carletta, et al., 1993; Okada & Otsuka, 1993).

2.4.5 Acoustic features

2.4.5.1 "Edit signal" at IP

Hockett (1967) proposed that DFs are marked by a sharp glottal cut-off at the IP. However, inspection of digitized DFs (e.g., Bear et al., 1992) has revealed no such localized and reliable signal. Similarly, evidence from perceptual experiments with gated or lowpass filtered speech (to eliminate semantic and syntactic cues) reveals that listeners cannot detect a DF until they have heard at least part of the first word in the RR, i.e. consistently *later* than the location of the proposed signal (Lickley, 1994.)

Glottalization at the IP has been noted to occur for some percentage of DFs, particularly for vowel-final fragments (Bear et al., 1992). However, the rate of glottalization differs across studies (Nakatani & Hirschberg, 1993; Lickley, 1994), and is by no means a reliable cue to the IP of DFs, since it frequently occurs elsewhere, in fluent speech (Bear et al., 1992).

2.4.5.2 Unfilled pause in IM

The duration of the IM is of particular theoretical relevance to models of speech production. Levelt (1983) proposed a "Main Interruption Rule," which states that speech is cut off as soon as trouble is detected by the speaker. This rule is based on the observation that speakers may cut themselves off at any point, even within a word, and that within-words cut-offs are not constrained to preserve morphological or syllable boundaries (Levelt, 1983; Nooteboom, 1980). Under Levelt's framework of speech production (Levelt, 1989), self-interruption is associated with a halting of the speech production process at all levels. Therefore, some minimum time is needed after the speech is cut off in order to plan the repair.

An unfilled pause in the IM has also been suggested as an important feature for speech perception. Howell and Young (1991) used synthesis to alter prosodic parameters of DFs, and found that a pause at the IM may help listeners discern the onset of the repair region. In addition this feature has played a role in approaches for automatic detection and correction of DFs in the ATIS[5] corpus (e.g. Bear et al., 1992; O'Shaughnessy, 1993; Nakatani & Hirschberg, 1994).

However, Blackmer and Mitton (1991) carefully analyzed speech to a Canadian radio call-in show, and found that over 30% of the DFs had IM durations below 100 milliseconds. This is less than half the time typically viewed as necessary for a planning pause (Goldman-Eisler, 1968). Furthermore, a number of DFs showed no pause at all. This result implies that replanning may occur even before speech is cut off in a DF.

2.4.5.3 Prosodic marking in RR

Levelt and Cutler (1983) judged "prosodic marking"--an increase in fundamental frequency (F0), duration, or amplitude--in the repair region of DFs from a pattern description task (Levelt, 1983). The authors found that marking occurred for roughly half of the repairs involving *error*, but only about 20% of the repairs involving mere *elaboration* (see Section 2.3.3.3 for a discussion of semantic aspects of DFs.) This suggests it may be more important to call attention to outright error than to inappropriateness.

[5.] See Chapter 3 for a description of this corpus.

However, these rates, while significantly different, reflected only tendencies, since a non-negligible number of appropriateness repairs were marked and nearly half of the error repairs were not marked. Similarly, Howell and Young (1991) examined hand-marked stress levels in corpus of British conversational speech and found that the RR had higher stress than the RM only 24% of the time. It was not clear in either study why only some error repairs were marked, although individual differences (tendencies to mark all DFs or no DFs) were partially responsible in both studies.

In perceptual studies using synthesized DFs, Howell and Young found that increases in the loudness and duration of the replacing material were helpful in the processing of alteration repairs. In automatic processing, however, because the rate of prosodic marking is fairly low, it may be best used as a *negative* cue in DF detection (Nakatani & Hirschberg, 1994). That is, if a hypothesized repair region has much *lower* stress than preceding material, it is not likely that the region actually corresponds to part of a DF.

2.4.5.4 Characteristics of filled pauses

As mentioned in the section on classification, filled pauses (limited here to “uh” and “um”) have a questionable *linguistic* status in some classification systems. To justify the inclusion of filled pauses under the scope of DFs considered in this thesis, it is necessary to consider filled pauses as linguistic elements. A number of findings support this view.

First, although filled pauses have variants in many (perhaps all) languages (Delattre, 1965; Allwood et al., 1989), their vowel quality varies predictably with the vowel inventory of the language (Esling, 1971).[6] Second, the intonation of clause-internal filled pauses scales predictably with the intonation of surrounding words (Shriberg & Lickley 1992a, 1992b, 1993). Third, “um” and “uh” differ with respect to the length of the following unfilled pause (Clark, 1994) suggesting that speakers may choose among two different filled-pause forms depending on the time they need to pause.

---

6. Interestingly, this variation appears to be salient: use of the filled pause from a first language when speaking a second language is quite noticeable to listeners, as reported anecdotally to the author by a number of sources.

A number of studies have focused on the phonetic characteristics of filled pauses. These studies have typically aimed to aid the automatic recognition of filled pauses, which have been noted as a source of error for speech recognition systems (e.g., Butzberger, Murveit, Shriberg & Price, 1992). Difficulties include the problem that the filled pause “uh” is homophonous with the common determiner “a,”, as well as the fact that filled pauses can occur at any location in a grammar and are therefore difficult to model in a language model.

O'Shaughnessy (1992) and Shriberg (1991) studied the ATIS corpus and found that filled pauses tended to be low in F0 and to show a level or slightly falling F0 pattern. More specifically, Shriberg and Lickley (1992a,b; 1993) found that clause-internal filled pauses in both the ATIS corpus and casual British English conversations were uttered at an F0 that could be predicted from the F0 of the closest preceding peak F0. This suggests that filled pauses (at least those occurring within a clause) are intonationally well-formed. In addition, Shriberg and Lickley found that the intonation of filled pauses was unrelated to temporal aspects of the speech, suggesting that when durational lengthening is imposed by planning demands, speakers nevertheless preserve intonational characteristics of the speech.

2.4.5.5 Characteristics of repetitions

Dickerson (1971) distinguished between two possible functions of repetitions in the speech of nonnative speakers of English. Repetitions could function “as a pause device through habitual usage,” but could also function to “smooth over the break” after a long pause. Heike (1981) proposed a similar distinction in his work on native speakers of English and German, terming the “pause device” repetitions as “prospective repeats” (used to gain time for lexical search), and the repetitions following a long pause as “retrospective repeats” (used to reestablish fluency after a break.) Heike's prospective repeats were “as a rule not accompanied by silent pauses but sometimes by filled pauses,” whereas the retrospective repeats were always preceded by a pause. However, these studies do not report actual measurements of pauses, nor do they consider intonation.

O'Shaughnessy (1993) studied prosodic aspects of repetitions in the ATIS corpus and noted that: 1) there was little prosodic change from the first instance of the repeated word to the second; and 2) repeated words were shortened up to 50% when R1 (the first instance of the repeated word) showed significant prepausal lengthening.

2.4.6 Pattern features

2.4.6.1 Word fragment in RM

Levelt (1983) and others have noted that speech may be cut off mid-word, even mid-syllable, yielding what will be referred to in this thesis as a "word fragment" or simply a "fragment" at the right edge of the RM (see Chapter 3 for additional discussion of fragments).

One puzzle in the literature is that the overall rate of fragments varies considerably across studies. When expressed as a percentage of the DFs that contain a fragment at the cutoff: Levelt (1983) reported 22% for pattern descriptions in Dutch; Lickley (1994) found 36% for casual conversation in British English; and Bear et al. (1992) found 60% for the ATIS corpus. It is not clear what accounts for this variation.

Fragments have received particular attention in studies of automatic processing. As noted in Bear et al. (1992), knowledge about the location of fragments would be an invaluable cue to both detection and correction of DFs. Fragments provide a useful cue for DF *detection* cue because they are present for many DFs and for no non-DFs. Fragments aid DF *correction* because they always mark the right edge of the RM, thereby specifying the right boundary of material to be deleted.

Unfortunately, fragments pose a difficult problem at the level of automatic speech recognition. Many systems do not model fragments because they are constrained to output only full words. In addition, fragments are frequently cut-off *content* words (Nakatani & Hirschberg, 1994); but because they are typically brief in duration, fragments often are better acoustic matches for short *function* words. This can cause difficulties for a language model.

2.4.6.2 Length of RM

Bear et al. (1992) found that the length of the RM (in words) in DFs in the ATIS corpus was typically very short. Roughly 85% of the DFs (excluding filled pauses in otherwise fluent stretches of speech) contained only one or two deleted words.

2.4.6.3 Editing phrase in IM

Studies of DFs (Maclay & Osgood, 1959; Hockett, 1967; Schegloff, 1979; Levelt, 1983) as well as studies of interjections and discourse markers (James 1972, 1973) note that the IM of DFs may be filled with an "editing phrase." These phrases can range from filled pauses (e.g. "uh") to discourse markers (e.g. "well") to interjections or explicit negations of previous material (e.g. "oops").

Much of this work has focused on the relationship between the editing phrase and the type of DF. It is generally agreed that the choice of editing phrase reflects the type and degree of contrast between the trouble element and the alternation in the repair. In general this work has been based on intuitions and made-up examples, rather than on analyses of actual DFs. Empirical results, however, are reported by Levelt (1983), who notes a higher rate of "high-contrast" editing phrases for error than for appropriateness repairs, and suggests that phrases which involve explicit negation may serve to highlight those corrections which are most relevant to accurate communication.

Although there has been notable emphasis on the presence of editing phrases in these and other studies, including studies of DF processing, editing phrases may actually be relatively infrequent in DFs. Levelt reports that 30% of DFs in his corpus contained editing phrases; however, this figure includes surface DFs consisting of *only* the editing phrase (most commonly, a filled pause). However, if such cases are not counted, the rate of editing phrases is much lower. Similarly, the rate of editing phrases in other studies has been reported to be closer to 10% (e.g., Lickley, 1994).

2.4.6.4 Retraced words in RR

Speakers often retrace to words preceding the trouble element in making a repair. In the following example, the speaker retraces to "leave" before replacing "eleven" with "noon":

```
which flights leave after eleven -- leave after noon
```

Levelt (1983) proposed that the locations to which a speaker may retrace are those for which the resulting RR produces a well-formed syntactic coordination with the original utterance. This is a

more constrained rule than one phrased only in terms of onsets of grammatical constituents, since retracing to certain constituent onsets nevertheless produces an ill-formed repair.

An important unanswered question, however, is what predicts whether a speaker will retrace at all, since retracing is not obligatory in effecting a repair. Although Nooteboom (1980) observed that phonetic errors show less retracing than lexical errors, little else is known about predicting the presence of retracing.

It is also unclear whether retracing serves a speaker-related or listener-related function. Levelt (1983) explains retracing in terms of a speaker trying to help his listener's "continuation problem." A listener is faced with the problem of determining the intended utterance from the disfluent one; retracing provides a way for the listener to hook up the repair correctly with the original utterance. Clark (personal communication), however, suggests a speaker-related explanation: speakers may retrace to the onset of the grammatical unit that they are having trouble formulating.

Retraced words have also played an important role in approaches to automatic DF detection and correction in speech applications (see Section 2.5.2) because retracing results in a surface pattern in which words or word strings are repeated in close proximity (e.g., "leave after" in the preceding example). For example, the approach of Bear et al. (1992) relies heavily on the presence of repeated words in close proximity; approaches such as that of Nakatani & Hirschberg (1994) and Heeman and Allen (1994) also make use of this feature.

2.5 Studies of Disfluency Processing

Although the focus in the literature has been on the *production* of DFs, a small number of studies have examined the *processing* of DFs--both by human listeners and by machines.

2.5.1 Human processing

A remarkable aspect of DFs is that they go largely unnoticed in everyday comprehension. Human listeners are so apt at filtering out DFs that the task of recording what was actually said in utterances containing DFs is a difficult and unnatural one, often requiring many passes at transcription (e.g., Deese, 1980). Evidence from laboratory experiments indicates that attending to DFs and comprehending what is being said are mutually inhibitory

processes (Martin & Strange, 1968). Listeners asked to locate within-constituent DFs of other speakers tend to displace them to constituent boundaries (Martin, 1967; Martin & Strange, 1968), suggesting that DFs are filtered out in a separate stream in processing. This is of course as it should be: the fact that listeners can attend to a message despite the presence of DFs allows for rapid and efficient communication. If we could not so easily process speech with DFs, we should have to devote great time and attention to the planning and delivery of each utterance.

Little is known about *how* listeners so easily cope with DFs. A small number of studies, however, suggest that lexical, syntactic, and prosodic information play a role (e.g., Deese, 1980; Fox Tree, 1993; Lickley, 1994).

2.5.2 Automatic processing

Unlike the situation for humans, DFs can cause considerable trouble for computer "listeners," including automatic speech recognition and natural language understanding systems. Although DFs have often been viewed as an unavoidable, unimportant source of noise in the input to such systems, there is increasing interest in the direct modeling of DFs. This is attributable to at least three factors. First, as automatic systems have evolved to handle spontaneous, natural-sounding connected speech, the rate of DFs in the input has increased. Second, improvements in the performance of systems for fluent regions of speech mean that DFs have begun to be associated with a larger proportion of system errors. Third, the availability of large corpora of spontaneous speech has facilitated the study of DFs by assuring an adequate number and range of cases for examination.

Early work in the direct modeling of DFs was that of Hindle (1983). Hindle assumed a reliable, phonetically-identifiable editing signal at the point of interruption, as well as hand-marked grammatical constituents. The identification of the edit signal triggered a series of editors which checked input strings for DF-like structures and eventually discarded unwanted words from the view of the parser. Although this method achieved high success rates, it was based on information (in particular, the presence of an edit signal) that is not present in actual speech (see Section 2.4.5.1). However, Hindle's work demonstrated that DFs are, syntactically speaking, highly rule-governed events.

More recently, a small number of studies have explored issues of processing DFs in a Spoken Language System (SLS). Such systems involve both speech recognition and natural language understanding. A common aspect of these recent studies is that they do not rely on the presence of an edit signal at the IP, as assumed in Hindle's work. The studies can be characterized (to a rough approximation) on the basis of the presumed location in the SLS architecture where the DF processing would occur.

One approach (Bear et al., 1992; see also Langer, 1990) can be best characterized as "post-parsing." Bear et al. attempted to detect and correct DFs in the ATIS corpus. An idealized recognition output (i.e. a transcription of what the speaker actually said) was assumed, as well as the availability of various parsing techniques. Simple "pattern-matching" techniques were used to identify locations of potential DFs. The "pattern matcher" searched for repeated words in close proximity as well as syntactic anomalies and certain cue phrases (e.g. "oops"). Syntactic and semantic parsing techniques were then used to discriminate actual DFs from "false positives" and to guess at the appropriate correction for each DF. Integration of pattern-specific acoustic information was also suggested.

A second approach can be characterized as "pre-recognition" (O'Shaughnessy, 1992, 1993, 1994; Nakatani & Hirschberg, 1994). These authors point out that because the speech recognition problem is not solved, assuming a correct transcription is unrealistic; therefore an effort should be made to determine what cues to DFs can be obtained directly from the speech signal. Both studies analyzed data from the ATIS corpus. O'Shaughnessy's work points to acoustic characteristics of certain DF phenomena (filled pauses and repetitions) which could aid their recognition. Nakatani and Hirschberg analyzed cues across DF types and identified acoustic cues which, in combination with lexical and other information, could aid DF detection and correction.

A third approach may be characterized as "post-recognition, pre-parsing" (Heeman & Allen, 1994). Heeman and Allen aimed to locate "modification repairs" in the TRAINS corpus (Allen & Schubert, 1991). The approach involved combining the pattern-matching approach of Bear et al. with an estimation of the likelihood of DF at any particular word transition. Statistics for word transitions were based on comparing the part-of-speech categories of the words on either side of the juncture to those compiled for a training set hand-labeled for DFs. This

approach uses many of the same local features as used in the Bear et al. approach, but does not rely on parsing the complete utterance. Such an approach is particularly useful if there is inadequate syntactic or semantic coverage for a parser in a particular domain.

The success rates of these approaches are dependent on the type and amount of information available in the speech input (e.g., in some studies features were hand-labeled in the input), differences in the rates and distributions of DFs in the speech corpora, and differences in the types of DFs the authors aimed to detect. For these reasons, it is not possible to directly compare results across the studies described. However, all of the studies indicate significant promise for direct DF modeling in applications, although much work remains to be done to determine the optimal cues and methods in automatic DF processing.

2.6 Chapter Summary and Directions for Research

2.6.1 Summary

Past work has shown that most DFs have a three-region surface structure. The majority of DF classification systems make reference in some manner to the change in wording that occurs across these regions. However, classification systems differ widely across studies, and raise a number of methodological concerns.

Descriptive studies of DF production have shown that overall rates of DFs, as well as rates of specific DF types vary based on features of the domain, the speaker, and the sentence in which a DF is produced. DFs have been found to show characteristic surface form in terms of syntactic, phonetic, and pattern features.

Studies of DF processing reveal that DFs are filtered out easily in human perception; although little is known about *how* this is achieved. DFs present difficulties, however, for automatic speech processing, prompting interest in direct modeling of DFs for spontaneous-speech applications.

2.6.2 Directions for research

The review illustrated that we currently have a large base of information on DFs. This information provides an important starting point for the present work. However, in order to

pursue a unified theory of DFs, four areas remain to be addressed. These areas guide the methodology and analyses in the remainder of the thesis.

1. Develop a theory-neutral classification system

A theory-neutral language for classification is crucial for replicability, cross-study comparisons, and communication among researchers in different fields. The DF representation and classification systems developed in this thesis (Chapter 4) represents an attempt to address this need.

2. Gather robust trends

The gathering of robust trends is critical for determination of appropriate features to include in a model. Although many features have been examined in past work, it is important to determine features that show regular *trends*, and to assure the statistical validity of trends by examining as much data as possible. This is the goal of many of the individual analyses in Chapters 5 and 6.

3. Describe relationships among features

In past work, there has been relatively little study of relationships among different features. Knowledge of feature relationships is essential for devising appropriate classification systems. It is also crucial for statistical analyses, since correlations among features must be determined and handled appropriately in hypothesis testing. It is additionally critical for generalizability: for example, we need to know which and how features of DFs change across speakers, or across domains. And finally, a good understanding of feature relationships should guide hypotheses about the nature of the explanatory variables underlying the features. As mentioned in Chapter 1, it is in terms of these variables, not features, that a unified theory must ultimately be phrased. Exploration of feature relationships is the goal of many of the analyses in Chapters 5 and 6.

4. Attain predictive power

Previous studies of DF production have been descriptive, or have used inferential statistics, but have had no predictive power. In order to devise and test theories, it is critical to model the trends. Direct modeling thus far on DFs has mainly been limited to work on automatic processing. In most cases, these models have been *nonparametric* models, essentially lookup

tables of conditional probabilities based on training data. Although these models do have predictive power, we cannot compress the knowledge encoded in a lookup table to an intuitive understanding of the phenomena.

A way to advance our understanding of disfluent phenomena is to pursue *parametric* models, so that the interpretation of parameters can suggest constraints for theory. Parametric models also provide simpler and more elegant models for use in automatic processing. In this thesis, parametric models are suggested for certain analyses (namely those which involve an ordinal random variable and include sufficient data) in Chapters 5 and 6.

Chapter 3: Speech Corpora

3.1 Chapter Overview

This chapter describes the speech corpora used in the present work. Section 3.2 outlines the requirements for the corpora from the perspective of the goal of the thesis as outlined in Chapter 1. Section 3.3 describes the three corpora chosen. Section 3.4 describes the checking and editing of original transcriptions in preparation for analyses of DFs. Section 3.5 summarizes the main points from each section.

3.2 Requirements for the Corpora

In choosing the set of spontaneous speech data for the present work, a number of requirements were considered. First, as stated in Chapter 1, this research required a large amount of speech data to assure coverage of the representational system and validity in analyses. Second, it was necessary that digitized speech waveforms be available, in order to allow checking of transcriptions as well as acoustic analysis. Third, because of the requirement of corpus size, availability of orthographic transcriptions was necessary to keep the research within the scope of available time resources. Fourth, only corpora in which the task did not involve face-to-face contact between speakers were considered, because this eliminates the factor of gestural cues (which, as mentioned in Chapter 2, affect DF production). Fifth, it was considered essential to include more than a single corpus, in order to distinguish properties likely to be associated with a particular domain from those potentially invariant over domains. A final priority was to choose corpora currently of interest to researchers in speech recognition and understanding. This would provide a link for future work on integrating results in real systems, and foster analysis by providing tools for extraction of some features by automatic methods.

Consideration of these factors led to the choice of three corpora: 1) the ATIS corpus of human-computer dialog in the air travel planning domain (MADCOW 1992, Dahl et al., 1994); 2) the SWITCHBOARD corpus of informal human-human telephone conversations on various topics (henceforth “SWBD”; Godfrey, Holliman & McDaniel, 1992; Wheatley et al., 1992); and 3) the AMERICAN EXPRESS/SRI corpus of human-human air travel planning dialogs (henceforth “AMEX”; Kowtko & Price, 1989).

ATIS was an ideal choice because in addition to meeting the criteria above, it was a corpus familiar to the author, who was involved in the data collection procedures. SWBD also meets the criteria listed above and complements ATIS nicely, since it represents a much more natural-sounding style of speech. The AMEX corpus, while smaller and not presently digitized, was chosen because it provides a useful comparison corpus, sharing the general task of air travel planning with the ATIS corpus, but involving human-human conversations as in the SWBD corpus.

A notable limitation of this set of corpora is the lack of within-subject control across corpora. An attempt was made to find speakers who participated in both the ATIS and SWBD tasks; however the number of definite cases was too small to be of use. A second limitation is a lack of sufficient disfluent speech per speaker for studying individual differences in the ATIS and AMEX corpora (due to a limited amount of speech per call in the AMEX corpus, and to a very low overall DF rate in ATIS.) However, analyses of individual differences are possible using the SWBD data.

3.3 Description of the Corpora

3.3.1 ATIS

3.3.1.1 Task and speakers

The ATIS corpus is a large, multi-site[1] corpus of human-computer dialog in the air travel planning domain. The corpus is distributed by the Linguistic Data Consortium (LDC). Further description of the corpus can be found in MADCOW (1992), Dahl et al., (1994), and in publicly available documentation files.

In the ATIS task, subjects were given hypothetical travel "scenarios," for example:

> You live about two hours from Philadelphia, and your family is coming to visit. So that you can pick them all up at the same time and drive the two hours only once, find flights for them that arrive within one hour of each other. The family members are coming from Baltimore, Denver, and Pittsburgh. After you have found the flights, check to make sure that the round trip fare for each is less than $1000.

[1] Participating sites included: ATT, BBN, CMU, MIT, SRI, and NIST.

Subjects “solved” the scenario by speaking to a computer. For example:

```
SUBJECT: Which nonstop flights go from Baltimore to
Philadelphia
```

The system responded with an answer such as:

```
SYSTEM: Here are the nonstop flights from Baltimore to
Philadelphia
```

along with a table containing the relevant information. Systems varied across sites and over the course of the collection period. In some cases a fully automatic Spoken Language System (SLS) was used; in other cases, particularly in earlier systems, the machine was simulated using a human “wizard” in the loop. Speakers were recruited at the various sites as described in the documentation; most were affiliated with the research institutions (employees or their acquaintances).

3.3.1.2 Recordings and transcriptions

Each query to the system was individually digitized and orthographically transcribed at the collecting site. The transcription conventions include notations that facilitate the extraction of DFs.

3.3.1.3 Selected data

The speech data selected for the present work (over 180,000 words) include the majority of utterances from the ATIS2 corpus, as well as ATIS3 training and development data distributed as of July, 1993.

3.3.2 AMEX

The AMEX corpus of human-human air travel planning dialogs consists of telephone speech between SRI employees and American Express travel agents. Details of the corpus are described in Kowtko and Price (1989).

3.3.2.1 Task and speakers

In this corpus, no task was set; rather, telephone conversations between SRI employees and American Express travel agents (i.e., calls involving real travel plans) were tape-recorded after agents obtained the employee's permission at the beginning of the call. Some employees called to make their own travel plans; others were administrative staff arranging plans for another employee, as in the excerpt below:

```
CLIENT: All right then coming back if if you ha- do you have a
late flight out on the twenty third? Probably not, huh?


AGENT: Well let me see.


CLIENT: He said either l- late on the twenty third or or any
time on the twenty fourth.


AGENT: Okay, there's an eight PM departure on Air Canada. It's
flight four sixty seven. It arrives into Toronto at eight
fifty five, connects to Air Canada flight seven fifty five.


CLIENT: No I'm it leaves Ottawa four sixty seven leaves Ottawa
at what time?
```

3.3.2.2 Recordings and transcriptions

Conversations were recorded in analog over the telephone line onto audio cassette tape, and are of variable acoustic quality. Speech was transcribed orthographically by a single transcriber with a background in linguistics. Although DFs were not marked in transcriptions, word fragments and filled pauses were indicated.

3.3.2.3 Selected data

Only the client's speech (and not the speech of the agent, either to the caller or to other agents) was chosen for analysis, because agent speech was highly ritualized and should be modeled separately. Sixty-eight conversations (over 12,000 words) were randomly selected; these represented sixty-six different speakers (two speakers were represented twice each). To the extent that it was possible to tell from the caller's voice (if that conversation was checked against

the tapes, or if the caller's name was mentioned in the transcription), slightly more than half of the callers were female, and about one third of the calls were made by administrative staff for another person.

3.3.3 SWBD

The SWBD corpus contains roughly three million words from over 2430 human-human conversations over the telephone on various topics. The corpus was collected at Texas Instruments and is distributed by the LDC. Further description is found in Godfrey, Holliman, and McDaniel (1992), in Wheatley et al. (1992), and in extensive documentation files distributed with the corpus.

3.3.3.1 Task and speakers

A set of 500 speakers representing all major dialects of American English participated in the task in exchange for a per-call remuneration. Speakers could participate as often as they desired; many speakers participated multiple times. Speakers were aware their speech would be recorded but informed only generally that TI speech researchers were interested in the conversations.

Speakers registered by choosing topics of interest (e.g., recycling, sports) from a predetermined set, and by indicating times they were available. They were automatically connected to another caller by a "robot operator" based on matching of registrants to interest topics and available times. An advantage of this procedure is the absence of experimenter bias. Conversations were therefore between strangers; however, transcribers rated the majority of conversations as sounding highly "natural."

Speakers conversed for up to ten minutes. A excerpt from a conversation on universal health insurance is below. The "prompt" was a recorded topic description played at the beginning of the call.

PROMPT: "DO YOU BELIEVE THAT THE US GOVERNMENT SHOULD PROVIDE UNIVERSAL HEALTH INSURANCE, OR SHOULD AT LEAST MAKE IT A LONG TERM GOAL? HOW FAR IN THAT DIRECTION WOULD YOU BE WILLING TO GO? WHAT DO YOU SEE AS THE MOST IMPORTANT PROS AND CONS OF SUCH A PROGRAM?"

SPEAKER A: Okay. Well I, how do you feel about that?

SPEAKER B: Uh, I think i-, it's a very good idea personally, because right now I don't have any health insurance at all. And I just know that it's fir-, I could get it where I work an-, but it would be like a hundred dollars a month which is a lot to me and I really just can't afford it. Uh, I have worked places where I had insurance but, you know, on a more, I work for a very small company now and I, I realize that, you know, it would, for that to be true you'd, you'd have to increase taxes and stuff like that but, uh, I know other people that are, you know, in the same situation and when it's just, uh, I, if you don't work for a large company, it's really hard to have insurance. Uh, I would be willing to, you know, pay more taxes. But, uh, and I, I think it's just, I think everybody should have the option to have health care. I mean I don't think, you know, but the only way to do that would be to, you know, have to tax more.

SPEAKER A: Yeah, the-, there would have to be some way to pay for it. The, the, I agree with you that, that, uh, it's no fun to, uh, not have insurance. I'm, I've had health insurance through my, uh, either my parents, uh, employment or my employment for many years but, uh, now I'm looking s-, at a layoff situation and I'm sitting here saying well how much is it going to cost me.

3.3.3.2 Recordings and transcriptions

Digital speech signals were collected directly from the telephone network. Therefore, although the signals are telephone bandwidth, the sound quality is quite good (in most cases adequate for pitch tracking.) The two channels from each conversation can be isolated, which facilitates analyses of a single speaker's speech.

The digitized waveforms containing both channels of an entire conversation are available on CDROM, along with orthographic transcriptions and automatically-generated word-alignment files. Transcribers were instructed to mark word fragments and to transcribe filled pauses, but there were no other special conventions for marking DFs.

3.3.3.3 Selected data

The selected data (over 40,000 words) include two conversations from each of 30 speakers (15 male, 15 female). The size of this selected data set was determined by conducting a pilot study to estimate the amount of data that would be needed to roughly equate the SWBD subcorpus to the ATIS subcorpus in terms of total DFs. The 30 speakers were randomly drawn from the set of speakers who participated in at least 10 conversations, until a matched set of males and females balanced for geographic origin was obtained. Although only two conversations from each of these 30 speakers were used in analyses, these frequent speakers were chosen so that additional data would be available for studies of individual differences in future work.

3.4 Editing of Transcriptions

Inspection of original transcriptions in all three subcorpora revealed significant discrepancies between transcriptions of what was said, and the level of detail and accuracy needed for the present work. This is not surprising, nor is it a criticism of the quality of the databases. DFs were not a focus in the original transcription of the corpora, and (as mentioned in Chapter 2) it is an unnatural task to listen for DFs, since we are highly practiced at filtering them out in everyday comprehension. The different corpora presented different issues for transcription quality, as described in the following three subsections.

3.4.1 Issues for ATIS

ATIS transcription conventions explicitly mark the presence of the DFs of interest in the present work. It was determined that these markings provide a highly reliable method for the *recall* of disfluent utterances. One thousand utterances containing no DF markings in original transcriptions were listened to for the presence of missed DFs. Half were drawn randomly from the entire subcorpus; the other half were drawn from the set of utterances made by only the "disfluent" speakers (i.e. those speakers who had made at least one DF according to original transcriptions). No DFs were found in either set, and only minor mistranscriptions of the fluent speech were found. Therefore the task of editing transcriptions for the ATIS subcorpus was limited to those utterances that contained DF markings in their original transcriptions.

Of the utterances in the subcorpus, 1760 originally contained some DF marking. All of these were listened to, and of these, 1342 utterances actually contained at least one DF. The discrepancy was largely due to the use of a single symbol ("-") for both machine truncations in recording and self-interruptions (this ambiguity has since been addressed in MADCOW transcription conventions).

Transcriptions of the disfluent utterances were edited to correct errors and to add information necessary for the present study. One necessary modification involved mispronounced words. As found in Chapter 6, ATIS has a high rate (relative to the total number of DFs) of speech errors; these were not reflected in original transcriptions because transcribers tended not to mark a word as *both* mispronounced *and* cut off, for example:

```
ORIGINAL TRANSCRIPTION: i'd like to go from bos(ton)- to
baltimore on u s air


REVISED TRANSCRIPTION: i'd like to go from [bals] boston to
baltimore on u s air
```

In the original transcription, the fragment was not marked with the prescribed symbols for mispronounced words; in addition, the fully pronounced instance of "boston" was completely missed. This difference is important for the present work since it distinguishes a speech error from a repetition (see Chapter 4).

Other modifications involved: 1) correction of inappropriate use of the fragment symbol for dialectal variants or for unreleased final stops (e.g. "oaklan-"); 2) insertion of the fragment symbol for words that were transcribed in full but not actually completed; 3) correction of errors in filled-pause transcription; and 4) insertions of fragments missed altogether. Overall, roughly 15% of the disfluent-utterance transcriptions were altered.

3.4.2 Issues for AMEX

Complete checking of AMEX transcriptions was difficult because the corpus was not digitized and the randomly-selected conversations were distributed throughout the audio cassettes. Therefore this corpus was only spot-checked. Six conversations were randomly selected and checked against the recordings. This revealed a low rate of missed DFs (under 5%)

and relatively minor discrepancies in the manner in which words in disfluent regions were transcribed (mainly an underrepresentation of fragments). This error rate is likely to be representative of the corpus as a whole, since all conversations were transcribed by the same person. The error rate was deemed tolerable for the present work, particularly since AMEX was included as a comparison corpus.

3.4.3 Issues for SWBD

All turns in the selected data that contained four or more words were checked against transcriptions. The majority of turns of three or fewer words consisted of only "continuers" or "assessments" (Goodwin, 1986) such as "uh-huh" and "yeah" which did not pose much opportunity for disfluency (as discussed in Chapter 5). The original transcriptions for these turns were found to be highly accurate in a spot check.

Over 25% of the disfluent-turn transcriptions were modified. The majority of problems were not due to completely missed DFs, but rather to precise transcription of a particular DF. A large category of incorrectly transcribed DFs involved missing fragments surrounded by repeated words, for example:

```
ORIGINAL TRANSCRIPTION: because as i said, our our schnauzer

REVISED TRANSCRIPTION: because as i said, our d(og)- our
schnauzer
```

These were audible by careful listening and quite visible in spectrograms. The difference is important since it distinguishes a simple repetition (which could represent, as discussed in Chapter 2, a hesitation form) from a replacement (an overt correction of error). Interestingly, fragments tended to be missed by transcribers in the cases just described more often than when at the end of a fresh start, indicating potential biases in detection.

A number of errors involved transcription of the filled pauses "uh" and "um." These were either missed entirely, or misplaced, which is not surprising given the results of the perception studies mentioned in Chapter 2. In addition, "uh" was sometimes used for grunts or unintelligible vocalizations, heavily released stops, or for "um." The difference between "um"

and “uh” is important, as suggested by Clark and colleagues (see Chapter 2), so these errors were corrected.

The SWBD transcription manual lacks a convention for indicating mispronunciation. Therefore speech errors were not adequately transcribed. For most mispronounced words, transcribers inserted a “real” word, resulting in either an ungrammatical word sequence, e.g.:

```
ORIGINAL TRANSCRIPTION: everybody just lives under an
umbrella the uncertainty with housing


REVISED TRANSCRIPTION: everybody just lives under an
umbrella[tiy] uncertainty with housing
```

or in a spurious DF pattern, such as the apparent repetition below:

```
ORIGINAL TRANSCRIPTION: they've been beefing up the police
patrols trying to put put more guys more cops higher...


REVISED TRANSCRIPTION: they've been beefing up the police
patrols trying to [por] put more guys more cops higher...
```

3.4.4 Ambiguous events

Across corpora, the transcription of certain DF-related events was found to be ambiguous at the level of the surface utterance. Ambiguity was particularly notable in transcribing word fragments and filled pauses.

3.4.4.1 Fragments

Ambiguity in the transcription of word fragments was found to arise in a number of situations. First, since coarticulation effects across word boundaries are expected in connected speech, a word directly preceding the interruption point of a DF may be fully articulated, but may sound cut off due to a lack of these effects across the interruption point (see also Lickley, 1994). Second, for single-phone words such as “I,” it is difficult or impossible to determine whether the word is prematurely cut off. Third, cases arise in which a word can be either a fragment of a corresponding word in the RR, or a full word that makes sense in the context, but

is replaced in the RR. This makes a difference in DF classification (see Chapter 4), since the latter reading corresponds to a substitution whereas the former does not. This situation arises for a large number of cases in which a word is followed by a contraction containing that word:

```
so that(-?) that's what brought us here
```

as well as to a variety of other rare cases, e.g.:

```
where's the stop(-?) where's the stopover

what are the cheap(-?) cheapest one way flights

so hers was the first one i had got(-?) gotten

if they raised the tax(-?) city taxes
```

Somewhat surprisingly, even with prosodic cues, it was not always possible to distinguish the full-word from the fragment readings.

3.4.4.2 Filled pauses

Ambiguities also occurred in the transcription of filled pauses. First, the filled pause "uh" is generally homophonous with the determiner "a." Either possibility may be appropriate in certain semantic and syntactic contexts, for example:

```
what type of a(?)/uh(?) plane is that flight
```

Added ambiguity is present for repeated a/uh when "a" is syntactically viable, since this can correspond to "a a," "a uh," or "uh a," the first of which is a repetition:

```
i need a(?)/uh(?) a(?)/uh(?) boston to denver flight
```

Second, the distinction between “uh” and “um” is often difficult to perceive before [m]-initial words when there is no intervening pause. Third, various vocalizations, such as voicing the release of a consonant, sound similar to “uh.”

In all situations described above, if there was no strong favoring of one possible transcription over the other after hearing the speech (using all available information, including prosodic and contextual cues) the original transcription supplied with the corpus was used.

3.5 Chapter Summary

Requirements for a data-driven, pretheoretical study led to the choice of three corpora of spontaneous speech: ATIS, AMEX, and SWBD. Advantages of this set include size, availability of digitized speech and orthographic transcriptions, representation of a range of speech styles, and use of AMEX as a comparison corpus. Disadvantages include the lack of within-subject control across corpora, and small disfluent-speech samples per speaker in ATIS and AMEX. Data were randomly selected from each corpus for hand-labeling. The selected data set consists of over 240,000 total words, representing over 600 different speakers.

In order to attain the accuracy and detail needed for analyses of DFs, original orthographic transcriptions were checked against the soundfiles and modified where necessary. While some modifications reflect misapplication of original transcription conventions, or lack of adequate conventions for representing DFs orthographically, other discrepancies suggest interesting biases in DF detection that are worthy of separate study. The process of checking transcriptions revealed that the transcription of some DFs is inherently ambiguous in certain contexts.

Chapter 4: Feature Annotation and Type Classification

4.1 Chapter Overview

This chapter details how the speech data described in Chapter 3 were annotated and classified in preparation for the analyses in Chapters 5 and 6. Section 4.2 describes the annotation of features, organized for convenience by feature dimension as established in previous chapters. Section 4.3 describes a system developed to encode pattern features of a DF in a single representation. Section 4.4 presents an algorithm for collapsing pattern representations into a small number of orthogonal classes or DF "types." Section 4.5 describes the resulting database of labeled DFs; this database forms the basis for the analyses in the following chapters.

4.2 Feature Annotation

The description of features is organized for convenience by feature dimension as established in earlier chapters. Recall from Chapter 1, however, that these dimensions reflect grouping only at the level of observable features; they have no reality at the level of analysis. The complete list of features annotated is summarized in Table 1. The third column refers to the relevant chapter section describing the annotation of the feature.

4.2.1 Domain features

As mentioned in Chapter 1, domain features are not a focus in this work. Domain features are best analyzed when one can control for individual speaker and independently vary aspects of the domain, which was not possible in the present study (see the description of corpora in Chapter 3). The domain feature examined was simply the source corpus for each DF (ATIS, AMEX, or SWBD). These corpora obviously vary along many dimensions. Hypotheses about which aspects of a corpus were likely to be responsible for differences in trends across corpora could not be tested directly, but are suggested in the relevant discussion sections based on the background literature presented in Chapter 2.

4.2.2 Speaker features

The examination of individual differences was limited to the set of 30 speakers in the SWBD corpus. These were the speakers for whom there was the greatest amount of

Table 1: Summary of Annotated Features

Feature Dimension	Feature	Section(s) Described
Domain:	corpus	4.2.1
Speaker:	identity	4.2.2.1
	gender	4.2.2.1
	speech rate	4.2.2.2
Sentence:	sentence length	4.2.3.1
	DF position in sentence	4.2.3.2
	other DF(s) in sentence	4.2.3.3
Acoustic:	duration	4.2.4.1
	fundamental frequency	4.2.4.2
Pattern:	length of RM	4.2.5.1
	word fragment at IP	4.2.5.2
	filled pause in IM	4.2.5.3
	editing phrase in IM	4.2.5.3
	discourse marker in IM	4.2.5.3
	retracing in RR	4.2.5.4
(Derivate-Pattern):	DF “type”	4.2.5.5, 4.3, 4.4

disfluent speech data available. The ATIS and AMEX corpora contained too little disfluent data per speaker for meaningful analyses of individual differences. In the case of the AMEX corpus this was due to a lack of enough speech overall. In the ATIS corpus, the problem was the very low *rate* of DFs (as discussed further in Chapter 5).

4.2.2.1 Identity and gender

Each speaker was assigned a unique identifier and noted as male or female.

4.2.2.2 Speech rate

As a rough estimate of speech rate, a single words/second statistic was automatically computed for each of the 30 SWBD speakers. For each speaker, turns containing at least 15 words (according to original transcriptions) were selected from the set of all utterances by that speaker in the original SWBD corpus (i.e. including conversations outside the two per speaker chosen for hand-labeling). Average speech rate was computed for each speaker by dividing total words by the total time speaking in the selected turns. This measure therefore includes words in the reparanda of DFs as well as unfilled pause time. Random spot-checking revealed that the accuracy of the automatic measure was satisfactory for 29/30 speakers. For the remaining speaker, the automatic measure was not reliable due to cross-talk in original recordings; therefore speech rate was not assigned for this speaker.

4.2.3 Sentence features

The unit marked as a "sentence" in this work can best be described as a unit that would be likely to be terminated by a period or question mark in conventional orthography. Although this definition is necessarily vague (as is the notion of a sentence in linguistics in general), the marking of sentences is likely to be reasonably consistent across the data examined because sentence boundaries were hand-marked by the same person (the author).

Full clauses joined by a coordinating conjunction were considered separate sentences, as in the following example ("/" indicates a sentence boundary):

```
we like to rent films / and recently we rented twin peaks /
```

Clauses joined by a subordinating conjunction, however, were considered part of the same sentence, e.g.:

```
i keep the plants by the window because it's warmer there /
```

In the interest of simplicity, all speech was accounted for using the same sentence unit. That is, every word was contained in exactly one unit, and no qualitative distinctions were made among units. It is important to note that for this reason, the notion of a sentence was necessarily stretched to cover cases that depart from the traditional notion of a sentence.

One such case involves sequences consisting only of one or more expressions such as “uh-huh” and “right,” which serve an interactional function in acknowledging another speaker's contribution. These elements, called “continuers” and “assessments” by Goodwin (1986), are in many accounts of discourse not considered real turns. By uttering these elements a speaker does not take the floor, but rather encourages the current speaker to continue.

Each sequence of words consisting of only continuers or assessments was nevertheless coded as a sentence, because it seemed unreasonable to group these elements with preceding or following material. Thus continuers or assessments occurring alone in a turn, or followed by additional speech, were sectioned off as a separate sentence as illustrated by speaker B’s turn in the following excerpt:

```
SPEAKER A: well i was in a rural area in wisconsin /

SPEAKER B: uh-huh /

SPEAKER A: and uh we had a good public school system /
```

Continuers and assessments were rare in the ATIS corpus, but frequent in the AMEX and SWBD corpora. In both AMEX and SWBD, nearly all sentences of two or fewer words consisted of only these elements.

A second case in which the definition of a sentence was stretched involves elliptic speech, or speech contextualized by preceding information. These cases were particularly common in the two travel-planning corpora, since speakers often provided less than a full-sentence answer in response to a prompt for information, as exemplified in the response of the client below:

```
AGENT: what day would you like to depart /

CLIENT: monday /
```

Similarly, cryptic speech, in which many words were left out, was not uncommon in the ATIS corpus and had to be represented in the sentence codings. For example:

```
origin boston, destination denver /
```

With the exception of fluent ATIS sentences, all speech in the HLD (i.e. both fluent and disfluent sentences) was hand-marked by the author for sentence boundaries during the process of checking transcriptions against recordings. Although sentence boundaries were indicated by standard punctuation in original AMEX and SWBD transcriptions, for consistency it was necessary to relabel them according to the present approach.

The set of fluent ATIS utterances (N=17,256) was too large to check by hand for utterances containing more than one sentence. To estimate the rate of multiple-sentence utterances (turns), 400 utterances were randomly selected. Of these, only two multiple-sentence utterances were found (in each case the utterance contained two sentences.) Therefore, "utterance" was used to approximate "sentence" for the set of fluent ATIS utterances.

4.2.3.1 Sentence length

Sentence length (in words) was computed two ways: total words and "efficient" words. The latter measure excludes words deleted in DFs. For example, the sentence below contains 24 total words, but only 16 *efficient* words after words in the RM and IM of DFs have been deleted (a line is drawn through the deleted words):

~~um~~ ~~i would like uh~~ i would like to ~~book a flight~~ book a flight for sunday from miami florida to las vegas nevada

The "efficient length" measure was used in all statistical analyses. This is because of the inherent confounding of length with any feature related to the presence of a DF, since by definition all DFs in this thesis contained one or more deleted words. Efficient length was computed automatically from the HLD using post-processing tools.

4.2.3.2 DF position in sentence

As mentioned in previous chapters, syntactic features are not a focus in this work. However, because previous work indicates the importance of syntax to DF production (see Chapter 2), a very coarse syntactic distinction was annotated: the position of each DF was hand-labeled as either sentence-initial or sentence-medial.

Hand-labeling of position was necessary because the conventions adopted for determining initial position skipped over any sequence of elements preceding the "syntactic onset" of the sentence. These skipped elements included coordinating conjunctions and

discourse markers, which do not constrain the following material syntactically. In addition, annotation of initial position skipped over any preceding material deleted in another DF, including any filled pauses.

The three following examples illustrate these conventions. In each case, the first word corresponds to the beginning of the hand-marked sentence, but all words up to the disfluent repetition are considered invisible to the position coding. Thus in all cases, the DF "the the" is considered sentence-initial:

```
well the the dog was old

and so the the dog was old

Rover was uh the the dog was old
```

In future work, position annotation could be automated to some extent by using information about the location of sentence boundaries, lexical lookup lists for conjunctions and discourse markers, and information on the location of words deleted in preceding DFs.

4.2.3.3 Other DF(s) in sentence

The presence of additional DFs in the same sentence as the DF being annotated was determined automatically from the HLD. (See Section 4.5 for a description of the file format from which this feature was directly obtained.)

4.2.4 Acoustic features

The acoustic features examined include duration and F0. Phone and word durations for speech from the ATIS corpus were obtained automatically using the SRI speech recognition system (Decipher^{TM}) in forced alignment recognition. In this procedure, durations are obtained from the forced Viterbi segmentation (Viterbi, 1967; Forney, 1973), where the recognizer is constrained to produce a known word string. Forced alignment is the optimal path (in the maximum-likelihood sense) assuming the words are known (and the underlying state machine models for them). Duration is extracted from the resulting segmentation.

Waveform analysis was performed using the Entropic ESPS/Waves 5.0 software on a Sun workstation. To facilitate waveform analysis of the SWBD files, the original conversation-sized, dual-channel files were automatically broken down into turn-length, single-channel files (based on the ".mrk" files distributed with the SWBD corpus). The breakdown program also produced Waves/ESPS format "label" files with time-aligned transcriptions based on the original SWBD word alignments, so that relevant locations within a turn could be easily located in waveform analysis.

4.2.4.1 Duration

Durations of filled pauses in the ATIS corpus were obtained using the output of the forced alignment recognition procedure described above. Durations of each word in a single-word repetition, and of any intervening pauses, were hand-labeled for the SWBD data with the aid of wide-band spectrograms. Although the SWBD corpus is distributed with alignments which are helpful in locating points of interest, the alignments were not accurate enough for the present work, particularly because the automatic alignment procedures used are not accurate in the environment of DFs and lengthened syllables.

4.2.4.2 F0

F0 tracks were generated by the ESPS software using default parameter settings. F0 was recorded at hand-measured locations as described in Section 6.4.4.

4.2.5 Pattern features

As defined in Chapter 1, pattern features represent aspects particular to *disfluent* speech. Pattern features and DF "type" were both determined directly from the pattern representation for each DF. The methods for labeling the DF pattern, and for computing DF type from that pattern, are discussed at length in Sections 4.3 and 4.4, respectively. Here, the specific features extracted from the pattern produced by the PLS and used in analyses in Chapters 5 and 6 are briefly listed. For descriptions of the symbols referred to, consult Section 4.3. In all cases the information was obtained automatically from the pattern representation using simple UNIX pattern processing languages (e.g. grep, sed, and gawk).

4.2.5.1 Length of RM

The number of words in the RM (length of the deleted string in each DF) was obtained directly from the DF pattern by counting the number of word-symbols preceding the IP. This measure does not include words in the IM, such as editing phrases, which are tallied separately (see below). Contracted words and word fragments were counted as full words.

4.2.5.2 Word fragment at IP

Presence of a word fragment preceding the IP was determined directly from the DF pattern by searching for the “-” symbol.

4.2.5.3 Words in IM

Filled pauses, editing phrases and discourse markers that appeared in the IM region were found by querying the pattern for any 'f', 'e', or 'p' symbols, respectively, that followed the IP and preceded any non-[f/e/p] word symbols. Reasons for distinguishing these three elements are given in Section 4.3.

4.2.5.4 Retracing in RR

The presence of retraced words (see Chapter 2) was determined by querying for any post-IP 'r' symbols preceding the 's-string' in the pattern.

4.2.5.5 DF “type”

DF type is considered a derivate-pattern feature because types were classified using an algorithm that takes as input the pattern representation for a DF, as explained in Section 4.4.

4.3 Pattern Labeling System (PLS)

This section describes the system developed for representing DFs based on pattern features, the “Pattern Labeling System” (PLS). The PLS is discussed in four sections. Section 4.3.1 outlines the goal and requirements of the system. Section 4.3.2 explains differences between the PLS, and the system upon which it was based (the “SRI system”). Section 4.3.3 provides a brief summary of the system in table form, noting the minimum information necessary to understand the analyses in Chapters 5 and 6. The last section is an

optional reading section, in which the PLS is described in further detail (following the organization of the summary). The detailed description is provided for those readers particularly interested in issues of DF representation, or desiring further explanation of specific aspects of the system after reading the summary.

4.3.1 Goal and requirements

The PLS was developed to represent the range of disfluent phenomena considered, in a manner appropriate to pre-theoretical, cross-corpus research. As defined in Chapter 1, the DFs within the scope were those produced by adult normal speakers of American English, and involving deletion of a contiguous sequence of linguistic material. The system was required to avoid the methodological concerns raised in Chapter 2. In particular, it aimed to: 1) use concise terminology; 2) avoid pre-theoretical hierarchical grouping of classes; 3) avoid reference to features (such as semantic features) that could not be reliably labeled from surface patterns for all cases, or that cannot be applied equally across domains; 4) produce orthogonal classes; and 5) provide complete coverage for the data examined. In addition, given the large number of DFs to be labeled, a practical requirement was that labeling time should be kept to a minimum.

4.3.2 Changes from SRI system

A previous system well-suited to meet many of these requirements was a system developed at SRI with colleagues John Bear, John Dowding, and Patti Price for work on automatic detection and correction of DFs in the ATIS corpus. This system, which will be referred to as the "SRI system," is outlined in Shriberg et al. (1992) and Bear et al. (1992), and described in further detail in Bear et al. (1993). However, to suit the needs of the present work, a number of modifications were made to the SRI system. These include refinements, extensions, and an assortment of minor but helpful notational changes. The refinements and extensions are discussed briefly below.

Two well-motivated refinements of the system were: 1) elimination of coindexing; and 2) allowing of variable-length substituted strings. Coindexing is described in Bear et al. (1993); it involves assigning an index to each word symbol to associate the word with the correct corresponding symbol on the other side of the IP. This allows reordering of words to be represented, for example:

```
give me your all all your nonstop flights from dallas
          1   2  2   1
```

Reordering has also been proposed as a DF class by other researchers, e.g., Allwood et al. (1989), who cite this example of reordering (translated from Swedish):

```
but then had  I   I had ...
          1  2   2  1
```

However, after labeling the large number of DFs in the present work, it was found that when reordering occurred (which was relatively rarely), it was similar to one of the two cases above. It appears that each of these cases could be construed as a type of DF not involving reordering at all. The first case seems to correspond to a speech error; the second case appears to be a deletion that happens to result in a surface pattern showing the same words in reverse order. (Treatment of these types of DFs in the PLS is explained in Section 4.3.4 below.) Therefore, reordering was viewed as an unnecessary, and more importantly, *unparsimonious* aspect of the system, since it grossly overgenerates the set of possible DFs. Elimination of reordering allowed for elimination of coindexing, which simplified the system and reduced labeling time.

The second basic refinement to the SRI system was to allow variable-length substituted strings, such as:

```
does the United flight 201 serve a sn(ack)- breakfast
```

In this case, it is clear that “breakfast” is being substituted for “a snack.” Because of the coindexing constraints in the SRI system, cases that did not contain word-for-word correspondences, such as that above, could not be explicitly labeled as substitutions. From the point of view of the SRI research, it is possible the correspondence could be detected by attaching syntactic and semantic information to labels. However, for the present work, no additional labels were used in analyses, and thus it was critical to capture the substitution relationship. Since coindexing is absent in the PLS as explained above, variable-length substituted strings pose no problem and are therefore allowed.

In addition to these two refinements to the SRI system, four enhancements to the system were added to provide full coverage for DFs examined in the present work. These additions include: 1) a method for labeling of overlapping or “complex” DFs; 2) integration

of labeling of filled pauses; 3) integration of labeling of speech errors; and 4) explicit conventions for handling ambiguous DFs, or those having more than one conceivable analysis. Details are provided in Section 4.3.4.

4.3.3 PLS in summary

Tables 2, 3, and 4 provide a quick reference guide to the PLS. Sections containing further information are noted in the "Section" column of each table. For convenience, the organization of the detailed description of the system follows that outlined in the summary tables.

Each DF in the PLS is delimited by "[ ]". The IP is marked by ".". Each word in the DF is then labeled using the symbols in Table 2. Each DF is corrected by two unordered operations, as indicated in Table 3. Special cases are handled by the PLS as indicated in Table 4.

Table 2: PLS in Summary: Pattern Symbols

Symbol	Explanation	Example	Section
Region-delimiting			
[ ]	onset RM, offset RR	(see all examples below)	4.3.4.1.1
.	IP	(see all examples below)	4.3.4.1.2
Syntactic-word			
r	repeated word	she she liked it [r . r]	4.3.4.2.1
s	word in substituted string	she my wife liked it [s . s s]	4.3.4.2.2
i	inserted word	she liked really liked it [r . i r]	4.3.4.2.3
d	deleted word	it was very she liked it [d d d .]	4.3.4.2.4
Extra-syntactic-word			
f	filled pause	she uh liked it / she uh he liked it [.f] [s . f s]	4.3.4.3.1
e	explicit editing term	she sorry he liked it [s . e s]	4.3.4.3.2
p	discourse marker	she liked well she liked it [r r . p r r]	4.3.4.3.3
Inter-sentence-word			
c	coordinating conjunction	she saw it and and she liked it [c . c]	4.3.4.4.1
Diacritics			
-	word fragment	she li- he liked it [s r- . s r]	4.3.4.5.1
~	misarticulated word	shle she liked it [r~ . r]	4.3.4.5.2
^	contracted word	she'd she'll like it [r^s . r^s]	4.3.4.5.3
=	substituted-string fragment	she thought highly she liked it [r s s= . r s]	4.3.4.5.4

Table 3: PLS in Summary: Correction Operations

Operation	Explanation	Example	Section
RM Deletor	Deletes all words in reparandum. Relies on IP location. Does not rely on symbol type.	~~he likes~~ . sorry she uh likes [~~s r~~ . e s f r]	4.3.4.6.1
f and e Deletor	Deletes filled pauses and explicit editing terms anywhere in the DF. Does not rely on IP location. Relies on symbol type.	he likes . ~~sorry~~ she ~~uh~~ likes [s r . ~~e~~ s ~~f~~ r]	4.3.4.6.2

Table 4: PLS in Summary: Special Cases

Case	Definition and Example	Solution	Section
Serial DFs	2 or more DFs are adjacent (do not overlap) Example: "he he liked uh loved it"	Treat as consecutive, individual basic DFs. he he liked uh loved it [r . r] [s . f s]	4.3.4.7
Complex DFs	2 or more DFs overlap. Example: "he she she liked it"	Treat as a complex of basic DFs in a hierarchical representation. See section for notational conventions. he she she liked it [R[s . s] . r]	4.3.4.8
Ambiguous DFs	DF has more than one possible analysis. See section for examples.	Apply 2 rules: 1) delete fewer over more words 2) assume less over more correspondence	4.3.4.9
Degenerate Cases	Labeling is impossible or inappropriate. See section for examples.	Flag cases but exclude them from analyses.	4.3.4.10

4.3.4 PLS in detail

4.3.4.1 Region-delimiting symbols

4.3.4.1.1 '[ ]' Bounds of DF

These symbols bound the DF by marking the onset of the RM and offset of the RR, respectively. Although these symbols are redundant for isolated, basic (noncomplex) DFs, they are needed to indicate the extent of each DF for some serial DFs and for complex DFs (see below).

4.3.4.1.2 '.' Interruption point

The IP is indicated by a '.'. Note that the IP is based on surface form only. We do not know when the speaker actually detected trouble, as discussed in Chapter 2. The RM consists of all material from '[' to '.' The IM is delimited in the full pattern by any extra-syntactic symbols ('f', 'e', or 'p') that occur after the IP and before any syntactic or inter-sentence symbols. The RR extends from the last of any post-IP 'f/e/p' to ']'.

4.3.4.2 Syntactic-word symbols

“Syntactic” words are defined for the present purposes as any words that play a role in the formal syntax of a sentence. These are contrasted to “extra-syntactic” words, which break from or are not part of the formal syntax, and to “inter-sentence” words, which are outside the scope of the sentence syntax.

4.3.4.2.1 'r' Repeated word

The symbol 'r' is used for any words in a DF that are exactly repeated before and after the IP. Pre-IP 'r's match up one-to-one, left-to-right, with post-IP 'r's. That is, no explicit coindexing is needed:

```
[rr.rr] = [r1r2.r1r2]
```

4.3.4.2.2 's' Word in substituted string

Each word in a substituted string is indicated by an 's'. Variable-length strings are necessary, since as described earlier, substitutions may not match up word-for-word, e.g.:

```
does United flight 201 serve  a sn(ack)-    breakfast
                            [s    s-   .    s]
```

Substituted strings must show syntactic and semantic correspondence. In a case such as the following, for example, “on” is not viewed as a substitution for “to” because the semantic selectional restrictions on “to” are different from those for “on.”

```
flights from boston to on friday
```

That is, we would not posit the speaker as having started to say the semantically nonsensical “flights from boston to friday.” (This particular example is handled as simply a deletion of “to.”)

In the absence of coindexing of corresponding words before and after the IP, it might seem unclear how multiple and adjacent variable-length s-strings within one DF could be correctly matched up. The answer is that empirically, multiple, adjacent substitutions do not seem to occur within the same DF. In fact, only one substituted string seems to occur per DF. This is an interesting finding and may have cognitive implications, i.e. that speakers do not change more than one thing at a time.

Exceptions to this finding, however, occur as an indirect consequence of grammatical phenomena that the PLS is not designed to handle. These include such changes as anaphora, and changes in tense, number, definite article form. For example, in the sentence:

```
our dog likes    he loves the beach
 ?  ?    s   .  ?   s
```

there is a true substitution (of “loves” for “likes”); however the additional substitution (“he” for “our dog”) is a case of anaphora that arises because the noun is repeated when the verb is changed. In the future, integration of syntactic information in the PLS could allow for better representation of such cases. For the present work, the cases were simply labeled as “s” since this was unimportant to analyses.

4.3.4.2.3 'i' Inserted word

Words inserted in the RR are marked with 'i':

```
please give me fares   round trip fares from pittsburgh
                [r .   i     i    r]
```

A notable difficulty in labeling insertions is that there is a fuzzy distinction between insertions that modify the propositional content, and insertions that represent a shift in level of language use, such as a parenthetical aside. Using syntactic information in the labeling process could help solve this problem. For the present purposes, inserted material that constituted a full clause was not labeled using 'i' but rather as an “aside.”

Note that in the present work there are no DFs consisting of *only* inserted words, i.e. cases in which modifying material is added as an afterthought (e.g., Carletta et al., 1993). Such cases contain no words in the RM or IM, and were declared beyond the scope of DFs covered (see Chapter 1) because they could not be reliably identified as disfluent.

4.3.4.2.4 'd’ Deleted word

Words deleted from the RM and having no corresponding words in the RR are indicated by 'd'. Deleted words most commonly occur at the beginning of a sentence; this has led to a DF type called “fresh start” or “false start” in many classification systems. However, as discussed in Chapter 2, deleted words may appear sentence-medially as well. For example, in:

```
flights from boston to    on  friday
                    [d.]
```

the word “to” appears to be the first word in an abandoned constituent; this constituent is not replaced by “on friday” because of the semantic correspondence requirement for substitutions; therefore the best we can say about “to” is that it is deleted.

4.3.4.3 Extra-syntactic-word symbols

Extra-syntactic words are those that do not play a role in the formal syntax of a sentence, but rather serve a discourse purpose. Three types of extra-syntactic words are distinguished by the PLS: 1) filled pauses (which have unrestricted distribution, and no semantic content); 2) explicit editing phrases (which are restricted to occur in the IM, or in some cases, just after the RR of DFs, and which have semantic content); and 3) discourse markers (which have a wider distribution than explicit editing phrases and which have been described as contributing “meaning,” unlike filled pauses; e.g. Schiffrin, 1987).

These classes were not distinguished in the SRI system, but are distinguished in the PLS because they are treated differently by the PLS correction operations. Filled pauses are deleted when found, and are labeled even if they do not occur in a longer DF (this was not the case in the work the SRI system was developed for). Editing phrases, like filled pauses, are deleted in DFs, but unlike filled pauses are not marked as DFs if they occur in an otherwise fluent stretch of speech. Discourse markers are unspecified for deletion, because whether or not a discourse marker is part of the intended utterance depends on the perspective of the researcher.

4.3.4.3.1 'f' Filled pause

Filled pauses (“uh” and “um”) are represented by 'f'. An issue concerning filled pauses is how to treat them when they occur inside a longer DF (whether in the IM or elsewhere) and how to relate this treatment to that of filled pauses in an otherwise fluent stretch of speech.

A filled pause occurring in an otherwise fluent stretch of speech is treated as a DF by the PLS:

```
show flights arriving in   uh   boston
                          [.f]
```

By convention, a '.' is placed before the filled pause; however, this is only for the purpose of aligning the filled pause with an identifier in the transcription. The '.' is not needed for filled-pause correction, because filled pauses are deleted wherever they are found, i.e. regardless of their location within a DF (see Section 4.3.4.6).

A filled pause occurring within the extent of another DF is not counted separately, regardless of the location within the longer DF at which the filled pause occurs. Although most filled pauses in longer DFs are found in the IM, the PLS makes no distinction between a filled pause in the IM of a longer DF, and a filled pause elsewhere. All are simply marked as 'f' within the longer pattern:

```
show flights arriving in   uh arriving in boston
             [r     r .  f    r     r]
```

```
show flights arriving uh in  arriving in boston
            [r      f  r .   r     r]
```

```
show flights arriving in  arriving uh in boston
             [r     r  .   r    f  r]
```

Note that this convention represents an appropriate, theoretically neutral solution. An alternative solution (which represents the implicit view in some past accounts) proposes that filled pauses that occur in the IM of a longer DF function as editing phrases, whereas those occurring elsewhere function as hesitations. There is to date no hard evidence for this view. It is a theoretically interesting possibility, but a rather intractable question unless independent evidence discriminating filled pauses of the two types can be found.[1]

By convention, filled pauses occurring in direct sequence are counted as one DF; those occurring outside the bounds of a longer DF are counted as a separate disfluency, even if adjacent to the longer DF. These conventions were adopted in order to consistently count DFs. The conventions do not preclude reanalysis; all information needed for reanalysis can be obtained automatically from the HLD.

4.3.4.3.2 'e' Word in explicit editing phrase

Each word in editing phrases such as “no,” “oops,” and “i'm sorry” is indicated by 'e'. These elements occur almost exclusively in the IM and explicitly refer to what was just previously said. Certain editing phrases (e.g., “i mean”) can also be used as discourse markers; in most cases the usage can be discriminated from context.

Occasionally, editing phrases are postposed, appearing after the RR (in many cases there is also an editing phrase at the IM),[2] e.g.:

```
list flights to boston    i mean to denver rather
          [r   s    .  e  e   r   s     e]
```

In these cases the editing phrases were indicated in both locations in the labeled pattern.

4.3.4.3.3 'p' Word in discourse marker

Words in discourse markers (Schiffrin, 1987; Redeker, 1991) such as “well,” “you know,” and “like” are indicated by 'p' (for “pragmatic expression, since 'd' is used for “deletion”) when they occur within the bounds of a DF. For example:

---

1. Shriberg and Lickley (1992b) suggest that intonation and pitch range may help discriminate these two proposed functions of filled pauses, but these studies remain at a preliminary stage.
2. Interestingly, these double editing phrases appear to be particularly common in corrections of read (e.g. newscaster) speech, as noted informally by the author.

```
they you know they really wanted to win
 [r  .  p   p      r]
```

As pointed out by Lickley, it is not clear whether these elements, when they appear in the IM (the most frequent region for discourse markers) function as editing expressions relative to the DF, or whether they simply contribute the same type of discourse-relevant meaning to the following material in the RR as they would in a fluent stretch of speech. This was left as an open question in the present work. Discourse markers very frequently occurred with no DF present, particularly in the SWBD data. In these locations they were not labeled as DFs.

4.3.4.4 Inter-sentence-word symbol

4.3.4.4.1 'c' Coordinating conjunction

Predating a theory of DFs, it is not clear whether coordinating conjunctions that join sentences, such as “and,” function differently than words internal to a sentence. Distributionally and intonationally, these inter-sentence elements seem similar to filled pauses and discourse markers. Indeed various authors have noted anecdotally that these elements may function as productive hesitation devices (e.g. Blankenship & Kay, 1964). In order to be able to separate out repeated or changed coordinating conjunctions from other repeats and substitutions in analyses, the coordinating conjunctions are denoted in the PLS as 'c'. These can be automatically collapsed back to 'r' or 's' if desired using information in the HLD. Usages not associated with sentence coordination (for example, simple noun phrase coordination) are labeled using the 'r' or 's' symbols.

4.3.4.5 Diacritics

4.3.4.5.1 '-' Fragment

Word fragments (as described in Chapter 3) are labeled with the appropriate correspondence symbol, followed by “-”. For example:

```
show fli-  flights to boston
     [r-   .  r]


show fli-  fares to boston
     [s-   .  s]
```

4.3.4.5.2 '~' Misarticulated word

The PLS is not intended for detailed analysis of DFs involving misarticulations, including speech errors, which represent a fundamentally different phenomenon from the other DFs studied here. However, annotation of corrected misarticulation errors was integrated into the PLS by using the diacritic '~' immediately following the word symbol to indicate a misarticulated word. That is, regardless of the way in which a word in a speech error was misarticulated, it was annotated as a misarticulated version of the word intended for that location, e.g.:

```
show me grand trouns-      ground transportation
         [r~    r~-    .   r          r]
```

Note that the '~' can combine with '-' as in the example above.

This convention allows locations of speech errors to be recovered, and provides minimal information about the form of the error. However, the PLS representation is not ideal for focussed work on misarticulation errors. It does not distinguish among different types of misarticulation. It also has no provision for indicating the significance of particular words in the surface utterance for the misarticulated word in a speech error. For example, there is no way in the PLS to attach significance to “june” in the speech error below:

```
show flights from jen-     from denver on june tenth
            [r   r~-   .   r    r ]
```

Note, however, that these limitations are consistent with the goal of the PLS. The PLS makes no claims about what gives rise to a particular DF; it simply represents observable surface errors and the corresponding corrections.

4.3.4.5.3 '^' Contracted word

Individual treatment of words in a contraction is necessary for consistency with other aspects of the PLS, since the PLS labels each word, and individual members of a contracted form may bear different relationships to corresponding words on the other side of the IP. For example:

```
she'd    she'll  like it
[r^s  .  r^s]
```

```
she'd    he'd  like it
[s^r  .  s^r]
```

The '^' allows contracted forms to be represented as one word for analyses where this is relevant (for example, in prosodic analyses). In addition, the '^' diacritic provides a useful solution for labeling frequently-occurring cases of the form:

```
she    she'd  like it
[r  .  r^]
```

Since “she” is not exactly repeated, alternatives to the labeling above are either to label the correspondence as [r-.r] or [s.s]; neither of these seems as appropriate as [r.r^].

4.3.4.5.4 ’=’ Substituted-string fragment

The '=' diacritic is used for cases in which a substitution is strongly suggested but the first string is one or more words short, e.g.:

```
um i guess we're going to talk   describe uh job benefits
                           [s= .   s]
```

Here the hypothesis is that “describe” was substituted for “talk about” or some other similar phrase consisting of material after “talk” that can function syntactically like “describe.”

A second example illustrates a substitution involving rephrasing of the substituted constituent:

```
we could spend that money you know

for star-     children that are starving
    [s-=   .    s      s   s     s]
```

The hypothesis here is that the speaker began to utter “starving children,” but stopped short at “starving” to rephrase the noun phrase. Because “starving” is cut off mid-word, it also receives the fragment diacritic.

4.3.4.6 Correction operations

DF correction in the PLS is formalized as involving application of two unordered operations. These operations are distinct only definitionally, because they affect different pattern symbols. Since operations are unordered, they may alternatively be viewed as a

single operation. No significance in terms of cognitive processing of DFs is attached to the operations.

4.3.4.6.1 RM Deletor

This operator deletes all words in the RM.

4.3.4.6.2 'f' and 'e' Deletor

This operator deletes all 'f' and 'e' symbols wherever they are found. Although typically these elements are in the IM, restricting their deletion to the IM is insufficient, since 'f' may occur in an otherwise fluent region or within an RM or RR, and 'e' may occur following the RR, as described earlier.

Note that absence of reference to the IM in defining the deletion of filled pauses fits nicely with the neutral view of filled pauses discussed earlier. Under this neutral view, there is no editing function necessarily attributed to filled pauses occurring in the IM of longer DFs. These filled pauses are deleted just like filled pauses in all other locations, including filled pauses in an otherwise fluent region.

As mentioned in Section 4.3.4.3.3, it is left as an open question whether 'p' words (discourse markers) are to be removed in DF correction. This will depend on the perspective of the researcher as to whether discourse markers should be retained as part of the speaker's intended utterance.

4.3.4.7 Serial DFs

Two or more DFs are in a 'serial' relationship if the last word in the first DF immediately precedes the first word of the next DF (i.e. there are no shared words in these DFs). For example:

```
a)where is  the  the   uh  stopover
            [r  .  r]  [.f]


b)which  how much    how much does that cost
   [d.] [r    r    .   r     r]
```

Case (a) above illustrates the treatment of filled pauses as described earlier: filled pauses are labeled as a separate DF if they occur outside the boundary of any syntactic words in a DF.

Case (b) illustrates serial DFs in which the first DF contains no words in the RR region. It is worth remarking that according to the PLS, a DF with no words in the RR region is in at most a serial relationship with a following DF; it can never be in a complex relationship (see below) with a following DF because it contains no words in RR to potentially overlap with a following DF.

Serial DFs are treated no differently than other basic DFs (those separated from other DFs by fluent surrounding speech) in the analyses in this thesis. However, the construction of the HLD allows for special inspection of such cases if so desired, since locations of adjacent word indices in consecutive DFs can be recovered automatically.

4.3.4.8 Complex DFs

Complex DFs are cases in which two or more IPs bound material that corresponds to both preceding and following material. For example:

```
he -- she -- she went
```

The material between the two IPs, i.e. the first “she,” corresponds both to the preceding “he” (as a substitution) and to the following “she” (as a repetition).

4.3.4.8.1 Compositional approach

The question is how to represent these cases. One option is to use a flat representation, treating each case as a single DF with multiple IPs. The alternative is to analyze each case as composed of individual, overlapping DFs. The compositional analysis was chosen for the following reasons.

First, a flat analysis is unparsimonious in that a single symbol cannot be assigned to the middle word; new symbols would need to be created, resulting in a proliferation of symbols and of patterns. Second, a flat labeling would require alternative formalizations of correction operations for complex DFs. Third, empirically it was found after labeling a large number of basic DFs that complex DFs could be broken down into basic “component” DFs. That is, the component DF patterns in these analyses were familiar patterns seen in the labeling of basic DFs. And fourth, an empirical analysis of complex DFs in Chapter 6 suggests that if overlapping DFs are analyzed as complexes of basic DFs, rates of

component types in the complex DFs can be predicted based on rates of those types seen for basic DFs.

These arguments based on system parsimony and empirical findings clearly favored the compositional approach. It should be noted, however, that the compositional analysis is not intended as a cognitive model for production or processing of complex DFs (although this is an interesting area for future work). It is strictly a formalism to allow the PLS coverage of these cases.

4.3.4.8.2 Formalism

A formalism was needed for the representation of compositional structure. For convenience, a notation using labeled brackets was adopted, similar to that used to represent parse trees in a flat representation. Note however that the structures represented are *not* parse trees, but rather analyses of the relationships among component DFs.

Each DF is enclosed in brackets, and correction of DFs proceeds outward from "lowest" to "highest" DF (see corresponding tree structure below). For each DF, as for basic DFs, the normal correction operations apply. Correction includes deletion of the RM, and deletion of any 'f' and 'e' symbols anywhere in the DF. The output of correction is placed immediately to the left of the corrected DF. However, the *symbols* used when labeling the output of a DF are determined by the role of the output of that DF as *input* to the next DF. For clarity, outputs of correction of a lower DF are written in uppercase letters. This allows for easy visual matching (as well as automatic matching) of symbols to the original words in the labeled-bracket notation; i.e. the original words in the DF match up left to right with the *lowercase* symbols. For example:

```
   he  she  she
[R[s . s] . r]
```

Each bracketed structure corresponds to a tree structure, as illustrated in Figure 2, in which the IP for each DF corresponds to a node in the tree. Nodes are binary-branching, with one branch extending to the RM, and the other to the IM+RR. The output of correction of a DF is written above the node for that DF, where it corresponds to the left branch for the next higher DF.

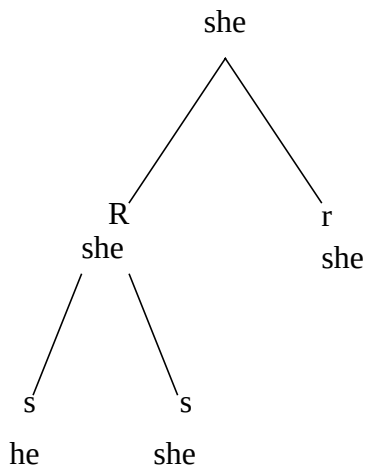


Figure 2. Hierarchical Representation of Complex DFs

4.3.4.8.3 Completely chained structures

When a lower disfluency is not fully contained within an upper one, the relationship is one of chaining. This was the case in the “he she she” example. There is an interesting ambiguity in the analysis of chained complex DFs. They can be analyzed as left branching or right branching:

```
left: [SHE[he.she].she]  [ R [ s .s ]. r ]

right:[he.SHE[she.she]]  [ s . S[ r . r] ]
```

In cases of ambiguous branching direction, a left-branching analysis was always preferred over a right-branching analysis. At present there is no theoretical significance attached to this decision; however, right branching can yield unusual basic disfluency patterns as well as incorrect accentuation. This is illustrated in the example below, where accented syllables are indicated in boldface type.

show me the ***flights*** -- ***delta*** flights -- delta ***fares***

A left-branching analysis yields the two component DFs:

lower DF: ***flights*** -- ***delta*** flights

upper DF: ***delta*** flights -- delta ***fares***

These are both reasonable accentuation patterns for the particular basic DFs. A right-branching analysis, on the other hand, yields:

lower DF: ***delta*** flights -- delta ***fares***

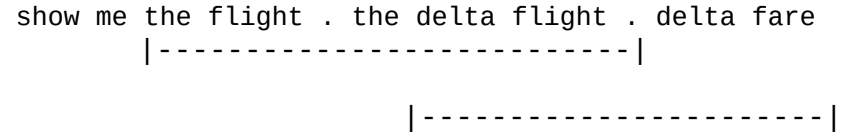
upper DF: ***flights*** -- delta ***fares***

in which the upper DF has both an unusual disfluency pattern (an insertion combined with a replacement) and quite noticeably, unexpected deaccentuation of the inserted content word.

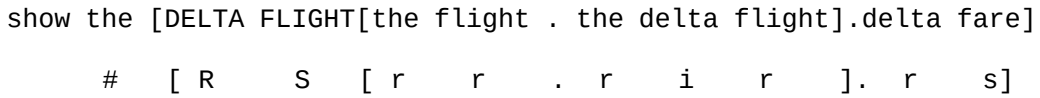
In the complex DFs labeled, any cases of truly ambiguous branching direction were stipulated as left-branching. However, at this point, no theoretical significance is attached to the preference for left-branching structures. It is possible that different branching directions are required for different disfluency patterns, or even for the same disfluency uttered in different contexts.

4.3.4.8.4 Partially chained structures

Partial chaining occurs when some, but not all, of the output of one disfluency is the input to another disfluency as in:



The problem here is that “the” is retraced in the (temporally) first DF, but not in the second. A left-branching analysis yields:



That is, the word “the” as output of the first DF has become a part of the fluent utterance; it is not carried over as input to the second DF. It is represented by “#” to indicate it is merely a word in the fluent portion of the sentence at the level of the analysis of the second DF.

Note that although at first glance it may seem preferable to solve this using a right-branching structure, such an analysis suffers the same problems as discussed above for fully-chained cases. That is, a right branching analysis yields, as the upper disfluency:

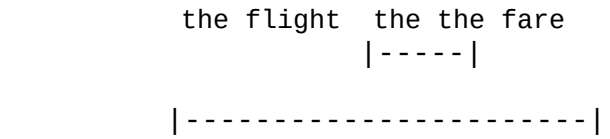
flight -- delta ***fare***

which is both a poor prosodic pattern (since the new, inserted content word “delta” is unaccented) and an unusual DF pattern (since it contains both an insertion and a substitution; see Chapter 6).

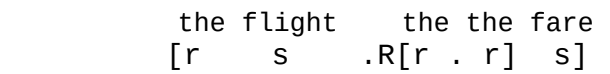
The left-branching analysis was therefore preferred, since there is no reason to assume that a complex disfluency is entirely encapsulated from surrounding speech until the output of the final disfluency correction,[3] and since this affords a convenient consistency in branching direction.

4.3.4.8.5 Nested structures

In nested complex DFs, one DF is entirely contained within another DF, such that one or more words of the containing DF surround the contained DF. For example:



Branching direction is not an issue for nested DFs. The contained DF must always be represented as the lower DF:



[3] However, there are likely to be constraints on allowing words to drop back into the fluent portion of the sentence; notably, a constraint that lines in a tree representation do not cross (i.e., that words drop back in a left-to-right, consecutive order, without trapping any words to be deleted in higher DFs between them). Discussion of the reasons behind this is beyond the scope here, but it is worth noting that this constraint was obeyed in the cases examined.

4.3.4.8.6 Binary vs. N-ary branching structures

This section, the last section on complex DFs, discusses the issue of using only binary versus allowing N-ary branching structures. It may be noted that for complex DFs of the form:

```
we    we   we
 r  .  r  .  r
```

i.e. cases involving homogeneous symbols, it may seem intuitively less appropriate to impose binary-branching structure (and hence, a hierarchical representation) when the DF could be represented using a flat N-ary tree, in which the last branch replaces all earlier ones. This feeling may be particularly strong for cases like the following (an example of repeated attempts to call one of three daughters):

```
li-   kath-  janet !
 s  .  s  .  s
```

since the intuitive analysis is that both “li-” and “kath-” are failed attempts at “janet” (rather than “kath-” a substitution for “li-”). However, despite such intuitions, the PLS uses only binary-branching structures. This is for both formal and empirical reasons. Formally, it is preferable not to introduce intuitive or theoretical biases into the representation, which is intended solely for classification. And, N-ary branching would add complexity to the formalization of correction operations. No information is lost, since binary-branching representations can be rewritten as N-ary if necessary. Empirically, there is evidence that treating cases with homogeneous symbols in the same manner as other cases is appropriate. In the analysis of complex DFs in Chapter 6, it is found that rates of homogeneous complexes can be predicted in the same manner as rates of nonhomogeneous complexes.

4.3.4.9 Ambiguous DFs

Ambiguity is defined not at the level of *labeling* but rather at the level of analysis. A DF is ambiguous if after using all available speech information (prosody, contextual, etc.), there is more than one conceivable *analysis* of the DF. Consistent with the goals of the labeling system, if there is more than one analysis for a DF, there will also be more than one labeling for that DF, since the labeling system was designed to distinguish alternative analyses.

“Analysis” of a DF refers to: 1) the determination of the extent of the RM region; or 2) correspondences of words in the RM region; or 3) both (1) and (2). Note that this kind of ambiguity is different from the type of ambiguity discussed in Chapter 3. In Chapter 3, ambiguity referred to uncertainty in the transcription of words in a DF; here, ambiguity pertains to analysis, given a particular transcription.

It is important to specify how ambiguous DFs were labeled because such cases occurred fairly frequently. It is also worth mentioning that the types of ambiguity encountered are not specific to the PLS, but are likely to be problematic for other systems as well. Ambiguous cases were handled using two conventions:

1. delete fewer words over more words
2. assume less correspondence over more correspondence

The first rule is consistent with the work of Bear et al. (1992) on automatic DF correction, where it was considered less desirable to delete words *not* intended to be deleted than to fail to delete the words intended to be deleted. However, this rule is the most consistent possibility more generally, since in a large number of situations, more material can always be (inappropriately) deleted, but deletion of less material has a natural lower bound. Note that the first rule dictates that in cases of ambiguity between a fluent and a disfluent reading, the fluent reading is chosen, since it involves deletion of no words. The second rule is consistent with a pretheoretical, data-driven approach. Symbols denoting correspondence between the RM and RR are 'r' and 's'; those not denoting correspondence are 'd' and 'i'. Application of these two rules determined a unique outcome for all cases of ambiguity.

4.3.4.10 Degenerate cases

A relatively small number of DFs were considered degenerate with respect to the PLS. DFs were degenerate if: 1) signal quality was degraded to the point of making words relevant to labeling indeterminate; or 2) the DF was beyond the scope of DFs considered by the PLS; or 3) the case in question did not seem appropriately labeled as a DF.

Cases labeled degenerate due to degradation of the signal involved unintelligible speech, waveforms truncated or degraded in recording, and regions in which background noise impaired the ability to determine what was being said. Cases representing phenomena beyond the scope of the PLS included uncorrected speech errors, uncorrected prosodic repairs, and DFs involving noncontiguous regions. Cases for which labeling as a DF seemed

inappropriate involved externally-induced interruptions (such as a cough, sneeze, or throat-clearing by the speaker), full-clause insertions (e.g. parenthetical asides); and cases of final ellipsis (in which a sentence was left incomplete because the ending was clear from context; i.e. the sentence was not meant to be deleted).

Degenerate cases were tallied, but excluded from all analyses. Total degenerate cases within the HLD are listed in Table 6 (Section 4.5); as can be seen, these cases comprised a small proportion of the data. However, as might be expected, relative proportions of degenerate cases were quite different for the different corpora. For example, over half of all degenerate cases in ATIS were due to machine truncations; over half in SWBD were cases of either “aside” or “final ellipsis.”

4.4 Type Classification Algorithm (TCA)

4.4.1 Goal and requirements

The PLS described in the previous section produces a large number of unique patterns, in keeping with the requirements in that section that DF representation should be flat, with no hierarchical groupings imposed. However, in order to analyze differences across DFs, a method was required for grouping patterns into a smaller set of classes. There are of course many different ways to do this, depending on the question at hand. This ability to regroup raw patterns in different ways is a strength of the current approach.

The goal of the particular method developed for grouping patterns, the “Type Classification Algorithm” (“TCA”), was to define a small number of classes that would roughly capture the kinds of distinctions used in previous classification systems. As noted in Chapter 2, most previous systems have made reference to the correspondence in wording between material in the RM and the RR. The TCA aimed to reflect similar distinctions by focusing on the PLS symbols denoting the nature of correspondences between words in the RM and those in the RR.

For simplicity in analyses, the TCA was required to produce orthogonal classes, and to have complete coverage for the set of basic, non-degenerate DFs labeled by the PLS. In addition, since speech errors were taken to represent a fundamentally different type of phenomenon from other DFs examined, the TCA was required to separate out all speech errors as a separate class.

4.4.2 Algorithm

Given the constraints above, an algorithm based on a natural ordering of relevant symbols was used. “Ordering” here refers simply to the precedence of symbols in determining the DF type. Relevant symbols in this algorithm are: 1) symbols involving correspondences between the RM and the RR (i.e.: r, s, i, d and c); 2) the misarticulation diacritic (~) for separating out speech errors; and 3) the filled-pause symbol 'f', because DFs could contain only a filled pause. Remaining symbols (e.g. '-', 'e', 'p') are irrelevant to the algorithm, because they occur independently over the presence of the relevant symbols. Note that 'f' is included, but not 'e' or 'p'; this is because filled pauses in an otherwise fluent sequence are included as DFs by the PLS, whereas this is not true for 'e' or 'p'.

The ordering is shown in Figure 3. The first symbol is '~,' to assure that all DFs

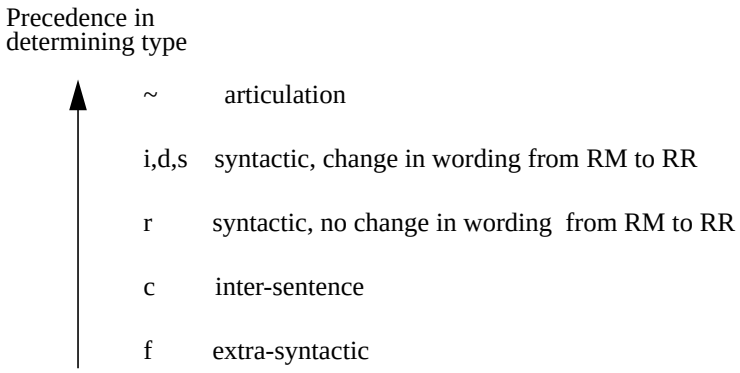


Figure 3. Ordering of Symbols for Determining Type

involving misarticulation are separated out. The ordering of remaining symbols may be informally viewed as representing something like the severity of the change in wording from the RM to the RR. However, no theoretical significance is attached to the notion of severity at present. It is not intended, for example, to reflect differences in the ability to process DFs.[4] Following the '~' symbol are symbols involving any type of change in wording from RM to RR. No distinctions are made within this level, since there is no basis within the PLS upon

[4.] The relative ordering of specific symbols does happen to be consistent with results from perceptual experiments (e.g. Deese, 1980; Fox Tree, 1993); however, significance of the ordering for cognitive models remains an issue for future work.

which to make such a distinction. Next in line is the 'r' symbol, which plays a role in the syntax of a DF, but involves no change from RM to RR. The 'c' symbol follows, with less precedence than 'r', because 'c' occurs outside the sentence syntax. Finally, 'f' is included as the symbol with lowest precedence, since it plays no role in the formal syntax of an utterance.

Using this ordering, the TCA produces eight orthogonal classes, referred to as DF "types" as indicated in Table 5. These types map very roughly onto the canonical set of types used in previous work (see Chapter 2). The algorithm makes reference only to the presence of the relevant symbols, not to their frequency, temporal ordering, or location with respect to the IP. As shown in Table 5, types are named for the symbol they are determined by. Note that the terms for types (an uppercase two- or three-character abbreviation) are distinct from the individual pattern symbols. For example, REP corresponds to any complete pattern containing at least one r, with optional 'c' and 'f' symbols, whereas 'r' indicates a repeated word within a pattern. Note also that because 's', 'i', and 'd' have the same degree of precedence, any patterns containing more than one of these symbols in the pattern are of type HYB (for hybrid). The TCA was encoded in a simple awk script, and each DF was automatically annotated as one of the eight types, based on its pattern representation.

4.5 Hand-Labeled Database (HLD)

This section describes the result of applying the methods described in the three sections above to the set of selected speech data introduced in Chapter 3. The resulting database is referred to as the "hand-labeled database" or HLD. This database forms the basis for the analyses in Chapters 5 and 6.

4.5.1 Summary statistics

Table 6 shows summary statistics for the HLD. The "total words" measure counts fragments and filled pauses as words. Under "total speakers," the total number of speakers having at least one DF is listed for each corpus. Note that it is not meaningful to compare percentages of speakers who made DFs across corpora, because of the large differences in the average amount of data per speaker in the three different corpora.

"Fluent sentences" are those containing no DFs of the types labeled by the PLS; "DF sentences" are those containing at least one DF. "Total DFs" are divided into basic (i.e. noncomplex) and complex DFs. Each complex DF was counted as a single DF (but analyzed

Table 5: Type Classification Algorithm

TYPE	Must Include	Must Not Include	May Include	Examples
ART	~		s i d r c f	shle she liked it [r~.r] i'd like to fry uh to fly from boston [rr~.frr] show grand trou- ground transport [r~r~-.rr]
HYB	2^+ from: s i d	~	r c f	(example of s + i): she liked he really liked it [sr.sir]
SUB	s	~ i d	r c f	she he liked it [s.s] she uh he liked it [s.fs] and she liked and he liked it [csr.csr]
INS	i	~ s d	r c f	she liked really liked it [r.ir] she liked uh really liked it [r.fir] and she liked and she really liked it [crr.crir]
DEL	d	~ s i	r c f	it was she liked it [dd.] it was um she liked it [dd.f] and it was and she liked it [cdd.c]
REP	r	~ s i d	c f	she she liked it [r.r] she liked uh she liked it [rr.frr] and she and she liked it [cr.cr]
CON	c	~ s i d r	f	she saw it and and she liked it [c.c] she saw it and uh and she liked it [c.fc]
FP	f	~ s i d r c		um she liked it [.f] she uh liked it [.f]

as composed of multiple overlapping DFs as described above.) Within the set of basic DFs, DFs were either non-degenerate, or degenerate.

"Clean DF sentences" are those containing no complex DFs and no degenerate DFs. This set was necessary, because both complex and degenerate DFs posed problems for certain feature annotations. For example, in the case of complex DFs, it is not clear how to

Table 6: Summary Statistics for Hand-Labeled Database (HLD)

Information	ATIS	AMEX	SWBD
Total Words	188,327	12,762	40,515
Total Speakers	523	66	30
- Speakers with DFs	356	65	30
Mean (St.Dev) Words / Speaker	359 (221)	193 (131)	1350 (694)
Total Sentences	18,675	1,821	4,583
- Fluent Sentences	17,256	1,332	3,112
- Disfluent sentences	1,422	489	1,471
Total DFs	1,694	745	2,586
- Basic DFs	1,574	672	2,320
- Not degenerate	1,491	661	2,186
- Degenerate	83	11	134
- Component DFs in Complex DFs	120	73	266
"Clean" Disfluent Sentences (no complex or degenerate)	1,227	423	1,228
Total DFs in Clean DF Sentences	1,457	594	1,934
Total Clean Disfluent + Fluent Sentences	18,483	1,755	4,340

count either “efficient” words or total DFs; in the case of degenerate DFs, no pattern features were labeled. Thus clean DF sentences were those for which all features in the HLD could be annotated for all DFs in the sentence. “DFs in clean DF sentences” is simply the total DFs contained in the set of clean DF sentences. Note that the majority of DFs were retained in this set.

4.5.2 Preparatory files

This section describes the procedures and formats used to create the preparatory files from which the database was automatically generated.

First, for each original turn-sized transcription, sentence boundaries and locations of IPs of DFs were hand-marked. This was done in the process of correcting transcriptions (as described in Chapter 3). For example, a single turn from a SWBD conversation about drug testing is shown below, with sentence boundaries ("==") and IPs of DFs ("/") inserted. The identifier in the first field indicates the SWBD conversation number, turn number, and channel:

```
2096-097-c1 i don't mind so much the fact that they test
people == but / uh not questioning the validity of the
results is / is a problem ==
```

A script converted these annotated turn transcriptions to a format containing one sentence per line. The script also assigned a unique, incrementing sentence identifier to each sentence, and replaced each slash with a unique, incrementing DF identifier:

```
SB-sent-07870 2096-097-c1 i don't mind so much the fact
that they test people


SB-sent-07880 2096-097-c1 but {44050} uh not questioning
the validity of the results is {44060} is a problem
```

Both sentence identifiers and DF identifiers were unique across the three corpora labeled, but the source corpus was explicitly noted in the sentence identifier (the "SB" in the sentence identifier above).

Another script operated on this one-sentence-per-line file, to produce a one-DF-per-line file, i.e. a file containing only those sentences having DFs, but listing each DF on a separate line. This was the file used for entering pattern labels for DFs. For instance, the example above was converted to:

```
$.f 44050 2096-097-c1 but ** uh not questioning the
validity of the results is {44060} is a problem


/r.r 44060 2096-097-c1 but {44050} uh not questioning the
validity of the results is ** is a problem
```

The first sentence does not appear, since it contains no DFs, and the second sentence appears twice since it contains two DFs. The first field was left blank for the insertion of the

pattern label (in the examples the pattern labels have been inserted). In addition to noting labeling the pattern according to the PLS, the DF was labeled as “sentence-initial” or “sentence-medial” by placing “$” or “/”, respectively, in front of the pattern. (For further details on this coding see Section 4.2.3. Note that the word “but” is skipped over in this coding as explained in that section). The identifier of the DF to be labeled was listed in the second field, and the original turn identifier in the third field. Remaining fields showed the transcription, with ** replacing the location of the identifier of the DF to be labeled (since sentences could contain more than one DF). This simple and consistent format proved extremely efficient for the process of DF labeling.

4.5.3 Automatic database generation

Tools were developed to use the files described in the preceding section as input, and automatically generate a database file containing the information shown in Table 7. The

Table 7: Contents of Database File

Field	Information
1	sentence (and corpus) identifier
2	DF index within sentence (1=1st DF, 2=2nd DF, ...)
3	total DFs in sentence
4	DF identifier
5	turn identifier
6	speaker identifier
7	pattern label
8	comment codes
9	total fields in transcription (including identifiers)
10	location of identifier for current DF
11	number of pre-IP symbols
12	pattern label for pre-IP symbols
13	words for pre-IP symbols
14	number of post-IP symbols
15	pattern label for post-IP symbols
16	number of post-IP symbols
17	transcription (in single-field format)

TCA operated as a post-processing script on this file, to automatically convert patterns to DF types.

The database file and a number of associated post-processed files make up the HLD. As described in Chapter 3, each of the transcribed turns in the HLD has an associated speech recording (digitized waveforms for ATIS and SWBD, and an audio tape recording for AMEX). The HLD forms the basis for the analyses in Chapters 5 and 6.

Chapter 5: Type-Independent Analyses

5.1 Chapter Overview

This chapter examines characteristics of DFs overall, without making reference to DF type. The set of data analyzed is that specified on the last line of Table 6 in the description of the Hand-Labeled Database in Chapter 4. This set includes the set of fluent sentences and the set of "clean" disfluent sentences[1] in each of the three corpora described in Chapter 3.

The chapter consists of five major sections. The first two sections examine rates of occurrence of DFs with respect to the total speech produced. Section 5.2 describes a model for predicting the likelihood that a sentence is fluent (contains no DFs) based on its length. Section 5.3 describes features affecting the likelihood of disfluency at each word. Relevant features include sentence length, word position, presence of another DF in the same sentence, and speaker.

The last three sections examine rates of pattern features in DFs overall. Section 5.4 presents a model for predicting the distribution of DFs having *k* deleted words. Section 5.5 compares rates of DFs containing a word fragment, overall and by DF position. Section 5.6 reports the frequency of DFs containing words in the interregnum--both overall, and by type of word (filled pause, editing term, or discourse marker).

Section 5.7 summarizes the main findings of the chapter, and discusses issues that cut across the chapter sections.

5.2 Rate of Disfluent Sentences

This section examines "disfluent-sentence" rate, or the probability of a disfluent sentence, in each of the three corpora. This measure of rate has a binary outcome. A disfluent sentence is defined as a sentence that contains one or more DFs; a fluent sentence is defined as a sentence containing no DFs. Section 5.2.1 proposes a simple model for predicting the

---

1. Disfluent sentences that were not "clean," i.e. those that contained complex DFs or degenerate DFs, were excluded because as explained in Chapter 4, these cases pose inherent problems for feature annotation.

probability that a sentence is fluent, given its length. Section 5.2.2 examines data from individual speakers to discover whether the behavior observed over speakers holds within speakers. Section 5.2.3 describes the rate of disfluent sentences overall in a corpus, taking into account the distribution of sentence lengths in the corpus.

5.2.1 Rate by sentence length

Figure 4 shows frequency histograms for fluent and disfluent sentences, by sentence length. Each curve connects points for frequencies at sentence lengths binned at four words; points are plotted at the center of the bin.

The curves shown with asterisks indicate the data for fluent sentences. The remaining four curves in each graph correspond to the set of disfluent sentences; these data are plotted four different ways. Separate curves indicate the disfluent data by total words and efficient words, where the latter measure does not count words deleted in the DF(s) in the sentence (see Chapter 4). For each of these two measures, the distributions are plotted using both absolute frequencies and normalized frequencies. Normalized frequencies equate the total disfluent sentences to the total fluent sentences, for better visual inspection of the shapes and locations of the distributions.

A number of observations can be noted from these distributions. First, comparing across the corpora, the shape of the fluent-sentence distribution for ATIS is roughly bell-shaped, whereas for AMEX and SWBD the fluent-sentence distribution peaks at one word and decreases rapidly with increasing length. Second, the absolute rate of DFs in ATIS is lower than that in AMEX or SWBD, as can be inferred from comparing the area under the unnormalized disfluent-sentence distribution to the area under the fluent-sentence distribution for each corpus. Third, in all three corpora, the disfluent-sentence distributions, including the distributions reflecting efficient length, are shifted rightward from the fluent-sentence distribution. This indicates that disfluent sentences are on average longer than fluent sentences, even after the words deleted in DFs have been removed. Fourth, in all corpora, the total-word distributions and efficient-word distributions for disfluent sentences show essentially the same shape; they differ by a fairly consistent horizontal shift (this is best seen in the normalized curves). This suggests that within the set of sentences known to be disfluent, the total number of words deleted does not change much over sentence length.

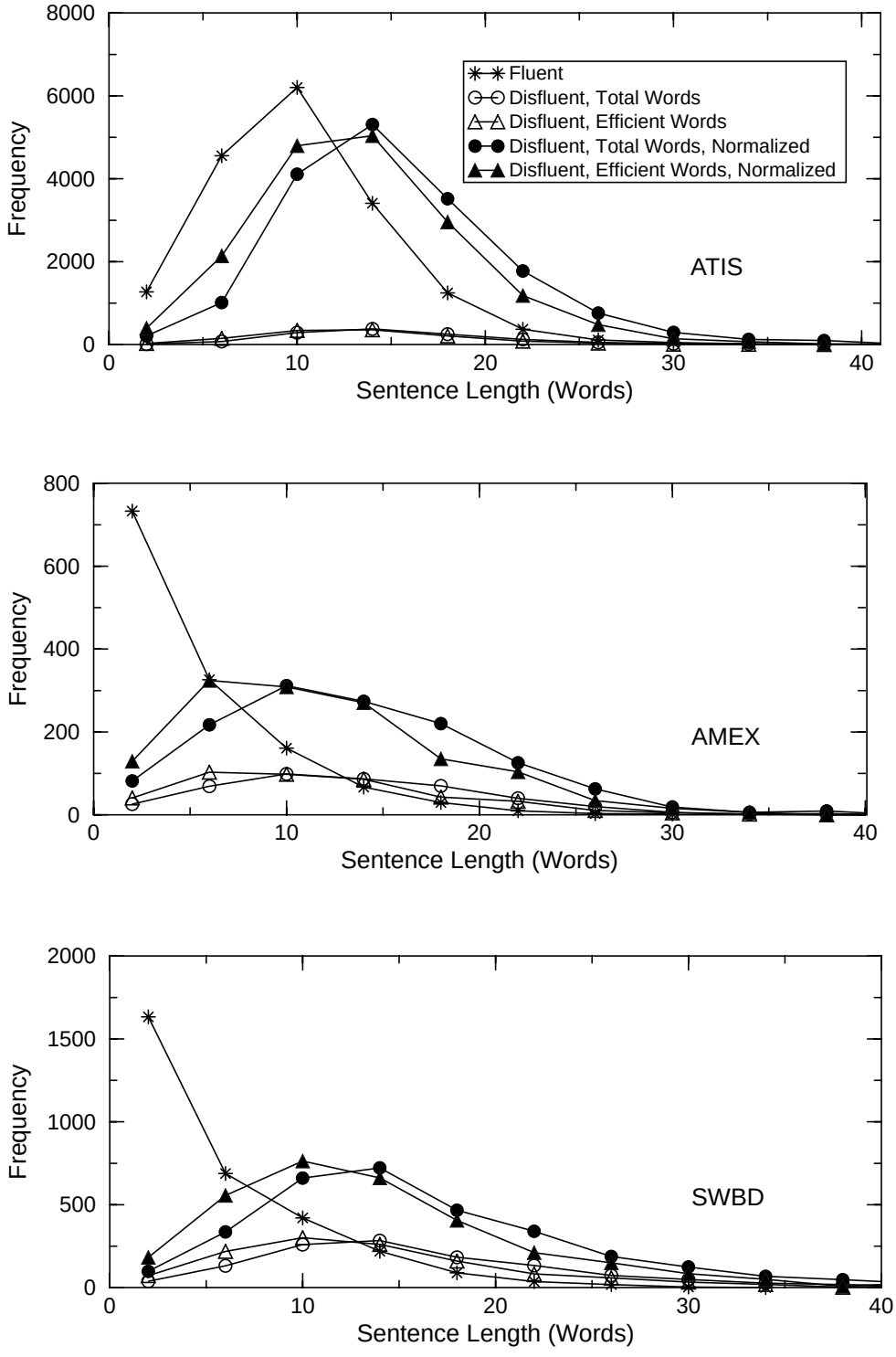


Figure 4. Distribution of Fluent and Disfluent Sentences by Sentence Length

The data in Figure 4 have been replotted in Figure 5 to indicate the probability of a fluent sentence at each length. The probability of a disfluent sentence is one minus the probability of a fluent sentence; the *fluent*-sentence measure is plotted because this simplifies the form of the models proposed later in this section.

Here and in all further analyses involving sentence length, the *efficient*-words measure is used, because there is an inherent correlation between presence of a DF and sentence length (since all DFs considered here contained one or more extraneous words; see Chapter 4). Efficient length avoids this confound by omitting words deleted in DFs from the word count for the sentence. It can be seen from Figure 4 that the pattern of results using total words would be similar to that using efficient words, since these distributions differ by a roughly constant horizontal shift. However, the efficient words measure is necessary for statistical analyses.

The points plotted in Figure 5 represent length bins of one word. A threshold of 20 total sentences was necessary for inclusion, although as can be seen from the histograms in Figure 4, most bins contained far more data than this. Error bars indicate the standard deviation for the binomial distribution at each length.

As shown, in all corpora, the probability of a fluent sentence decreases steadily with sentence length. This is not surprising, since we expect that additional words should increase the opportunity for disfluency. The question is how to model this rate of decay.

Before turning to the question of modeling the trends, however, an important observation should be noted from Figure 5. As the probability of a fluent sentence decreases, generally the error bars increase. This is chiefly attributable to a decrease in overall sample size,[2] as can be construed from Figure 4. Low sample size also accounts for the generally larger overall bars for the AMEX corpus. An noticeable discontinuity, however, occurs for all three corpora at sentence lengths exceeding about 15-20 words. Not only do the error bars increase rapidly in this region, but also there is considerably greater noise in the trends. This may indicate a qualitative difference in the nature of very long sentences, or a breakdown of the sentence-

[2.] This measure, however, also increases slightly for proportions near .50.

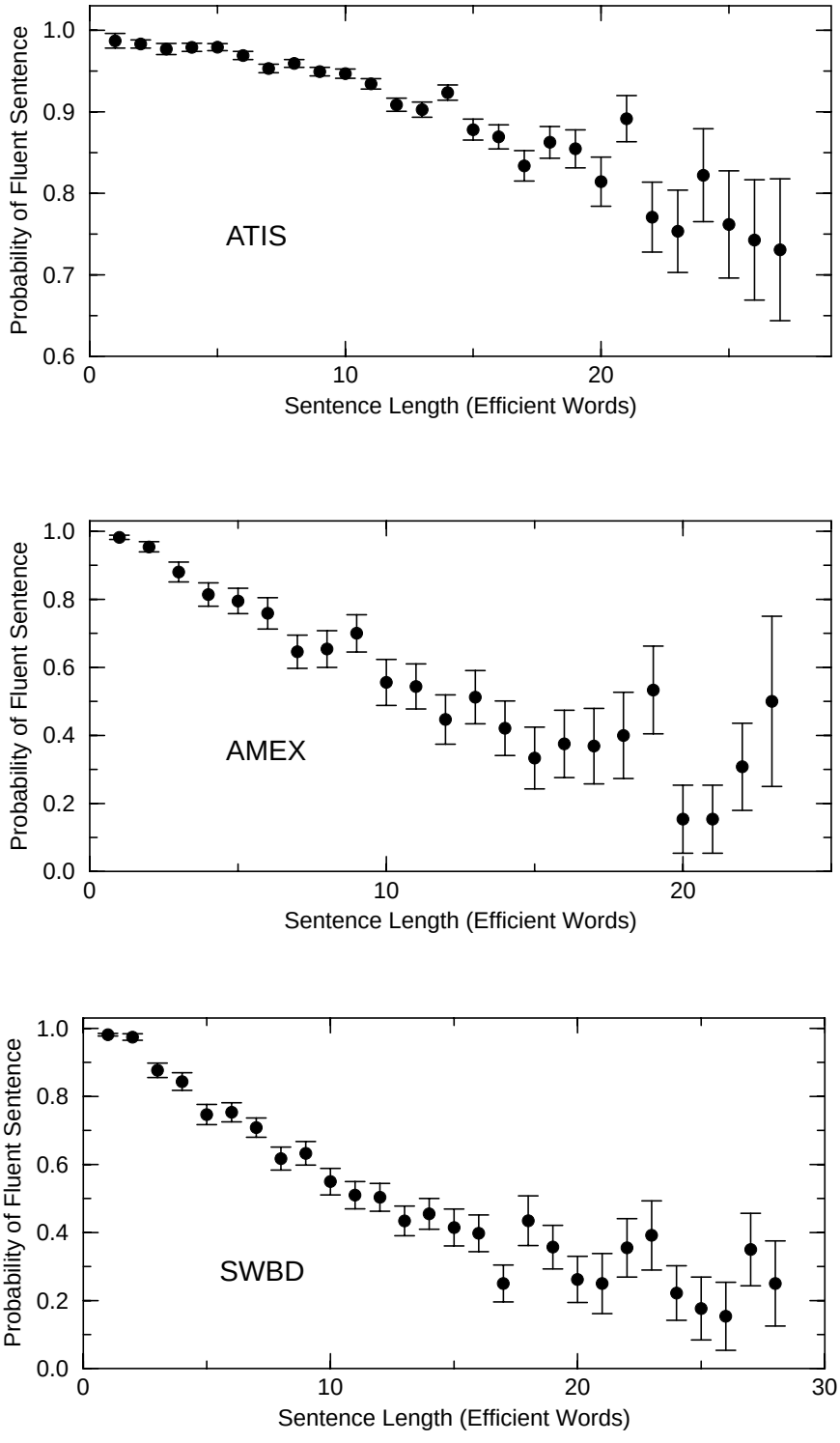


Figure 5. Probability of a Fluent Sentence by Sentence Length

labeling system for the coding of long sentences. It was not possible to investigate this behavior in the current study, but the effect deserves attention in future work.

The focus turns now to modeling the trends. From Figure 5 we see that the function describing the probability of a fluent sentence given sentence length must be some type of monotonically decreasing trend. The function is constrained theoretically to range between 1 and 0. Empirically however, the function need only describe the data in the range observed; thus, violation of a theoretical boundary constraint is admissible if the violation occurs outside the range of observed data.

Three simple functions were considered:

$$\text{exponential: } y = C * b^{j}$$

$$\text{linear: } y = C + b * j$$

$$\text{inverse: } y = 1/(C + b * j)$$

where j represents sentence length (in efficient words), b is a free parameter (equal to the slope in the function), C is free parameter (equal to the constant in the function), and y represents the probability of a fluent sentence.

In all functions, the parameter b reflects the rate at which sentence fluency decays with each additional word in a sentence. Thus, b can be interpreted as a "per-word fluency rate" parameter. Note however that this parameter reflects an aggregate rate over all words in a sentence and over all sentences in a corpus; it does not necessarily reflect the rate at any particular word. The constant C is a scaling factor. If other than 1, it represents a "sentence-level" effect, i.e. any effect which reduces fluency to the same degree for all sentences, regardless of their length. This could be the case, for example, if there is a constant probability of producing a DF at the beginning of a sentence for all sentence lengths.

Each of these simple functions has a reasonable intuitive interpretation. Differences among the functions can be thought of as differences in the manner in which the per-word fluency rate parameter changes with additional words in a sentence.

An exponential model will provide the best fit if the per-word fluency rate is uniform over sentence length. An exponential model is unique among the models considered in meeting the theoretical boundary constraints, since it ranges between 1 and 0.

A linear model will provide the best fit if the fluency rate at each word *decreases* as sentence length increases--that is, if adding a word to a long sentence reduces fluency to a greater degree than does adding a word to a short sentence. A linear model violates the lower theoretical bound of 0 probability, since at some high value of *j*, it must cross the abscissa. However, as can be roughly inferred from Figure 5, in the ATIS corpus the upper limits of observed values of *j* fall within the region in which *y* would remain above 0 for a linear fit. Therefore, a linear model may provide a reasonable fit for the range of observed data in this corpus.

An inverse model will provide the best fit if the fluency rate at each word *increases* as sentence length increases, i.e. if the decay in fluency is mainly due to words occurring early in a sentence. An inverse function asymptotes, as required, at high values of *j*. However, it increases rapidly to values above 1.0 at low values of *j*. Because short sentences do fall within the range of observed data, this behavior is likely to affect the empirical fit of the function.

The three functions were fit for the data from each corpus. The fits excluded values of *j* at which the trends became noisy in Figure 5. The thresholds chosen corresponded to natural breaks in sample size.[3] The maximum sentence length used for the ATIS and SWBD corpora was 19 words; for AMEX, which contained less data overall, the maximum length used was 18 words.

Linear regressions were performed using the method of least-squares. Although the reliability of the estimate of fluent-sentence rate at each value of *j* varies with sample size, regressions were unweighted (the contribution from each *j*-bin was counted equally.) Use of unweighted regressions was warranted because noisy data had been removed, and because given the skewed distribution of fluent-sentence lengths in AMEX and SWBD, a weighted regression would fit the data for short sentences at the expense of the data for long sentences.

[3.] The form and parameter values of the models change little when noisy data points are included, but prediction error increases at high values of *j*, making it difficult to compare models.

The fit for the nonlinear functions was performed after the data were transformed to a linear space.[4] Parameter values and prediction error obtained for the regressions were converted back to a linear space for comparison.

For the exponential and linear fits, the value of *C* was close to 1.0 for all three corpora. This is the value of the y-intercept if there is no vertical scaling of the function, since 0-length sentences must contain 0 DFs. These models could therefore be simplified to single-parameter models (by setting *C* to 1). The finding that *C* is approximately equal to 1 suggests the lack of sentence-level effect, i.e. the lack of a factor that would affect all sentences to the same degree, regardless of their length.

Table 8 shows the prediction error for each of the five fitted functions. Since the values shown have been transformed back to a linear space, model fits for a particular corpus can be directly compared.

Table 8: Prediction Error for Models of Fluent-Sentence Rate

Corpus	2-Parameter Exponential $C * b^j$	2-Parameter Linear $C + b*j$	2-Parameter Inverse $1 / (C + b*j)$	1-Parameter Exponential b^j	1-Parameter Linear $1 + b*j$
ATIS	.0157	.0150	.0165	.0172	.0164
AMEX	.0419	.0507	.0708	.0496	.0514
SWBD	.0426	.0578	.0727	.0493	.0596

4. For the exponential model, the transformation was:

$$y = C * b^j \implies \ln(y) = \ln(C) + \ln(b) * j$$

For the inverse model, the transformation was:

$$y = 1/(C + b * j) \implies 1/y = C + b * j$$

As shown in the table, the 2-parameter inverse model is associated with the highest prediction error of all 2-parameter models for all corpora. For AMEX and SWBD, the inverse model provides a poorer fit than the single-parameter models. The poor fit of the inverse model is attributable to large deviations at low values of *j*. The inverse model was therefore eliminated from further consideration.

Both AMEX and SWBD are best fit by an exponential function. The single-parameter exponential function provides a better fit for both corpora than a linear function with an additional free parameter. This suggests that the per-word fluency rate in these corpora may be uniform over sentence length.

ATIS appears to be (just slightly) better fit by a linear than an exponential function. Although the linear model is unrealistic theoretically for high values of *j*, it provides a good fit for the observed data. This result suggests that the per-word fluency rate in ATIS may decrease at higher values of *j*. The behavior of the per-word fluency rates in all corpora is investigated in more detail in Section 5.3.1.

The parameter values from the fits provide a way to directly compare DF rates across the three corpora. To keep matters simple, the estimates from the single-parameter exponential model are used for all corpora. Although the ATIS corpus is slightly better fit by a linear function, it is reasonable to use the parameter obtained in the exponential fit because the exponential function approximates a straight line for high fluency rates. Figure 6 shows the fit of this model for all three corpora. The fit was performed in transformed space as explained above; the function in Figure 6 is:

$$y = exp(B*j)$$

where

$$B = ln(b)$$

Table 9 shows the slope, 95% confidence limits and correlation coefficients for the linear fits in the transformed space. From Figure 6 and from Table 9, we see that the trends for AMEX and SWBD are extremely close. Statistically, we cannot reject the null hypothesis of identical slopes for the two corpora, since the slopes differ by only .0007, which falls well within

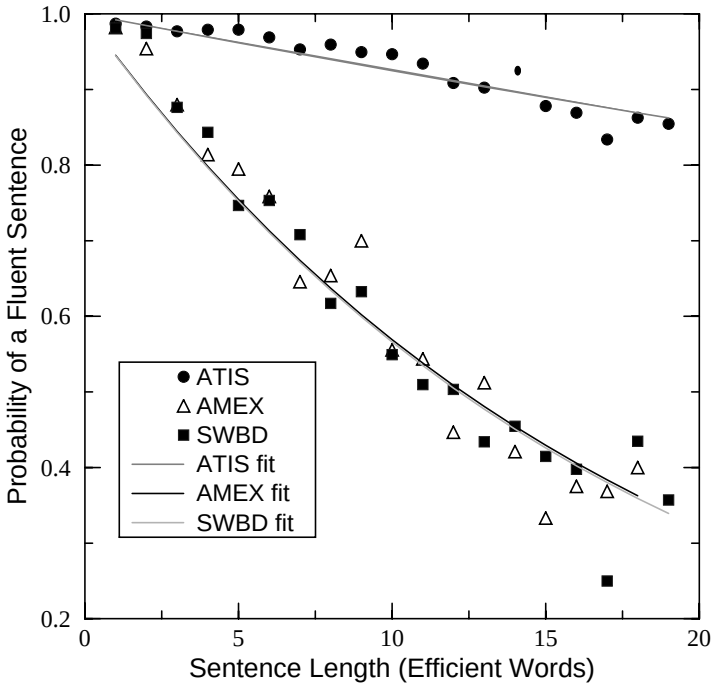


Figure 6. Probability of a Fluent Sentence by Sentence Length: Fit of Exponential Model

Table 9: Linear Fit (in Transformed Space) for 1-Parameter Exponential Model

Corpus	Degrees of Freedom	Slope = $ln(\boldsymbol{b})$	95% Confidence Limits for Slope	Correlation Coefficient
ATIS	18	-.0078	+/- .0015	.90
AMEX	17	-.0563	+/- .0048	.98
SWBD	18	-.0569	+/- .0044	.98

the 95% confidence limits for both corpora. The difference between the ATIS slope and each of the other slopes (greater than .0048) exceeds confidence limits for all corpora. Thus the slope for the ATIS corpus is significantly lower than the slope for either of the other two corpora.

The correlation coefficients in Table 9 represent the proportion of the variability in fluent-sentence rate that can be explained by a linear relationship (in transformed space) with sentence length. As seen in the table, this correlation is particularly high for AMEX and SWBD. When transformed back to a linear space, the slope reflects the overall probability of fluency at each word; one minus this probability is the likelihood of *disfluency* at each word, as shown in Table 10:

Table 10: Per-Word Rate Predictions of Exponential Model Assuming No Cooccurrence Effect

Corpus	Per-Word Fluency Rate ***b***	Per-Word Disfluency Rate 1 - ***b***
ATIS	.9922	.0078
AMEX	.9453	.0547
SWBD	.9447	.0553

Thus in the ATIS corpus, there is less than a 1% probability of disfluency at each word; in AMEX and SWBD this probability is roughly 5.5%.

Two important qualifications must be made concerning these per-word rates. First, the rates are overall or aggregate rates, summed over all sentences and over all speakers within a corpus. Second, it is not necessarily the case that the per-word DF rate will equal 1-*b* as determined from the results in the fit for the fluent-sentence data. The per-word DF rate can be predicted from *b* only if the occurrence of each DF is independent of the occurrence of all other DFs. If DFs tend to cooccur within sentences at rates significantly higher or lower than predicted by the binomial distribution, then the relationship between the sentence-based and word-based rates will be more complex. These two issues are explored in detail in later sections of this chapter.

5.2.2 Rate by sentence length by speaker

It is important to know whether the exponential behavior of the trends just described holds for individual speakers. The exponential behavior over speakers could result, for example, from an association between average sentence length produced, and per-word DF rate within speakers.

To address this question, data from the 30 SWBD speakers were used. As explained in Chapter 4, these are the only speakers for whom there were enough data to address questions of individual differences. Of the 30 speakers, the 14 speakers having the most overall speech were used, since the breakdown by sentence length requires a large amount of data. It is particularly difficult to obtain enough values at high sentence lengths because frequency falls off rapidly as length increases, as was shown earlier (see Figure 4).

To obtain enough observations to estimate the rate parameter at different lengths, data were binned into three length groups: 1-4 words, 5-9 words, and 10-14 words. The estimation of the fluent-sentence rate was computed as the number of fluent sentences in the bin divided by the total sentences in the bin. A data point was included only if there were at least 10 total sentences in the bin. Some speakers did not have this threshold number of sentences in the 10-14 word bin; for these speakers, only the points at 1-4 and 5-9 were plotted.

Results are shown in Figure 7. Each curve connects points for a particular speaker. Although there are insufficient data for statistical analysis, from inspection results are not inconsistent with a hypothesis that the exponential trend holds within speakers. Each curve that has three points shows a steeper slope between the first two bins than between the second and third, and ranges between 1 and 0. These observations suggest an exponential model should provide a better fit than a linear model or inverse model, respectively, for data from individual speakers.

It is apparent from Figure 7 that different speakers have different per-word fluency rates, since slopes differ in value. This result is consistent with literature on differences in overall DF rates across speakers, as discussed in Chapter 2. An important additional observation is that the slopes for speakers whose data stop at the 5-9 word bin are interspersed among the slopes for speakers having observations in the 10-14 word bin. As described above, lack of a third point indicates a lack of long-sentence data for the speaker. Therefore, there does not appear to be an association between a speaker's tendency to produce longer sentences and his or her fluency rate. This result adds weight to the hypothesis that the aggregate exponential behavior reflects a conglomeration of individual exponential trends rather than an association between fluency rate and average sentence length within speakers.

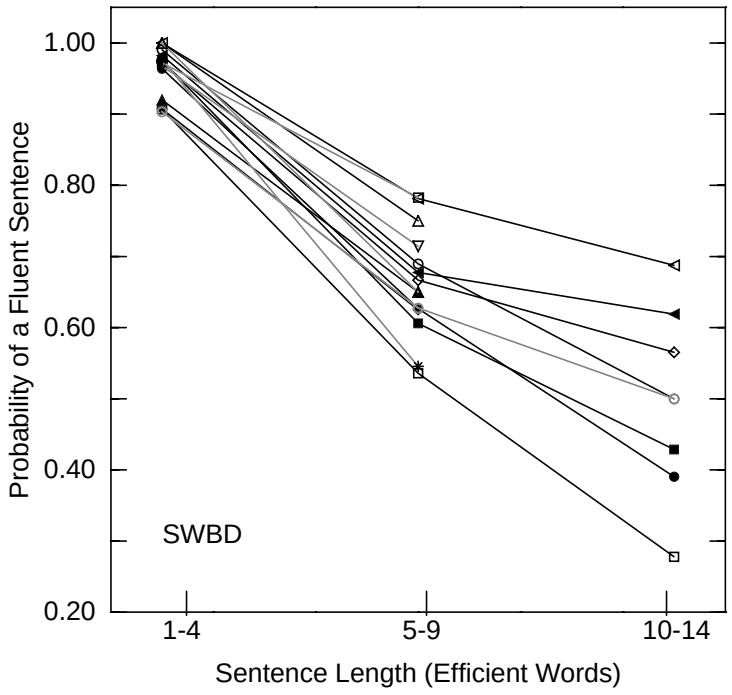


Figure 7. Probability of a Fluent Sentence by Sentence Length, by Speaker.
(Lines connect points for individual speakers.)

5.2.3 Rate in corpus overall

Section 5.2.1 presented a model for determining the likelihood of a disfluent sentence given its length. It is also useful to know the rate of disfluent sentences in a corpus as a whole, when the actual distribution of sentence lengths in the corpus is taken into account. Figure 8 shows, for each corpus, the probability of a *disfluent* sentence as a function of the maximum sentence length included. As in the earlier figures, length is measured in efficient words. The probability measure is the ratio of disfluent sentences to total sentences in each maximum-length bin. Note that while the length measure on the abscissa is cumulative, the probability measure on the ordinate is not. The latter measure is a fraction rather than an absolute count; it is not inherently constrained to increase monotonically.

The probability measure rises rapidly for all three corpora up to about the 15-20 word maximum-length bin, at which point it begins to level off. The rise reflects the increase in disfluent-sentence rate for longer sentences as predicted by the model in Section 5.2.1. The

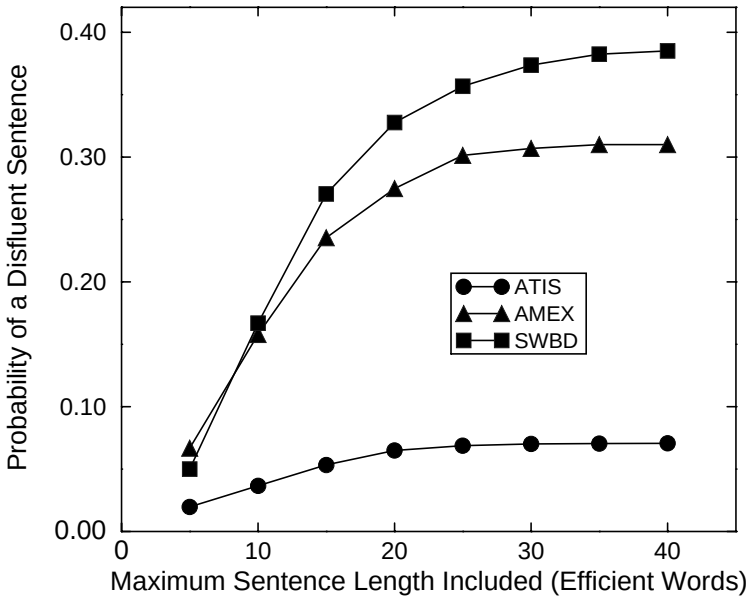


Figure 8. Probability of a Disfluent Sentence by Maximum Sentence Length Included

leveling off is attributable to two factors: 1) decreasing sample size at high lengths; and 2) the cumulative measure on the abscissa (i.e. additional sentences included at each new bin make up a decreasing proportion of total sentences as length increases.)

As expected, the values for ATIS are substantially lower than those for the other two corpora. The new knowledge gained from Figure 8 is that AMEX and SWBD, which were found earlier to have statistically indistinguishable *b* values in the disfluent-sentence model, differ on the current measure. When only sentences below about 10 words are included, the rate of disfluent sentences for these corpora is roughly the same. As longer sentences are added, however, the rate of disfluent sentences in SWBD increases steadily over that for AMEX.

This behavior reflects a difference in the distribution of sentence lengths in the two corpora. Mean and median sentence lengths (for all sentences, fluent and disfluent) for AMEX were 6.16 and 4 words, respectively, compared with 7.41 and 6.00 words for SWBD. These distributions are significantly different. As can be inferred from Figure 4, the distributions are highly skewed; therefore a Kruskal-Wallis one-way analysis of variance by ranks was used to test for differences in the location of the distributions. Results showed a significant difference between the two groups, $\chi^2(6094) = 22.88$, $p<.0001$. The higher percentage of disfluent

sentences overall in the SWBD corpus is therefore attributable to a higher percentage of long sentences in that corpus.

5.2.4 Section summary

The probability of a fluent sentence given its length is predicted quite well by a single-parameter exponential model. If DFs are assumed to occur independently, the parameter in the model corresponds to a per-word fluency rate that is uniform over sentence length. The per-word *disfluency* rate is equal to one minus the per-word fluency rate. The exponential model provides a good fit for the AMEX and SWBD corpora. ATIS, however, is slightly better fit by a linear model, suggesting the per-word disfluency rate in ATIS may increase with sentence length. Parameter values for the exponential model indicate that AMEX and SWBD have statistically indistinguishable per-word disfluency rates, while the rate for ATIS is significantly lower.

Inspection of trends for individual speakers in the SWBD corpus suggest that the exponential behavior of the trend observed over speakers is likely to reflect an aggregation of individual exponential trends, rather than an association between fluency rate and average sentence length within speakers. Speakers appear to differ, however, in absolute fluency rates.

When disfluent-sentence rate is measured taking into account the frequency of sentence lengths in a corpus, the rate in the SWBD corpus exceeds that in the AMEX corpus. This result is not inconsistent with the finding that AMEX and SWBD have similar parameter values in the exponential model: the higher rate in SWBD is attributable to overall longer sentence lengths in that corpus.

5.3 Rate of Disfluency per Word

This section investigates the probability of a DF at any particular word. For purposes of discussion, this per-word disfluency rate will be referred to as *d*:

Definition: *d* = per-word disfluency rate

In the previous section, we saw that a simple model could predict the likelihood that a sentence was fluent given its length. The model's parameter, *b*, predicted the fluency rate at each word. Therefore, *d* is predicted by 1-*b*. However, as discussed at the end of Section 5.2.1, however, two caveats apply to the prediction of *d* from 1-*b*. First, 1-*b* is an aggregate rate, computed over all

words in a particular length bin. Second, the relationship between 1-*b* and *d* holds only if DFs occur independently. If DFs cooccur in the same sentence at a rate higher or lower than that predicted by the binomial distribution, then *d* cannot be directly determined from 1-*b*.

The following sections describe analyses in which *d* is directly measured, and in which features affecting *d* are investigated. Section 5.3.1 describes *d* as a function of sentence length. Sections 5.3.2 and 5.3.3 examine the relationship between *d* and DF position (sentence-initial or sentence-medial), across and within speakers. Sections 5.3.4 and 5.3.5 examine the relationship between *d* and a combination of position and sentence-length features, across and within speakers. Sections 5.3.6 and 5.3.7 investigate the question of whether DFs occur independently, across and within speakers.

5.3.1 Rate by sentence length

Given the results in Section 5.2, we have expectations about *d* if DFs occur independently. We expect that *d* will increase with sentence length for ATIS, because ATIS was best fit by a linear function. We expect that AMEX and SWBD will show a constant *d* over sentence length, since these corpora were best fit by an exponential function. And, we expect that the value of *d* for AMEX and SWBD should be equal to 1-*b*.

In examining *d* as a function of sentence length, it is helpful to first show Figure 9, which indicates the average number of DFs in a sentence as a function of sentence length. This measure is the ratio of total DFs to total sentences within a length bin. Note that the measure is gradient; it *counts each DF* in a sentence (whereas the disfluent-sentence measure discussed earlier made only binary distinction between no DFs and one-or-more DFs in a sentence).

The behavior of the per-*word* rate of disfluency (*d*) as a function of sentence length corresponds to the trend in Figure 9, multiplied by 1/*j*, (where *j* is equal to the number of efficient words in a sentence) as shown in Figure 10. Figure 9 indicates that in absolute terms, the average number of DFs in a sentence grows roughly linearly with sentence length for all three corpora. The trends for AMEX and SWBD appear to grow at similar rates, while the trend for ATIS grows considerably more slowly.

When the trends are multiplied by 1/*j* to obtain the per-word rate, a number of observations may be noted. First, the resolution at low values of *j* in Figure 10 allows

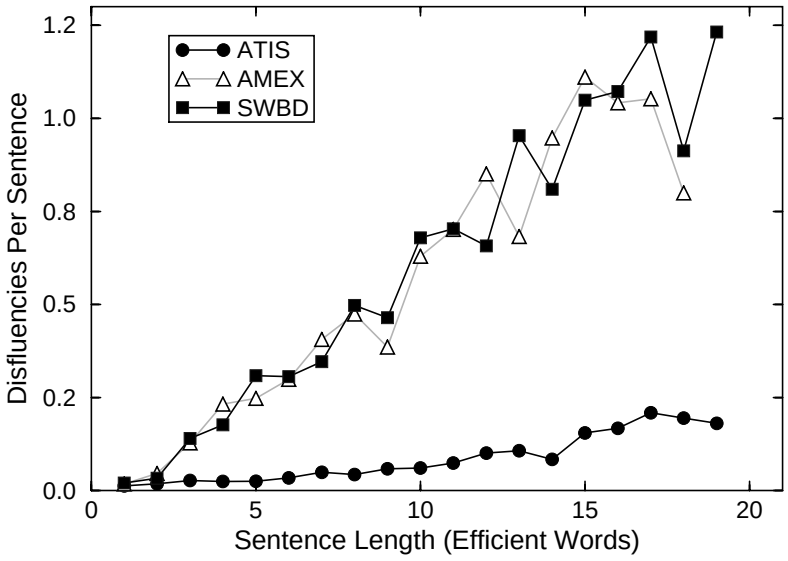


Figure 9. Rate of Disfluency Per Sentence by Sentence Length

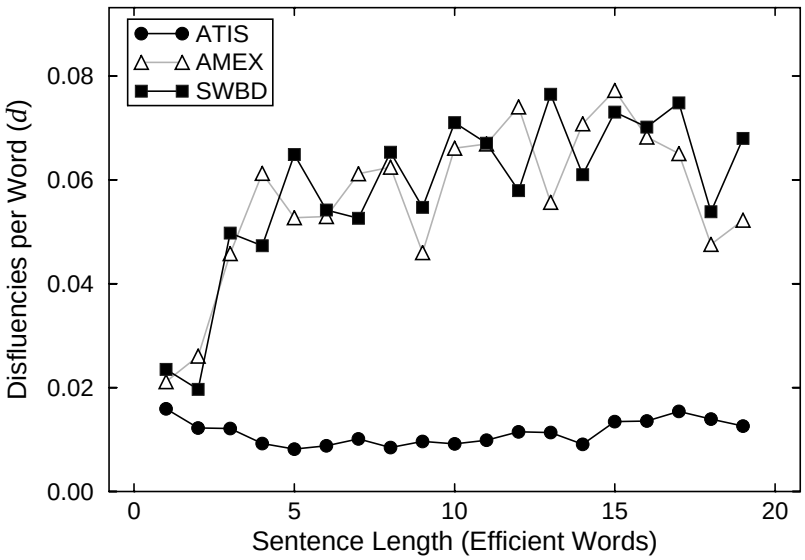


Figure 10. Rate of Disfluency Per Word by Sentence Length

observation of a discontinuity in all three trends for short sentences. Interestingly, the pattern of the discontinuity for ATIS differs from that for the other two corpora.

For ATIS, *d* decreases from 1 to about 5 words, then increases thereafter. From Figure 10 it can be seen that the discontinuity corresponds to a roughly constant per-sentence value (i.e. the average number of DFs in a *sentence* is not decreasing from 1 to 5 words; rather, only the per-word rate decreases). Inspection of the short sentences revealed that the stable per-sentence rate for short sentences did not appear to be attributable to particular speakers. However, 83% of the DFs in sentences less than five words long were sentence-initial filled pauses (for example "uh to Boston"). This is a substantially higher rate both of filled pauses per total DFs, and of initial DFs, than observed generally (see Chapter 6). Further discussion of the discontinuity is provided in Section 5.3.2, where trends are broken down by DF position.

The discontinuity at low sentence lengths for AMEX and SWBD is in the opposite direction. As seen in Figure 10, *d* is notably lower for 1- and 2-word sentences than for the remainder of the lengths. This behavior is explained in part by the preponderance of continuers and assessments such as "uh-huh" at these sentence lengths (see Chapter 4 for a description of the sentence labeling for these cases). The disfluency rate for sentences consisting only of continuers or assessments is extremely low. This is not surprising, since these elements do not involve planning of propositional content.

A second observation from Figure 10 is that if data at short lengths are excluded, trends for AMEX and SWBD, though noisy, appear to be fairly flat over sentence length. The trend for ATIS, however, appears to rise steadily over sentence length. To assess the relationship between *d* and sentence length, a linear regression was performed for each corpus, after removing all sentences containing less than five words. The regressions were fit to the estimate of *d* for each bin.[5] The null hypothesis in each case corresponds to an absence of association between sentence length and *d* (i.e. to a slope that is not significantly different from zero). Results are shown in Table 11.

---

5. This method does not take into account implicit sampling variation in the estimate of rate at each length; however, regressions using the raw data (i.e. using each sentence as an observation) do not change the pattern of statistical results.

Table 11: Test of Hypothesis that Per-Word DF Rate (d) Changes with Sentence Length

Corpus	Degrees of Freedom	Critical Value of ***t*** at $\alpha = .05$, two-tailed	***t*** -Value of Regression Coefficient	Correlation Coefficient r^2
ATIS	13	2.160	5.70	.84
AMEX	12	2.179	1.16	.32
SWBD	13	2.160	1.49	.38

As shown, for ATIS the *t*-value of the regression coefficient exceeds the critical value at $\alpha = .05$. For AMEX and SWBD, however, the value of the test statistic does not exceed the critical value. Therefore we reject the null hypothesis for ATIS, but not for AMEX or SWBD. For ATIS, 84% of the variability in *d* can be explained by sentence length (in the range from five to 19 words.) For AMEX and SWBD, sentence length is of no definite use in predicting *d*, since the slope for *d* over sentence length was not found to be significantly different than zero.

These results are consistent with the results predicted from Section 5.2.1, where fluent-sentence rate in AMEX and SWBD was best fit by an exponential model for fluent-sentence rate, while the rate in ATIS was best fit by a linear model. The exponential model predicted the uniform *d* observed for AMEX and SWBD. The linear model predicted the increasing *d* observed for ATIS. The increasing *d* in ATIS has the effect of flattening out the bow in the disfluent-sentence curve, since although each additional word represents a decreasing percentage of the overall word count, this is compensated for by an increasing likelihood of disfluency at the longer lengths. Interestingly, the result for ATIS is consistent with the work of Oviatt (1994), who found DF rates in human-computer dialog to increase with utterance length (also measured in words; see Chapter 2). It was not possible to determine here whether the differences in the behavior of *d* across corpora were associated with differences between human-human and human-computer dialog, or with other variables. In future work, careful study of differences among sentences as sentence length increases, as well as studies of additional domains, could help shed light on this question.

Rates for AMEX and SWBD appear to be very close. Although rates did not significantly differ for the parameter in the disfluent-sentence model, the comparison of *d* values between the corpora must be tested separately, since as explained above, *d* cannot be directly determined from 1-*b* unless DFs occur independently. For each corpus, *d* was estimated as the

average of the *d*-estimates at each length bin. Results were .0587 for AMEX and .0612 for SWBD. The critical difference for $t(27)$ at α=.05, two-tailed, is 2.052. The *t* statistic obtained was .73, which does not exceed the critical value. Therefore, there is no evidence upon which to reject the null hypothesis that AMEX and SWBD have the same overall rate of disfluency at each word.

For AMEX and SWBD, we can also compare the value of *d* to that predicted by 1-*b* in the exponential model.[6] Recall that if DFs occur independently, *d* should be equal to 1-*b*. The estimates of *d* (.0587 and .0612, for AMEX and SWBD, respectively) are slightly higher than the corresponding estimates of 1-*b* shown in Table 10 (.0547 and .0553, respectively). One possible explanation for the difference is the exclusion of the very short sentences in the estimate of *d*, since these sentences have low DF rates. A second possibility is the presence of a cooccurrence effect. In this case, since *d* exceeds 1-*b*, the predicted cooccurrence effect is a positive correlation between *d* and presence of another DF in the same sentence.[7] The possibility of a cooccurrence effect is investigated in detail in Section 5.3.6.

5.3.2 Rate by position

This section examines *d* as a function of word position. As explained in Chapter 4, word position was labeled using a binary distinction between sentence-initial position and sentence-medial position.[8] While simplistic, the sentence-initial versus sentence-medial distinction (henceforth simply "initial" and "medial") nevertheless captures minimal phrase-related information, since initial position is likely to correspond to the beginning of a phrase, whereas medial position may or may not correspond to the beginning of a phrase.

---

6. The comparison between *d* and 1-*b* is not made for ATIS, because of the complication that *d* changes over sentence length for this corpus.

7. A positive correlation would result in fewer disfluent sentences for a fixed *d*, since DFs would be distributed among fewer total sentences. This would increase *d* relative to the value predicted by 1-*b*.

8. Although the specific location of each DF (the word index within the sentence in which it occurs) is available in the HLD, use of this information in the current work would be difficult to interpret due to a lack of phrasing information.

Figure 11 shows rates of initial and medial DFs in the three corpora. Rate is computed as the number of observed DFs in the position, divided by the number of potential sites for the position. This measure corresponds to *d* of the previous section, but restricted to the set of words in a specific position. The number of potential sites for initial DFs is equal to the total number of sentences. The number of potential sites for medial DFs is equal to the number of total words minus the number of total sentences. The appropriate total-words measure is efficient words (i.e. ignoring words deleted in DFs), since words within another DF do not constitute a potential DF site in these analyses. Error bars indicate the standard deviation of the proportion as predicted by the binomial distribution.

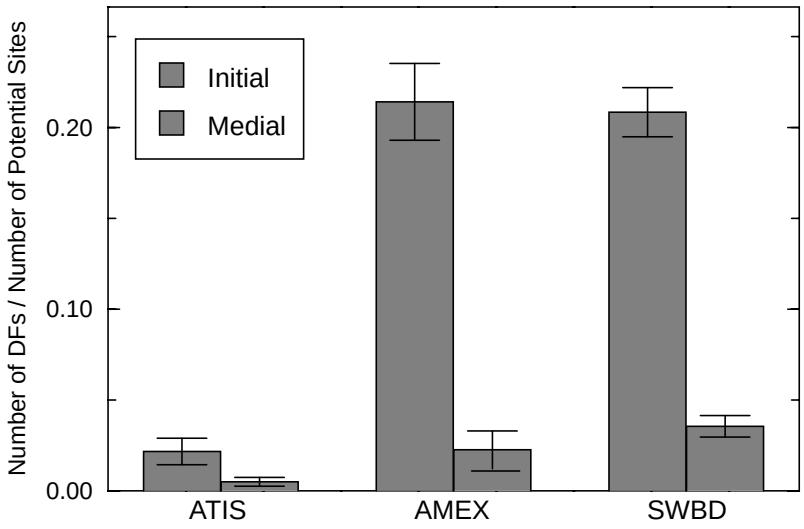


Figure 11. Rates of Initial and Medial DFs

A clear observation from Figure 11 is that rates for initial position exceed those for medial position in all three corpora. A comparison of initial to medial rates within each corpus yields z=24.24, 32.21 and 44.55 for ATIS, AMEX and SWBD, respectively; p<.0001 in all cases. Therefore, in all corpora, DFs are significantly more likely to occur in initial position than in medial position.

This result is consistent with past work showing DFs to be more likely to occur early in a phrase (see Chapter 2). DFs in initial position may reflect cognitive processing, for example, planning of the sentence. However as noted in Chapter 2, the onset of a sentence is confounded with the onset of a turn. Thus it should be kept in mind that DFs in initial position could serve a discourse rather than a processing function. Results also have implications for speech applications. For example, automatic DF detection could be aided by assigning different conditional probabilities of disfluency to initial and medial positions.

A second observation from Figure 11 is that ATIS has a lower overall rate of DFs than AMEX or SWBD. This is a predictable result given the lower per-word DF rate for ATIS seen in the previous section. The new information in Figure 11 is that the lower overall rate corresponds to a reduction of *both initial and medial* DFs. ATIS shows significantly lower rates than the other corpora for both initial DFs ($z = 39.49$ and 47.13) and medial DFs ($z = 19.34$ and 45.91; all p values $< .0001$.)

A third finding is that AMEX and SWBD do not differ in rates of initial DFs ($z=.49$, $p>.05$), but do differ significantly in rates of medial DFs ($z=5.93$, $p<.0001$). The theoretical implication of this result remains an issue for future work. Mathematically, this result must be reconciled with earlier results, in which the two corpora were found to have overall per-word DF rates (*d*) that did not significantly differ. The overall per-word DF rate is equal to the sum of initial and medial DFs, divided by the total words in the corpus. Because the rate of initial DFs is so much higher than that of medial DFs, in absolute terms this rate is more influenced by initial DFs than by medial DFs. Therefore the difference in medial DFs between the corpora has little effect on the statistical comparison of overall rates.

5.3.3 Rate by position by speaker

In the previous section we saw that the rate of initial DFs significantly exceeds that of medial DFs for all three corpora. The question addressed in this section is whether this behavior holds within speakers. It is possible that initial DFs could be more frequent in the aggregate results as a consequence of a higher rate of initial than medial DFs for certain speakers, but equal rates of initial and medial DFs for other speakers.

To investigate this possibility, data from the 30 SWBD speakers were examined. In Figure 12, each point reflects data for a single speaker. The value on the abscissa is the speaker's rate of initial DFs; the value on the ordinate is the same speaker's rate of medial DFs. In each case, rate is computed as the total observations in the relevant position, divided by the total potential sites. The equivalence line (y=x) is also indicated.

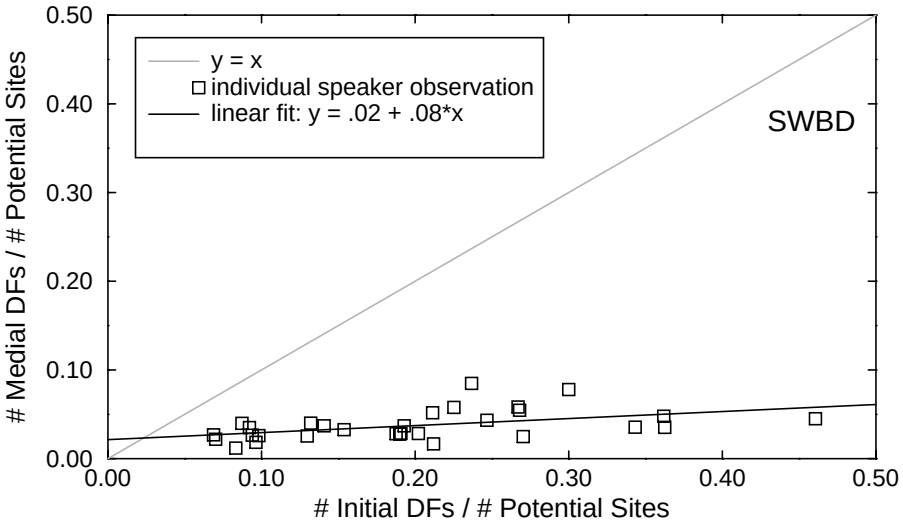


Figure 12. Rate of Initial and Medial DFs by Speaker

As shown, speakers range in absolute rates for both initial and medial DFs. All speakers, however, have points below the equivalence line. Therefore, it is not the case that the high initial rate is due to only a few speakers making large number of initial DFs. Rather, all speakers produced a much higher rate of initial than medial DFs.

A linear fit to the observed data is also shown. The slope is significantly different from zero: $t = 2.86$, $p<.05$. This suggests some association between a speaker's rate of initial DFs and rate of medial DFs. However, it was not possible to evaluate the significance of this correlation, because the internal variability for each speaker in the production of the two types could not be determined due to lack of sufficient speech data per speaker. Future work could address this question by using a much larger set of speech data.

5.3.4 Rate by position by sentence length

The previous two sections have shown that there are two different values for the likelihood of a DF at each word--one for initial words and another for medial words. This section examines how each of these rates varies with sentence length.

Figure 13 shows rates of initial and medial DFs by sentence length. Two rate measures are shown; the rate per sentence and the rate per word. These measures correspond to the per-sentence and per-word measures described earlier. The solid curves in the graphs indicate the rate of total DFs, or the sum of the initial and medial values at each length bin.

A number of observations can be noted from Figure 13. First, in the per-sentence plots, the overall trends rise with sentence length. This result is expected, since longer sentences present additional opportunities for the occurrence of a DF. An unexpected finding, however, is that in all three corpora, the trends for *initial* DFs also rise with sentence length. This finding is unexpected because each sentence has only one possible point for an initial DF. In all three cases, despite the noisy curves, the rise is significant, as shown in Table 12.

Although the rising initial curves in the per-sentence plots were unexpected, in retrospect, they explain an earlier result. Recall that when the 2-parameter exponential and linear functions were fit to the data on DF-sentence rate, *C* came out close to 1.0 for all three corpora. We can now see why this was the case. It is *not* because initial DFs have the same probability of occurrence as medial DFs--there are indeed two different probabilities, as we saw earlier. Rather, the reason for the lack of a constant term in the exponential model is that *both* rates--initial as well as medial--depend on the length of the sentence. Therefore, both rates are rolled into the base in the exponential model.[9]

This result, in which sentence-initial DFs are more likely to occur before longer sentences, presents an important area for future work in psycholinguistics. A better understanding of the result awaits an analysis of the sentences in terms of aspects more illuminating than simply the number of words, for example in terms of semantic and syntactic makeup.

[9.] Inspection of Figure 13 suggests, however, that the *form* of the trends for initial and medial DFs differ. Initial DF rates appear to grow quickly and level off, whereas medial rates show the opposite behavior. However, there were not enough data to separately model these trends.

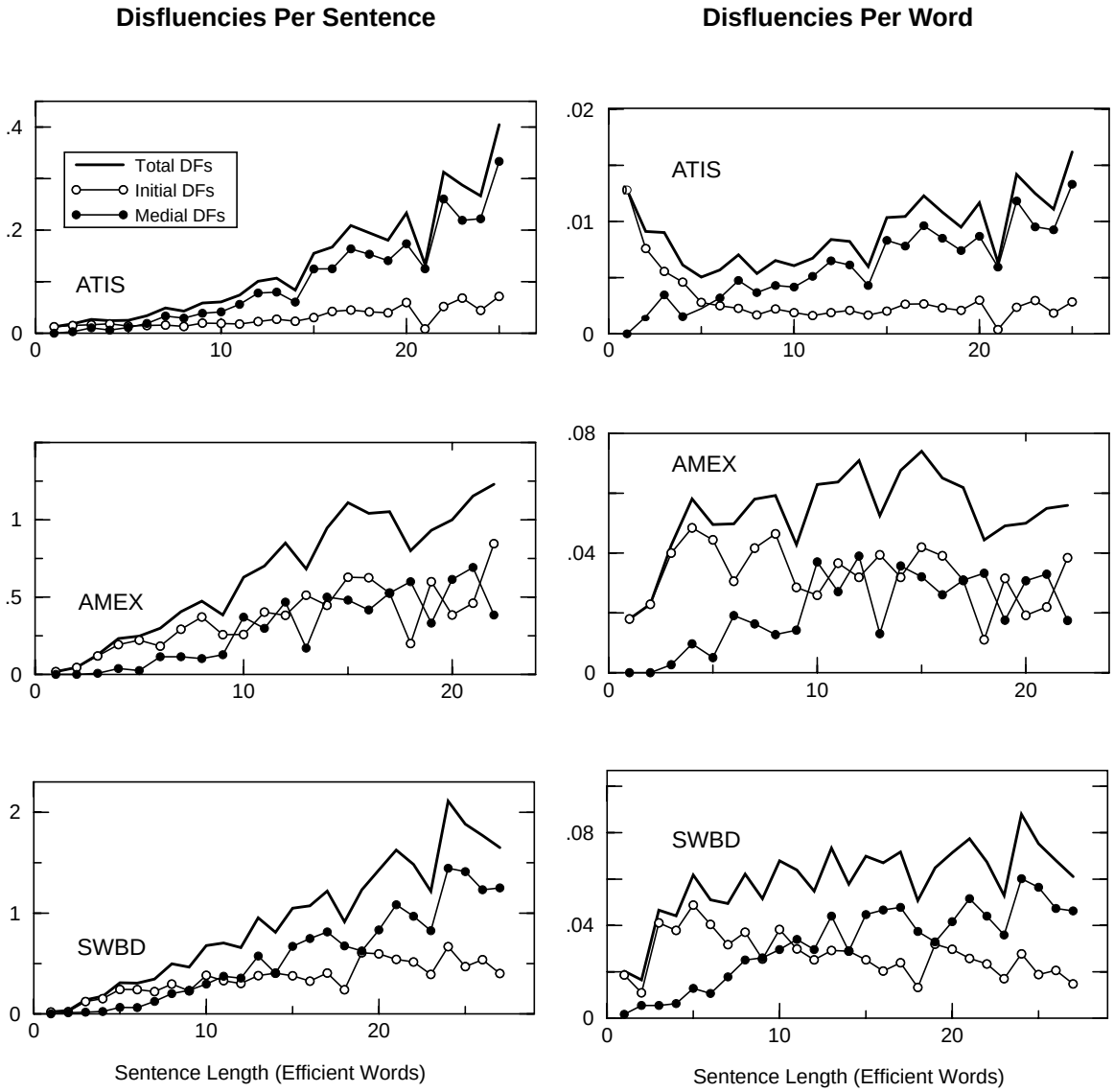


Figure 13. Rate of Initial and Medial DFs by Sentence Length

Table 12: Test of Hypothesis that Per-Sentence Rate of Initial DF Changes with Sentence Length

Corpus	Degrees of Freedom	Critical Value of ***t*** at α = .05, two-tailed	***t*** -Value of Regression Coefficient	Correlation Coefficient r^2
ATIS	22	2.074	6.37	.80
AMEX	20	2.086	6.12	.81
SWBD	25	2.060	7.86	.84

The result could be of immediate use, however, in speech applications. As noted in Chapter 2, automatic DF processing methods have not capitalized on nonlocal features of DFs. The present findings suggest that detection of sentence-initial DFs could be improved by adjusting the initial-DF probability based on an estimate of the length of the sentence it precedes.

Two additional observations can be noted from inspection of the per-word plots in Figure 13. First, the discontinuities for short sentences described earlier appear to be attributable to discontinuities in initial, not medial trends. As shown, the majority of DFs found in short sentences are initial; medial DFs do not tend to appear until the sentence is about five words long. Interpretation of this result requires a better understanding of differences between the short and longer sentences, as well as information on the types of DFs associated with each position.

Second, the per-word plots indicate that the uniform per-word DF rate (*d*) for AMEX and SWBD reflects a summation of trends moving in opposite directions. In both corpora, the per-word rate of initial DFs decreases with sentence length, while the per-word rate of medial DFs increases with sentence length. In the range of observed sentence lengths, these changes in these two rates happen to almost exactly cancel. It is not clear why this should be the case, since the rates are logically independent. A question for future work is whether this behavior is coincidental, or whether it reflects a constraint on the relationship between initial and medial DF rates.

5.3.5 Rate by position by sentence length by speaker

This section tests the hypothesis that the increase in the probability of an initial DF for higher sentence lengths holds within speakers. It is possible that the trends observed in the previous section reflect a correlation between speakers who produce longer sentences, and those who produce initial DFs. Such an association could result in the aggregate behavior observed, without implying a relationship between initial DF rate and sentence length for any particular speaker.

To address this question, data from the 30 SWBD speakers were used. One- and two-word sentences were omitted due to the discontinuity in the initial-DFs curve in the bottom right graph in Figure 13. As discussed in Section 5.3.4, this discontinuity may indicate that short sentences (consisting mainly of only continuers or assessments) provide little opportunity for

disfluency. Inclusion of these sentences would in any case have skewed results in the favor of the hypothesis. Sentences of 20 or more words were also omitted, because of too little data per speaker at high lengths.

To obtain enough data per speaker (since this analysis involved measuring only initial DF rates), sentence length was grouped into two bins: 3-8 words and 9-19 words. These bins roughly divided the data in half by frequency, since longer sentences were overall less frequent than shorter sentences. For each speaker, the rate of initial DFs in each bin was computed as the total initial DFs divided by the total sentences. The rate in the shorter bin was compared to that in the longer bin using a Wilcox matched pairs signed ranks test. Results showed a significantly higher rate of initial DFs in the longer sentence set, $z=3.73$, $p<.001$, two-tailed. This suggests that the association between sentence length and rate of initial DFs holds within speakers.

5.3.6 Rate of cooccurrence

This section investigates whether DFs occur independently of other DFs. As noted earlier, if DFs occur independently, then 1-*b* in the disfluent-sentence model should equal *d*, or the overall per-word DF rate. It was found earlier that for AMEX and SWBD, which both showed a uniform *d* over sentence length, *d* was only slightly higher than predicted by 1-*b* in the disfluent-sentence model. Thus, the slightly higher observed *d may* indicate that DFs tend to cooccur within sentences at rates higher than predicted under an independence assumption.

To investigate this possibility, a subset of cooccurrence phenomena were examined. To simplify matters, the analysis was limited to cooccurrences of initial and medial DFs within the same sentence. No distinction was made within numbers of medial within a sentence (since this rate is confounded with sentence length). Although this method only partially measures cooccurrence, it accounts for the majority of observed cases and avoids the complicated statistics required for a more comprehensive analysis.

Table 13 shows the breakdown of total sentences in each corpus, by the presence/ absence of initial and medial DFs. Cell values are raw counts, expected values, and residuals, respectively. As shown, in all corpora there is a higher observed than expected value in the “presence/presence” cell (and correspondingly in the “absence/absence” cell), suggesting that a

Table 13: Breakdown of Sentences by Presence/Absence of Initial/Medial DFs

		Initial Present	Initial Absent
ATIS:	Medial Present	84 (19.4) 64.6	837 (901.6) -64.6
	Medial Absent	306 (370.6) -64.6	17256 (17191) 64.6
AMEX:	Medial Present	66 (31.3) 34.7	109 (143.7) -34.7
	Medial Absent	248 (282.7) -34.7	1332 (1297.3) 34.7
SWBD:	Medial Present	237 (117.8) 119.2	437 (556.2) -119.2
	Medial Absent	515 (634.2) -119.2	3112 (2992.8) 119.2

DF is more likely to occur in a particular sentence if there is another DF in that sentence (although no cause and effect relations can be inferred from this result).

Table 14 shows results for tests of significance and for degree of association in the three corpora. In all corpora the rate of cooccurrence of an initial DF and one or more medial DFs in the same sentence is significantly different from chance, as seen by the Chi square test. However, in all cases the *degree* of association is fairly low, as seen by the low values for the phi coefficient. (The phi coefficient tests for association in contingency tables having two dichotomous variables. It is normalized for sample size and ranges between zero and one, with one indicating complete association.) As shown in Table 14, the degree of association was below .20 for all corpora.

The finding that there is a small but reliable cooccurrence effect is consistent with earlier results in which we saw that the value of *d* (the per-word DF rate) was slightly higher than that

Table 14: Tests for Cooccurrence Effect

Corpus	χ^2 Coefficient (Test of Significance)	ϕ Coefficient (Degree of Association)
ATIS	230.63, p<.00001	.112
AMEX	52.00, p<.00001	.172
SWBD	150.81 p < .00001	.186

predicted by 1-*b* in the exponential model of disfluent-sentence rate. The difference is small and in the direction predicted by a positive association between presence of initial and medial DFs. Since phi values can be compared directly across corpora, it is worth noting that AMEX and SWBD are similar in value while for ATIS the value is somewhat lower.

We cannot infer cause and effect from these results. One possibility is that production of an initial DF causes a DF later in the sentence because the speaker is somehow distracted from the earlier DF. A second (and not mutually exclusive) possibility is that whatever is responsible for the later DF is also reflected in the planning of the sentence. This second possibility is conceivable given the earlier results showing the rate of an initial DF to increase with sentence length, a nonlocal feature. A third possibility is that the cooccurrence of DFs in the same sentence is an epiphenomenon of a correlation between sentence and speaker, since DFs within a sentence are by definition produced by the same speaker. The first two possibilities remain interesting areas for future work. The third possibility, that the effect is better explained as a speaker-specific effect, is investigated in the next section.

5.3.7 Rate of cooccurrence by speaker

As just mentioned, the cooccurrence effect witnessed for the aggregate data could reflect a speaker-specific effect. DFs could congregate in certain sentences if those sentences were produced by a speaker with a particularly high DF rate. To address this possibility, data from the 30 SWBD speakers were used. For each speaker, a 2x2 contingency table like that in the previous section (Table 13) was constructed. The expected probability of cooccurrence under an independence assumption was computed as the product of the marginals in the table (i.e. the probability of an initial DF times the probability of a medial DF). The observed probability for

the speaker was computed by dividing the total cases of cooccurrence by the total number of sentences.

In Figure 14, each open square indicates the expected and observed probability of cooccurrence for a particular speaker. The filled square corresponds to the point for aggregate data, as determined from the contingency table for the SWBD corpus in Table 13 in the previous section.[10] The equivalence line y=x is plotted for reference. Points on this line reflect the predicted relationship between expected and observed probabilities of cooccurrence under an independence model.

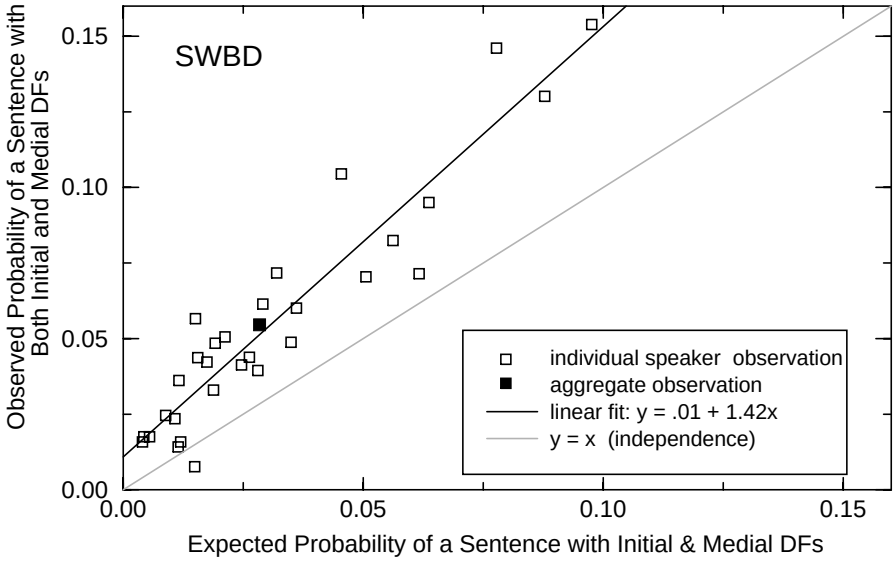


Figure 14. Observed versus Expected Probability of Cooccurrence of Initial and Medial DFs, by Speaker

Under an independence assumption, we expect points for individual speakers to be evenly scattered above and below the independence line. As shown however, 29 of the 30 speakers have higher observed than expected values, a reliable result by Sign test, $p<.001$. There

10. Note that the aggregate point is not directly determined from the speaker-specific values, since the aggregate point does not reflect data blocked by speaker.

is thus a significant tendency for initial and medial DFs to cooccur in a sentence, within speakers.

In addition, the higher the expected value, the further in general the observed is from the equivalence line. Therefore this trend can be modeled. Since from inspection, there is no evidence for anything other than a linear relationship, data were fit with the function

$$P(I+M) = C + s * P(I) * P(M)$$

where:

P(I+M) = probability that a sentence will contain both an initial and a medial DF

P(I) = probability of an initial DF overall

P(M) = probability of a medial DF overall

s = free parameter (slope)

C = free parameter (constant)

The result of the fit was:

$$P(I+M) = .01 + 1.42 * P(I) * P(M)$$

The 95% confidence region for the slope, with 29 degrees of freedom, was 1.22-1.62. This region does not include 1; therefore the fitted slope is significantly different from y=x or the independence assumption. The fit is fairly tight; r^2=.94. The close linear fit explains why the aggregate point lies close to the fit for the individual speaker data. These results suggest, quite interestingly, that the cooccurrence effect observed for the aggregate data reflects individual cooccurrence effects that are all of roughly the same *degree*.

The probability a DF in a particular sentence is therefore conditional upon the presence of another DF in the sentence. For example, under independence, the conditional probability of a medial DF given an initial DF is predicted by the probability of a medial DF overall:

$$P(M \mid I) = P(M)$$

However, the empirical results assign a different probability, which takes into account the probability of an initial DF:[11]

$$P(M \mid I) = .01/P(I) + 1.42 * P(M)$$

These results should be of interest for cognitive models. We saw earlier that individual speakers produce different absolute rates of initial and medial DFs (Figure 12). We could not determine whether these rates were correlated within speaker due to lack of sufficient data. We can see in the present section, however, that the *degree* of association between occurrences of initial and medial DFs is fairly consistent across speakers. The interpretation of these findings is left as a question for future work. Interpretation will require first gaining an understanding of the cause and effect relationship underlying the cooccurrence effect, as discussed earlier. These results also have implications for automatic processing. For example, if there is strong evidence for the presence of an initial DF in a sentence, the probability of a medial DF changes from that in the corpus overall, to the value shown above.

5.3.8 Section summary

The rate of disfluency per word shows a regular trend for all corpora if data for very short sentences are omitted. The rate is uniform over sentence length for AMEX and SWBD, but increases within increasing sentence length for ATIS. These results are consistent with predictions from the disfluent-sentence model presented earlier. The magnitude of the rate parameter for AMEX and SWBD is slightly higher than that predicted by the disfluent-sentence results, suggesting a small cooccurrence effect.

Disfluency rates vary significantly with word position. In all corpora, DFs are much more likely to occur in initial than in medial position. The higher rate of initial DFs is also found within speakers in the SWBD corpus. The rate of initial DFs does not differ significantly between AMEX and SWBD, but medial DFs are more likely in SWBD. Despite a higher relative rate of medial DFs in SWBD, SWBD does not differ significantly from AMEX in overall per-word DF rate because that rate is influenced to a greater degree by initial DFs. The lower overall per-word DF rate in ATIS reflects a lower rate of both initial and medial DFs.

[11.] Derived by substituting the definition $P(B \mid A) = P(A+B)/P(A)$ into the equation for $P(I+M)$.

The per-sentence DF rate of both initial and medial DFs increases with sentence length for all corpora. The increase in per-sentence rate of initial DFs explains the lack of a constant term in the disfluent-sentence model of Section 5.2.1: both initial and medial per-sentence rates are rolled into the base in the exponential model. The increase in initial DF rate with longer sentences is reliable in all three corpora; it is also found to hold within speakers in the SWBD corpus. The magnitudes of the per-sentence rates of initial and medial DFs are such that when multiplied by $1/j$ to obtain the per word rate, both initial and medial rates increase with j in ATIS, yielding an overall per-word rate that increases with j. For AMEX and SWBD, however, the per-word rate of medial DFs increases with sentence length, while the per-word rate of initial DFs *decreases* with sentence length. The increase for medial DFs is roughly canceled by the decrease for initial DFs, yielding an overall per-word rate that is uniform over j for these corpora.

In all corpora, there is a small but reliable cooccurrence effect: an initial DF is more likely to occur in the presence of a medial DF in the same sentence, and vice versa. The degree of association is similar for AMEX and SWBD, but lower for ATIS. The cooccurrence effect is also seen for individual speakers in the SWBD corpus. Across speakers, the observed rate of cooccurrence is well described as a linear function of the cooccurrence rate expected under an independence assumption.

5.4 Rate of DFs with *k* Deleted Words

Whereas the analyses thus far have examined rates of the occurrence of disfluency with respect to the total speech produced, the analyses in the remainder of this chapter examine rates of pattern features with respect to the total DFs produced. The present section examines the pattern feature "deletion length," or the number of words hypothesized to be deleted in the DF (see Chapter 4). This length will be referred to as k:

Definition: k = number of words deleted in a DF

Section 5.4.1 investigates whether longer sentences are associated with longer deletion lengths. Section 5.4.2 describes the raw data, collapsed over sentence length, both including and excluding DFs consisting of only a filled pause. A simple model is proposed to predict the proportion of total DFs that have k deleted words, and the value of the parameter in the model is compared across the three corpora.

5.4.1 Rate by sentence length

This section asks whether sentence length should be taken into account when describing deletion length. Although the efficient words measure for sentence length does not include words deleted in a DF, it is nevertheless possible that higher values of *k* are associated with longer sentences. This is because of the correspondence between words in the RM and RR (see Chapter 4). In many types of DFs, the material deleted is repeated, replaced, or expanded upon in the RR. Therefore, the length of the intended sentence tends to be at least as long as the removed material. If high values of *k* are frequent, these would tend to occur in sentences of at least *k* words.

To check for an association between *k* and sentence length (in efficient words), for each corpus a linear fit was performed to relate the mean value of *k* to sentence length. Short sentences were omitted because of the unusual behavior of these sentences as discussed earlier. In addition, data were analyzed two ways. One method used all DFs in a corpus. The second method omitted DFs consisting of only one or more filled pauses (type "FP" according to the TCA; see Chapter 4). The second method was included for two reasons. First, because FPs are frequent in all corpora (as described in Chapter 6) and because the majority of FPs contain only one deleted word, the high number of single-word deletions contributed by FPs could mask an effect of sentence length. Second, it was not clear at the outset whether extra-syntactic words should be combined with syntactic words in characterizing deletion length (see Chapter 4 for definitions of "extra-syntactic" and "syntactic").

Table 15 shows results of the linear fits. The null hypothesis corresponds to the lack of a relationship between sentence length and *k*. As shown, the *t* statistic for all slopes does not exceed the critical value of *t*; therefore we cannot reject the null hypothesis for any of the six slopes. These results suggest that, across corpora, DFs occurring in longer sentences have deletion lengths similar to those occurring in shorter sentences.

5.4.2 Rate over sentence length

Since the previous section showed that deletion length is not related to sentence length, analyses in the present section examine data collapsed over sentence length. In Figure 15, the raw frequency of DFs with *k* deleted words is plotted for each corpus (note that therefore the

Table 15: Test of Hypothesis that Mean Deletion Length Changes with Sentence Length

Corpus	DFs Included	Degrees of Freedom	Critical Value of ***t*** at α = .05, two-tailed	***t*** -Value of Regression Coefficient
ATIS	all	13	2.160	.5060
ATIS	no FPs	13	2.160	.8272
AMEX	all	12	2.179	.1644
AMEX	no FPs	12	2.179	.4755
SWBD	all	13	2.160	1.231
SWBD	no FPs	13	2.160	1.363

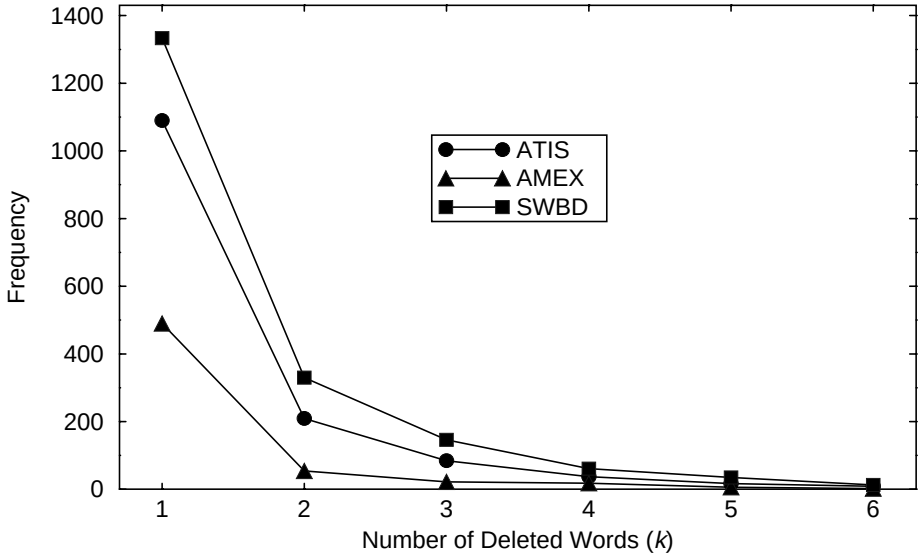


Figure 15. Frequency of DFs by Deletion Length

height of the curves reflects overall sample size, i.e. the total DFs in each corpus). Although *k* is a discrete variable, points for each corpus are connected in the figure for ease of inspection

For all corpora, the trends show a steep and asymptotic decay in frequency as *k* increases; however, the points at *k* = 1 are discontinuous (higher than expected) with the rest of the trend. With the exception of the point at *k* = 1, the trends are well fit by an exponential

model, since when plotted in a *k*, *ln(y)* space, the trends fall lie roughly on a straight line, as shown in Figure 16.

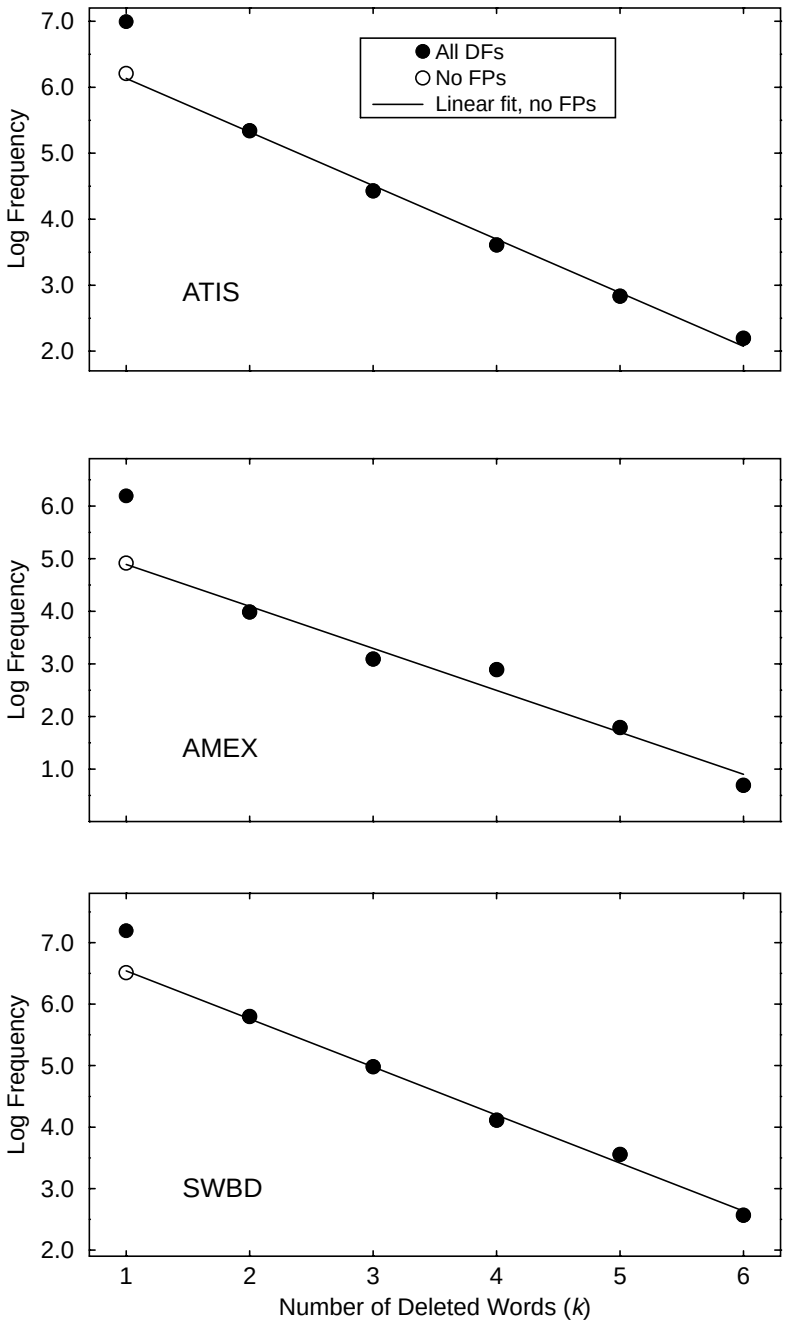


Figure 16. Rate by Deletion Length, with and without FPs: Fit of Exponential Model

Furthermore, the overshoot at $k = 1$ can be accounted for if FPs are removed from the frequency counts. In Figure 16, filled circles correspond to frequencies when all DFs are included; open circles correspond to frequencies when FPs are excluded. Open circles are visible only at $k = 1$; at other values of k, open and filled circles almost exactly overlap. This reflects the fact that DFs of type FP (see Chapter 4) almost exclusively contain a single FP (e.g. “uh”); longer deletion lengths (e.g. “uh uh uh”) are relatively infrequent. As shown, the open circle lies roughly on the line formed by points at $k > 1$.

The removal of FPs is not arbitrary. As discussed in the previous section, it is not clear whether filled pauses should be combined with syntactic words in a model of deletion length. These results suggest that filled pauses should *not* be grouped with syntactic words in characterizing deletion length.

Each of the trends in Figure 16 can be modeled by the function:

$$y = C * q^{k}$$

where y is the percentage of total DFs within a corpus that have k deleted words, and q is a free parameter corresponding to the rate at which frequency decays with increasing k. This is a single-parameter model, because the constant, C, is a deterministic function of q. Since the sum of the percentages for all deletion lengths within a corpus must equal 1, C can be rewritten as[12]:

$$C = (1-q)/q$$

This model is consistent with results of previous studies on deletion length. As noted in Chapter 2, Bear et al. (1992) reported that most DFs have deletion lengths of only one or two words. Given the present results, we can see that this is the case because frequency decays rapidly with k.

It is not clear how to interpret the exponential decay rate. In mathematical terms, q is simply the ratio of the frequency of DFs having k +1 words to the frequency of DFs having k words, for all k. For cognitive models, the interpretation of q will rely on an understanding of

---

[12.] Derivation: consider the sum of all probabilities over k: $\Sigma\,(C * q^{k}) = C * (q + q^{2} + q^{3} + ...)$
It is known that the sum of the geometric series: $1 + x + x^{2} + x^{3} + ...$ is $1 / (1-x)$.
Since the probabilities must add to 1, $C * q * (1 / (1-q)) = 1$.

factors that influence the presence of additional words in the RM. The role of words in the RM depends crucially on DF type (see Chapter 4). For example, for the DF type DEL, deletion length may reflect latency in detection of a problem; we might expect to see a uniform exponential decay in this case. However, for the DF type REP, a longer deletion length corresponds to a longer retracing (see Chapter 2); this is likely to have a different interpretation than a delay in detection of trouble.

Having determined a suitable model, we turn now to the value of the parameter *q*. Table 16 shows the slopes for the fits in Figure 16 (i.e. linear regressions in the *k*, *ln(y)* space), with confidence limits.

Table 16: Linear Fit (in Transformed Space) for Slope in Deletion Length Model

Corpus	Degrees of Freedom	Slope = *ln(q)*	95% Confidence Limits	Correlation Coefficient r^2
ATIS	4	-.8115	+/- .0644	.9983
AMEX	4	-.7979	+/- .1690	.9885
SWBD	4	-.7811	+/- .0613	.9984

As shown, the values for the slopes in the three corpora are quite close, and each slope falls within the 95% confidence limits for the other two. Thus, a good fit can be obtained for all corpora using a single value for *q*. Using the average of the three slopes in the transformed space yields *q* = roughly .45. The corresponding value for *C* is (1-.45)/.45), or 1.22. This single function predicts results for all three corpora, as shown in Figure 17. In addition to the three corpora of the present study, data from the Bear et al. (1992) study of DFs in the ATIS database are plotted. The Bear et al. study involved about half as much speech data as the present ATIS corpus, and largely but not completely overlaps with it.

The finding that *q* does not differ over corpora suggests *q* may be domain-independent. If found to remain invariant over additional corpora, *q* is an important parameter to include in a unified theory. The interpretation of a "universal-*q*" is not immediately clear; this awaits a better understanding of the variables underlying the exponential decay rate.

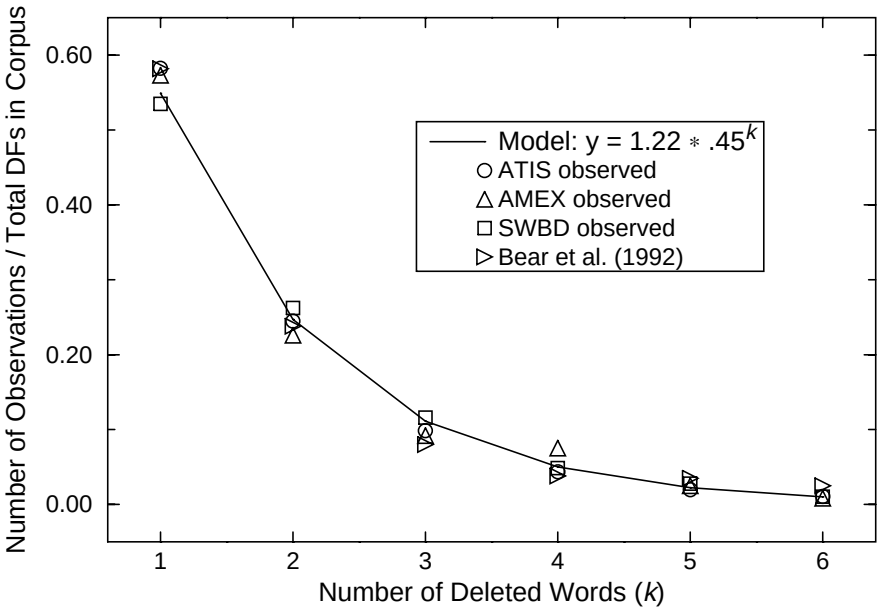


Figure 17. Fit of Deletion-Length Model to All Corpora Using Same Parameter Value

One possible interpretation of universal-*q*, however, is that it reflects a constraint on how many words can be deleted in a DF. Since *q* corresponds to the rate at which frequency decays with increasing *k*, it determines the range of *k* values seen in the majority of DFs. For all corpora, less than 1% of the total DFs were found to have more than six deleted words. Thus, a possible explanation for the invariance in *q* is that producing or comprehending DFs with deletion lengths longer than six words taxes the limits of cognitive processing. Although this is the type of constraint that we would expect to show up as a universal across corpora, more work is needed in order to fully understand this parameter.

These results should also be useful for automatic detection and correction of DFs. For example, they could aid in the detection of sentence-initial deletions (the DF type DEL as explained in Chapter 4, called “false starts” in many classification systems). These DFs are difficult to detect because they often lack cues to disfluency, such as repeated words (Bear et al., 1992) or similar syntactic units in succession (e.g. Hindle, 1983). The model of deletion length presented here predicts that the probability of a DF decreases by .45 for each word in the RM. Therefore, a sentence-initial deletion becomes exponentially less likely as the hypothesized interruption point occurs each additional word away from the beginning of the sentence.

These results could also aid DF *correction*. As noted in Bear et al. (1992), detection algorithms can yield likely points of interruption, but there is often ambiguity in how far back the RM extends. This is particularly true if adequate semantic parsing is not available, since deletions of varying lengths may yield grammatically acceptable but semantically nonsensical parses. An approach adopted in Bear et al. was to make the shortest deletion possible, starting at the word immediately preceding the hypothesized interruption point and extending back one word at a time. Such an approach is consistent with the present findings; however, the present findings additionally provide a specific probability to attach to the hypothesis of each additional word. This information could be combined with other sources of information in weighting likelihoods of competing deletion lengths.

5.4.3 Section summary

The number of words deleted in a DF is unrelated to the length of the sentence in which the DF occurs. This is true across corpora, and even if filled pauses are removed from the analyses.

Across corpora, when the set of all DFs in a corpus is examined as a function of deletion length, the distribution falls roughly on a straight line in the *k*, log-frequency space. The exception is the point for $k = 1$, which in all corpora lies considerably above that predicted by points for $k > 1$. The overshoot at $k = 1$ disappears, however, when filled pauses are removed from the set of DFs considered. For the set of DFs excluding FPs, the percentage of DFs having *k* deleted words is predicted by a single-parameter exponential model. Furthermore, the value of the model parameter appears to be invariant across corpora.

5.5 Rate of DFs with a Fragment

This section examines the rate of word fragments in DFs (see Chapter 3 for definitions and descriptions of fragments). Section 5.5.1 describes the rate of fragments for the set of DFs overall in each corpus; Section 5.5.2 describes the rate of fragments by DF position.

5.5.1 Rate overall

This section examines the rate of fragments in each corpus overall, where fragment rate is computed as the number of DFs containing a fragment divided by the total DFs in the corpus. The relevant set of total DFs, however, should be those for which a fragment is possible. As described in Chapter 4, fragments occur across DFs types. However, there is one DF type for which fragments are never labeled: the type FP (DFs consisting of only one or more filled pauses). Therefore, filled pauses are excluded from the total DF count in the following analyses of fragment rate.

Figure 18 shows the probability of a fragment (per DF, but excluding FPs) in each corpus. Error bars indicate the standard deviation of the proportion.

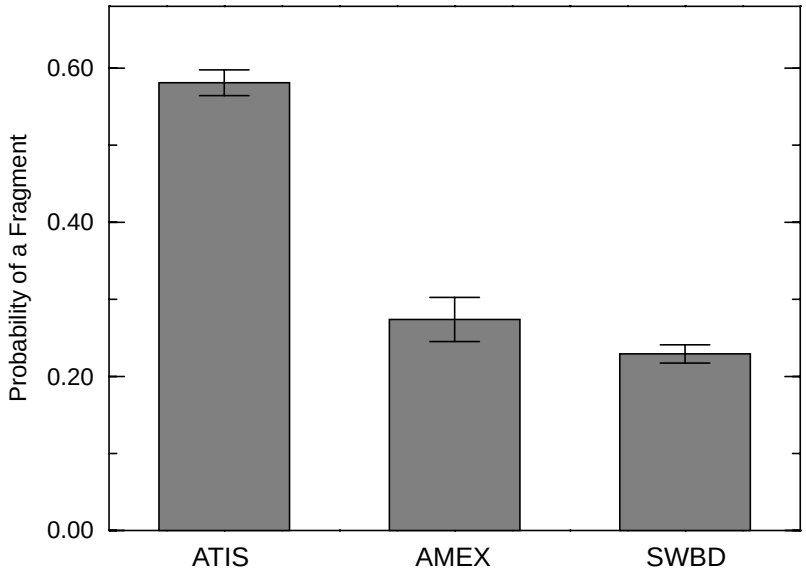


Figure 18. Probability of a Fragment

As shown, ATIS has a higher overall probability of a fragment (.58) than the other two corpora (.27 and .23 for AMEX and SWBD, respectively). A test of proportions shows that ATIS has a significantly higher rate of fragments than AMEX, $z=8.44$, $p<.0001$; from Figure 18 it can be inferred that ATIS also significantly differs from SWBD. The AMEX and SWBD rates are not significantly different, $z=1.53$, $p>.05$.

These results are consistent with past work (see Chapter 2) showing a large discrepancy between the high fragment rate in ATIS (e.g. Bear et al., 1992) and the lower rate in studies not involving human-computer interaction (e.g. Levelt, 1983; Lickley, 1994.) It is not clear what accounts for this difference between domains; however, DF type is likely to play a role. This question is investigated further in Chapter 6.

The results have particular implications for automatic processing. As noted in Chapter 2, fragments are a helpful cue for both DF detection and correction. They are helpful for *detection*, because fragments (nearly) always indicate disfluency.[13] They aid correction because they (nearly) always mark the right edge of the RM.[14]

However, as also noted in Chapter 2, the detection of fragments presents a problem for automatic speech recognition systems that are constrained to output only full words. Recognition of fragments could provide a large win for corpora similar to ATIS; however, the benefits of fragment detection will be reduced for corpora similar to AMEX or SWBD, since there are fewer fragments overall in these corpora. This information should be taken into account when allocating resources for DF detection and correction in a particular domain.

5.5.2 Rate by DF position

Interestingly, the rate of fragments also shows a systematic association with DF position. The conditional probability of a fragment given the DF is sentence-initial differs from the conditional probability of a fragment given the DF is sentence-medial.

Figure 19 shows the same data as presented in Figure 18, but broken down by DF position. Error bars indicate the standard deviation for the proportion. The rate measure is the number of DFs in the position that contain a fragment, divided by the total number of DFs in the position.

---

13. This point should be qualified: on rare occasions, words are left incomplete in the "final-ellipsis" situations described in Chapter 4. In this case they are part of the intended utterance. It should also be noted that by "fragment" here is meant the fragments as transcribed according to the conventions in Chapter 3.

14. Strictly speaking, this is not true for speech errors as labeled by the PLS, since patterns like "gran- troun- ground transportation" may occur.

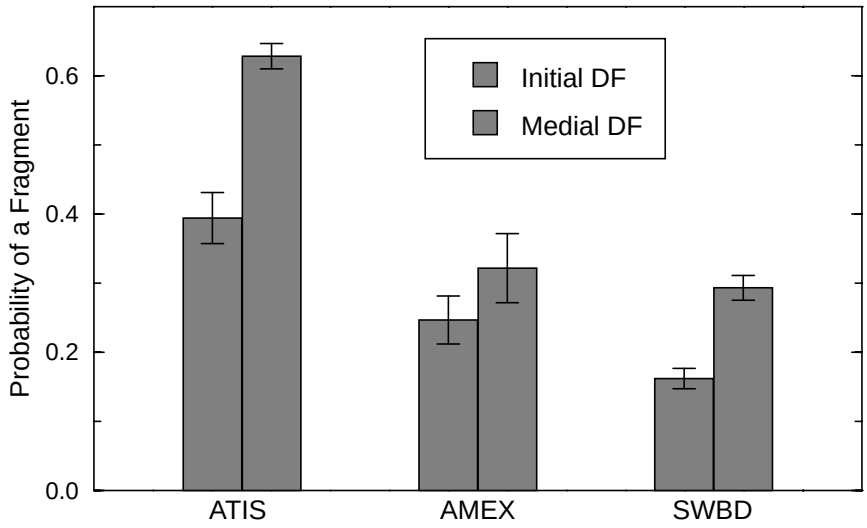


Figure 19. Probability of a Fragment by DF Position

Comparisons of the initial and medial rates within each corpus show a significant difference for AMEX and for SWBD, $z=5.61$ and $p=5.57$, respectively; $p<.05$ in both cases. The comparison for AMEX is not significant, $z=1.26$, $p>.05$; however, the trend appears similar to that of the other corpora but with larger error bars due to the smaller dataset. Thus the AMEX comparison is likely to not be significant due to a lack of power. Tests of proportions also showed that the higher overall rate of fragments in ATIS seen in the previous section reflects a significantly higher rate of fragments in both initial and medial DFs.

It is not clear what explains the higher rate of fragments in medial than in initial DFs, but this is worth pursuing in future work--particularly since it occurs in all corpora. One possibility is that there may be a higher rate of word fragmentation for content words than for function words, as noted by Nakatani and Hirschberg (1994). This could help explain the present result if content words are more likely to occur in medial than in initial position. This question was not further pursued in the present work, but could easily be examined using the information in the HLD. The results in this section suggest that automatic detection of fragments in speech applications could be aided by adjusting the likelihood of a fragment based on the position of the hypothesized DF.

5.5.3 Section summary

Fragment rates in AMEX and SWBD do not differ significantly, while ATIS has a significantly higher rate--roughly double the rate in the other two corpora. The rate of fragments is systematically associated with DF position: in all corpora, medial DFs are more likely to contain a fragment than initial DFs. The overall higher rate of fragments in ATIS reflects a higher rate of fragments in both initial and medial DFs.

5.6 Rate of DFs with Words in Interregnum

This section examines the rate of DFs that have one or more words in the interregnum (see Chapter 4). Potential words include filled pauses (“um” and “uh”), explicit editing phrases (e.g., “oops,” “no”), and discourse markers (e.g. “well,” “you know”). Section 5.6.1 describes the rate of DFs having one or more of any of these types of words in the interregnum; Section 5.6.2 examines rates of each type of word.

5.6.1 Rate overall

This section examines the rate of words in the interregnum. DFs of the type FP (see Chapter 4) are excluded from the analyses, since by definition these DFs always contain a word in the interregnum. The percentage of DFs in each corpus (excluding FPs) having one or more words in the interregnum is shown in Figure 20. Error bars indicate standard deviations.

As shown, across corpora the rate of words in the interregnum is low; only about 12 to 15% of the total DFs contain these elements. This result is important for both cognitive models and automatic processing. For cognitive models, results suggest that human processing of disfluency is unlikely to rely on the presence of words in the interregnum, a notion that runs counter to some previous accounts (see Chapter 2). Similarly, although these elements can alert an automatic system to the presence of a DF in a small number of cases, for the majority of DFs other means for DF detection must be devised.

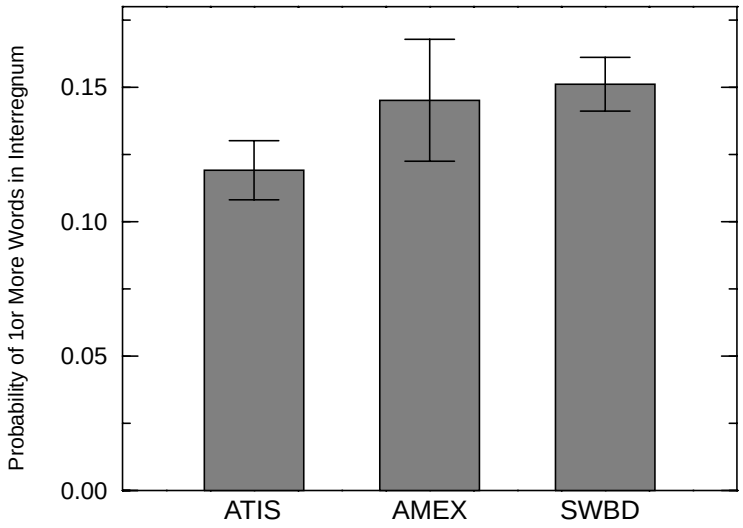


Figure 20. Probability of One or More Words in Interregnum

A second observation is that the rate for ATIS is lower than that for the other two corpora. The rates for AMEX and SWBD are not significantly different, $z=0.24$, $p>.05$. The rate for ATIS is significantly lower than that for SWBD, $z=2.09$, $p<.05$. The comparison between ATIS and AMEX is not significant; however from inspection of the error bars it seems likely this reflects a lack of power as a result of the small sample size for AMEX.

These results suggest a domain difference in rate of words in the interregnum; a better understanding is likely to involve inspection of DF type. Results are also noteworthy for automatic processing, which can make good use of these elements (when present) as cues to the interruption point of a DF. The implication is that although there may be some increase in the rate of these elements for corpora more “natural” than ATIS, the increase is likely to be minimal.

5.6.2 Rate by type of word

Figure 21 shows the data in Figure 20, broken down by type of word (filled pause, explicit editing term, or discourse marker). Error bars are standard deviations; sample size for each bar is equivalent to that for the total DF set within each corpus in Figure 20, since interregnum words can cooccur in DFs. A notable observation is that in all three corpora, the most commonly occurring interregnum word is a filled pause. It is not clear why this is the case.

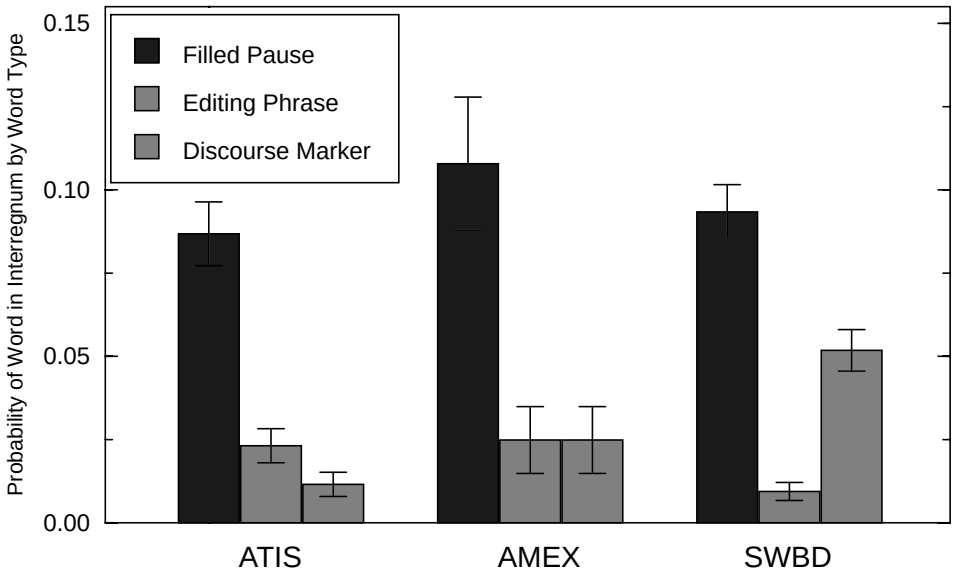


Figure 21. Probability of Word in Interregnum by Word Type

One possible reason, however, is that filled pauses are semantically neutral, and therefore can occur in a wide variety of DF types, including frequent types such as repetitions (e.g. “we uh we”).

A second observation is that the relative rates of explicit editing phrases and discourse markers vary across corpora. In ATIS, editing phrases are more common than discourse markers; in AMEX rates are equal; in SWBD discourse markers are more common than editing phrases. This pattern of results may indicate a greater need for explicit editing in task-oriented dialog (ATIS and AMEX) where precise content is important, and a greater need for discourse markers to manage interaction in human-human dialog (AMEX and SWBD). However, further investigation, including examination of DF type, is required to fully understand this result.

A third point, not observable from Figure 21, is that filled pauses and discourse markers are *unreliable* indicators of longer DFs. In all corpora, over 80% of the total filled pauses occur in isolation, i.e. *not* in the interregnum of a longer DF. Discourse markers were also prevalent in otherwise fluent stretches of speech in AMEX and SWBD, although total numbers are not known since these cases were not labeled for the thesis. Thus, the only fairly reliable cue to the

interregnum of a longer DF is the explicit editing term, which occurred in less than 4% of the DFs in each corpus.

5.6.3 Section summary

The rate of words in the interregnum is low in all three corpora. AMEX and SWBD have statistically indistinguishable rates, while the rate for ATIS is significantly lower. Rate by type of interregnum word shows filled pauses are the most common type in all three corpora, but relative rates of editing phrases and discourse markers vary. Editing phrases are more prevalent in the task-oriented corpora; discourse markers are more prevalent in the human-human dialogs. Neither filled pauses nor discourse markers are reliable indicators of a longer DF, since both occur frequently in otherwise fluent stretches of speech.

5.7 Chapter Summary and Discussion

5.7.1 Summary

The five analysis sections showed regular trends in DFs along a number of different dimensions. The rate of occurrence of DFs was found to show regularities both in the rate of disfluent sentences and in the rate of disfluency per word. Section 5.2 provided a model to predict disfluent-sentence rate from sentence length. Section 5.3 found regular trends in the per-word DF rate by sentence length, by DF position, by cooccurrence of DFs in same sentence, and by combinations of these features. Where tested, trends held for individual speakers.

The rate of pattern features that apply across DF types was also found to show regularities when rate was measured relative to the set of total DFs. Section 5.4 found that a single function and parameter value predicts the distribution of deletion lengths in all corpora. Section 5.5 found reliable differences in the rate of fragments, by corpus and by DF position. Section 5.6 found reliable differences across corpora in the rate of words in the interregnum, both overall and by the type of interregnum word.

5.7.2 Discussion

The overall results lend strong support for the goal of the thesis as stated in Chapter 1: DFs can be modeled along a number of dimensions. For cognitive models, these results imply

that DF production can offer constraints for theories of speech production. Findings also suggest that direct DF modeling could improve the automatic processing of spontaneous speech in applications.

Comparison of results across the chapter sections indicate important themes for future work. These include: 1) interpretation of domain differences; 2) interpretation of domain universals; 3) use of parametric models; and 4) use of a large database.

5.7.2.1 Domain differences

As described in Chapter 3, in choosing the corpora for the present study the AMEX corpus was included as a comparison corpus because it shares domain features with both ATIS and SWBD. Like ATIS, AMEX consists of task-oriented dialog about air travel planning. Like SWBD, AMEX consists of human-human dialog over the telephone. Results show an overwhelming tendency for AMEX to group with SWBD rather than with ATIS. This is an important result for both cognitive models and applications. For cognitive models, it will be important to understand the underlying variables associated with a domain that influence DF production. For applications, we see that research on DFs in the ATIS corpus, which has been the focus of many past studies (e.g., Bear et al., 1992, Nakatani & Hirschberg, 1994) may not predict results for certain aspects of DFs produced in more natural speech styles.

Differences between ATIS and the other two corpora are not only quantitative but also qualitative. ATIS has a lower overall DF rate, but also differs in terms of the relationships among features. For example, in ATIS, the per-word DF rate rises with sentence length whereas in AMEX and SWBD the per-word rate is independent of sentence length. This difference shows up as a difference in the form of the best-fitting model of disfluent-sentence rate (exponential for AMEX and SWBD, but linear for ATIS). Hypotheses about the underlying variables responsible for differences across domains must therefore be able to account for the qualitative as well as quantitative variation in rates across corpora. And, for speech applications, qualitative differences suggest that the nature of the algorithms used for automatic processing methods may differ across corpora.

Across analyses, an unexpected and remarkable finding is the *degree* of similarity between AMEX and SWBD, particularly given that these corpora involve separate sets of

speakers. For example, AMEX and SWBD have statistically indistinguishable values for: disfluent-sentence rate, per-word DF rate, rate of initial DFs, degree of cooccurrence effect, rate of fragments and rate of words in the interregnum. These values do not appear to be universal over domains, since all of these values differed for ATIS. A question for future work, then, is whether the similar values observed for AMEX and SWBD characterize human-human telephone dialog more generally, or whether they reflect some similarity between these domains that is not readily apparent.

5.7.2.2 Domain universals

Although on many measures, ATIS differs from AMEX and SWBD, in other cases results show similar trends for all corpora. Perhaps the most striking result is the invariance of the parameter in the model of deletion length. However, other similarities across corpora include: 1) the shape of the distribution of disfluent sentence lengths; 2) the finding that per-word rates of initial DFs rise with sentence length, 3) the presence of a cooccurrence effect; 4) the higher rate of fragments in medial than in initial DFs; 5) the lack of association between sentence length and deletion length; 6) the better fit of the deletion-length model when filled pauses are removed; and 7) the preponderance of filled pauses in the set of observed interregnum words. If these findings are found to hold for other domains, they will be particularly important to interpret for a unified theory. Universals are also useful in automatic processing, since algorithms based on these trends should achieve success across applications.

5.7.2.3 Parametric models

As stated in Chapter 2, an important way to advance our understanding of DFs is to attain predictive power by directly modeling the observed trends. More specifically, it was argued that *parametric* models are preferred over nonparametric models such as lookup tables of probabilities. Parametric models provide a way to compress the information in a lookup table to an intuitive understanding of the phenomenon. Such models also allow direct comparison of datasets based on the form of the model and/or the parameter values.

Results indicate that parametric models can play an important role in DF research. For example, the parametric model provided in the analysis of disfluent-sentence rate showed AMEX and SWBD to be identical in both the form of the model and in the value of its

parameter, whereas ATIS differed in both of these aspects. The model parameter was interpreted as an aggregate per-word fluency rate, which guided further analyses on per-word rates and rates of cooccurrence. Similarly, the model of deletion length showed all three corpora to be similar both in the form of the model and in the value of its parameter, and a possible theoretical interpretation of the invariant parameter was offered.

Parametric models provide the predictive power necessary to devise and test cognitive theories. They are also useful for applications, since they provide predictions for data not seen in a training set, and may be more straightforward to implement than nonparametric models.

5.7.2.4 Large databases

Finally, results stress the importance of using large databases in DF research. Large databases are crucial for at least three reasons. First, the per-word rate of DFs in natural-sounding speech styles (such as SWBD), is likely to be on the order of 5%-6%; the rate for less natural-sounding corpora (such as ATIS) is likely to be considerably lower. Findings suggest that datasets of the sizes used here for ATIS and SWBD are not unnecessarily large, since a number of analyses indicate a lack of power for the smaller AMEX corpus. Second, breaking down the set of disfluent data by features and combinations of features further reduces the amount of data in each condition. Third, some features of DFs occur reliably but at inherently low rates. Although these rare cases may be unimportant to analyses using inferential statistics, they are important for modeling. For example, estimates of the frequencies of DFs having high deletion lengths was crucial to providing enough range in the deletion-length measure to adequately model the trend. In future work, therefore, it is important to study corpora in which adequate numbers of DFs are available.

Chapter 6: Type-Dependent Analyses

6.1 Chapter Overview

This chapter consists of analyses that refer to DF type, where DF type is defined by the Type Classification Algorithm as described in Chapter 4. The analyses are presented in four main sections.

Sections 6.2 and 6.3 examine selected trends from Chapter 5, but additionally broken down by DF type. Section 6.2 describes the rate of occurrence of the different DF types. Type distributions are examined overall, by speaker, and by DF position. A section on filled pauses examines the distribution of filled-pause forms (“um” and “uh”) by position and by speaker. Section 6.3 examines the rate of pattern features by DF type, including the rate of types with *k* deleted words, the rate of fragments for different types, and the rate of cooccurrence of selected features in substitution DFs. Sections 6.4 and 6.5 focus on specific DF types. Section 6.4 discusses the rate and composition of complex (overlapping) DFs. Section 6.5 discusses acoustic properties of the simplest DF types: filled pauses and repetitions. Section 6.6 summarizes the main findings from each section, and discusses issues that cut across the chapter sections. Comparisons of findings with selected results from Chapter 5 are also discussed.

6.2 Rate of Basic DFs by Type

This section examines the rate of occurrence of DFs by type. The DFs analyzed are “basic” DFs, i.e. those in the set of “clean” disfluent sentences; see Chapter 4. Section 6.2.1 describes the rate of types overall. Section 6.2.2 reports the rate of types by individual speaker in the SWBD corpus, and examines speaker features that correlate with differences in these type distributions. Section 6.2.3 examines the relationship between DF type and DF position. Section 6.2.4 looks specifically at filled pause forms (“uh” and “um”) and reports an association between form and DF position, both across and within speakers.

6.2.1 Rate over speakers

This section examines the rate of DF types in the corpora overall, collapsing over speakers. The first question raised concerns the relative frequencies of types. Use of relative frequencies allows us to visually compare the shapes of the distributions for corpora that vary in overall per-word DF rates (see Chapter 5). In Figure 22, rate is shown as the number of observations of each type divided by the total DFs in the corpus. Thus, frequencies are percentages of total DFs in a corpus and sum to 1.0. Types are (arbitrarily) arranged on the abscissa from highest to lowest frequency when summed over the three corpora. Note this is a categorical scale; points are connected for each corpus only for ease of visual inspection.

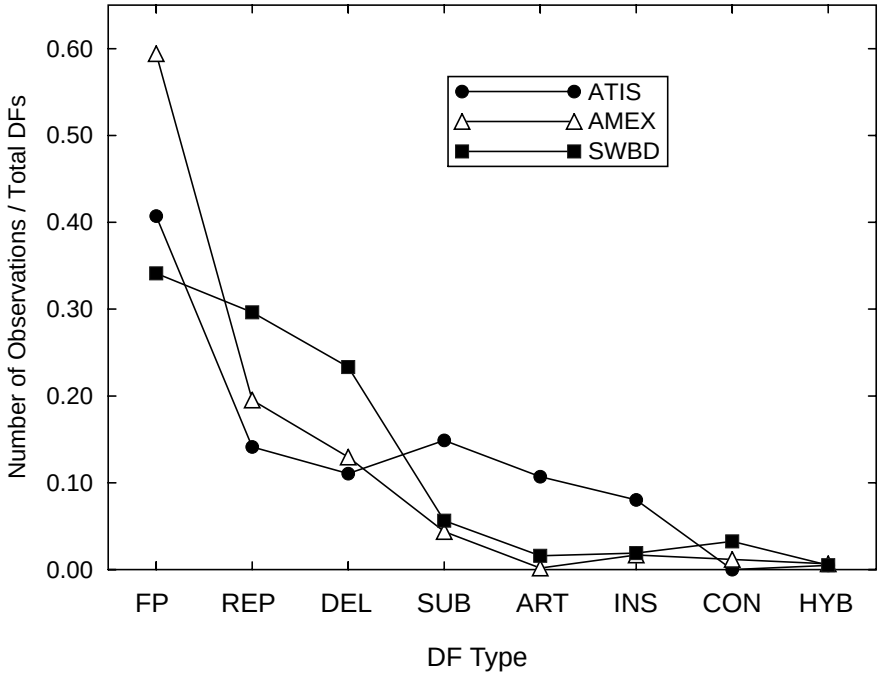


Figure 22. Rate of DF Types (per Total DFs)

The first question investigated is whether the distribution of types within each corpus is significantly different from uniform. The null hypothesis is that within a corpus, each type is equally likely; the alternative hypothesis is that some types are reliably more likely than others.

Results of the appropriate tests are shown in Table 17. The Chi square coefficient is provided as a test of significance. In addition, since the Chi square coefficient is sensitive to sample size, the Cramér coefficient is provided as an index of the magnitude of association between corpus and type. The index ranges between zero (no association) and 1 (complete association); therefore it can be used to compare strengths of association across corpora differing in sample size.

Table 17: Tests of Hypothesis that Types are Equally Likely Within a Corpus

Comparison	Degrees of Freedom	χ^2 Coefficient	p	Cramér Coefficient
ATIS	1,7	583.97	<.0001	.448
AMEX	1,7	454.92	<.0001	.619
SWBD	1,7	1075.11	<.0001	.527

As shown, for all corpora the type distribution is significantly different from that expected under the null hypothesis. Of further note is that AMEX has the greatest degree of association as reflected by the Cramér coefficient. This coefficient is maximized by sharper distributions, i.e. by maximal differences in rates of individual types. Inspection of Figure 22 indicates that the high degree of association for AMEX is largely attributable to a high rate of the type FP. FPs alone account for roughly 60% of all DFs in that corpus. ATIS has the lowest degree of association; after a high rate of FPs, rates for REP, DEL, SUB, ART and INS are fairly similar.

Given that the distribution of types in each corpus differs from a uniform distribution, a second question is whether the distributions differ across corpora. A test comparing the three distributions yields $\chi^2(2,7)=646.89$, $p<.0001$, indicating differences among at least two of the corpora. Results of pair-wise comparisons between the corpora are shown in Table 18.

As shown, the null hypothesis is rejected for each pair-wise comparison. Thus, the type distribution in each corpus is significantly different from a uniform distribution, and also significantly different from the distribution in either of the other two corpora. However, note that the value for the Cramér coefficients in Table 18 are considerably lower than those in Table 17. This indicates that although the distributions differ significantly across corpora, there is nevertheless some similarity across corpora in the *manner* in which trends deviate from a uniform

Table 18: Tests of Hypothesis that Corpora Have Same Type Distribution

Comparison	Degrees of Freedom	χ^2 Coefficient	p	Cramér Coefficient
ATIS vs. AMEX	1,7	182.10	<.0001	.298
AMEX vs. SWBD	1,7	126.18	<.0001	.223
ATIS vs. SWBD	1,7	481.47	<.0001	.377

distribution. From inspection of Figure 22 it appears this similarity is largely attributable to the consistently high rates of FPs, and to a lesser extent, to the high rates of REP and DEL. The details of these distributions will not be discussed in this section (but see below). However, it is interesting to note that the comparison corpus, AMEX, shows a rate profile similar to ATIS for certain types (FP, REP, and DEL) but a profile similar to SWBD for other types (SUB, ART and INS.) This result suggests similarities between specific domains in *relative* rates of DF types; however, it is difficult to interpret a difference in relative rates without also having an understanding of absolute rates, i.e., rates of each type per word.

Rates per word (number of observations of a type, divided by the total words in the corpus) are shown in Figure 23, which is otherwise arranged identically to Figure 22. Within each corpus, the sum of the rates in Figure 23 equal the overall per-word DF rates (*d*) from Chapter 5. The notable difference between Figure 22 and Figure 23 is the lower overall vertical scaling for the ATIS distribution. As described in Chapter 5, ATIS has an overall per-word DF rate, as estimated by 1-*b*, of .007, whereas ATIS and AMEX have rates of roughly .055.

Figure 23 shows an important result. A distinction can be drawn between a group containing the types SUB, ART, and INS, and a group containing the types FP, REP, and DEL.[1] SUB, ART, and INS are types which have similar, and low, per-word rates in all corpora.[2] FP, REP and DEL are DFs that are present in higher numbers in the two human-human corpora.

---

1. Rates of the types CON and HYB are discussed later in this section.

2. The rate for articulation errors is unreliable in AMEX because the author did not fully check the transcriptions. It is notable that for ATIS and SWBD, rates of articulation errors were quite close: .00083/word for ATIS and .00077/word for SWBD. However, these rates are considerably higher than the .00015/word rate reported by Deese (1980). This could be due to differences in the speakers or domain, or it could reflect differences in the labeling of misarticulations.

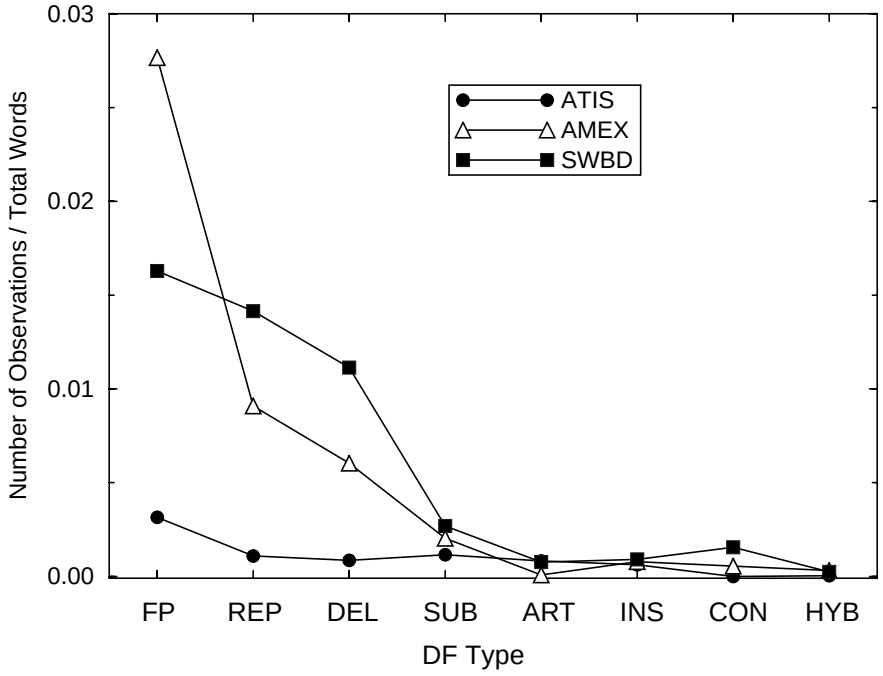


Figure 23. Rate of DF Types (per Word)

The trends in Figure 23 suggest a possible difference in the *source* of these two different groups of types. The types that show low and roughly equal per-word rates across corpora may reflect basic problems in message construction. The types that are more prevalent in human-human dialog may function to coordinate exchanges with a conversational partner. Distinguishing DF types along these lines is an important area for future work.

These results also have methodological implications. The type groupings suggested do not tend to appear in previous classification systems. This may be because they are unlikely groupings based on surface similarities of the types. As discussed in Chapter 2, surface similarity has guided groupings in many previous systems. The particular oddity in this case is the grouping of DEL with FP and REP.

FP and REP have been grouped together in pervious classification systems because both lack a change in wording from the RM to the RR (see Chapter 2). DEL, on the other hand, has often been grouped with DFs like SUB, since it involves a change in wording. Thus the trends in Figure 23 illustrate an advantage of the systems developed for DF representation and classification in the present work (Chapter 4). These systems avoid intuitive hierarchical grouping of basic classes, and aim instead to group and regroup basic classes based on empirical trends such as those encountered in the present section.

An interesting question to explore in future work is *why* DEL is grouped with FP and REP. One possibility is that DEL serves a function common in human-human dialog, but different from the proposed hesitation function served by FP and REP. An alternative possibility is that DEL serves a function similar to FP and REP. In this case, an important issue is what determines the speaker's use of one of these forms over another. This issue is discussed further below.

Some mention should be made of the two least frequent types across corpora: CON and HYB. The overall rate of CON is low in all three corpora. This behavior is at least partially explained by the restricted context in which CON occurs. By definition (see Chapter 4), CON occurs before a sentence. Typically it is found *within* a turn rather than at the start of a turn. Therefore, it tends to occur in turns having at least two sentences.

Figure 24 shows distributions of the number of sentences per turn in the three corpora.[3] As shown, less than 40% of the turns in SWBD contain multiple sentences; for AMEX and ATIS this rate is even lower. Therefore, a low overall rate of CON is expected given the limited number of non-turn-initial sentences in the corpora.

In addition, Figure 24 is consistent with the relative rates of CON in the three corpora. As can be (barely) seen from Figure 22, the per-DF rate of CON is lowest for ATIS and highest for SWBD. The results in Figure 24 are consistent with this ordering: ATIS shows the fastest decline in the frequency of multiple-sentence turns; SWBD shows the slowest decline.

Finally, the type HYB was extremely infrequent across corpora. This low rate has potentially important implications for the classification system described in Chapter 4. Recall that

---

[3.] Modeling of these trends showed they follow a roughly exponential rate of decay (Shriberg, 1994).

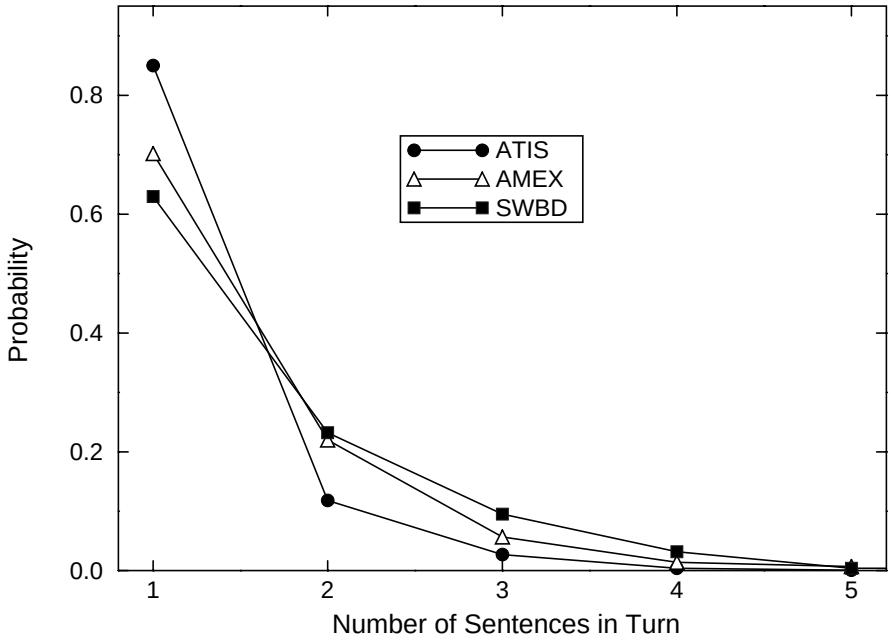


Figure 24. Distribution of Turn Lengths (in Sentences)

the symbols 's', 'i', and 'd' (representing operations of substitution, insertion, and deletion, respectively) were placed in the same level of the symbol ordering used for classifying DFs. HYB was a class created for cases in which two or more of these symbols cooccurred in a pattern.

The scarcity of observed cases of HYB suggests we may consider eliminating this type in future classification systems. This elimination of HYB is additionally supported by the finding that in all observed cases of HYB, an alternative analysis was possible (DEL or SUB). Interestingly, the elimination of HYB from the Type Classification Algorithm would attach theoretical significance to the symbol ordering proposed in Chapter 4. This is because the world of possible patterns would in that case exclude patterns containing more than one symbol from the same level in the ordering. Since the three operations at the same level of the ordering (insertion, deletion, and substitution) all involved a change in wording from RM to RR, this could have implications for cognitive models. It may suggest that speakers do not change more than one thing at a time when

making a repair (see also Chapter 4). The question of whether to eliminate HYB is therefore both a methodological and theoretical issue for future work.

6.2.2 Rate by speaker

This section examines the relative rate of DF types by speaker. Data are from the 30 SWBD speakers; in the ATIS and AMEX corpora there were not enough DFs per speaker available for analysis. Note, however, that because the three corpora have different type distributions, the results in this section may not predict rates by speaker in the other two corpora.

6.2.2.1 "Repeaters" versus "Deleters"

For each of the 30 SWBD speakers, a normalized type distribution was computed by dividing the number of occurrences of each type by the total DFs produced. Figure 25 shows type distributions for the 20 speakers having the most data. Lines connect rates for a particular speaker. The sum of the rates over the eight DF types for each speaker is 1.0. However, the type FP is not shown in Figure 25 (FP rates are described in a later section); therefore the rates shown in Figure 25 sum to less than 1.0 for speakers that produced FPs. The aggregate distribution (from Figure 22) is also indicated.

The distributions in Figure 25 indicate that the aggregate distribution from Figure 22 does not describe the behavior of individual speakers. Rather, the aggregate data reflect an average of two clearly different distributions. The differences between these distributions lie in the rates of the two most common DF types included in the figure: REP and DEL.[4] Certain speakers, who will be referred to as "repeaters," show much higher rates of REP than DEL. Other speakers, who will be referred to as "deleters," show much higher rates of DEL than REP. Rates of REP and DEL show a significant inverse correlation within speakers, r^2=-.7203, p<.01.

4. FPs are excluded from Figure 25 as mentioned in the text. The remaining types occur at consistently low rates across speakers. It is possible that there are systematic differences among these types by speaker; however, a much larger set of data would be needed in order to discern such differences.

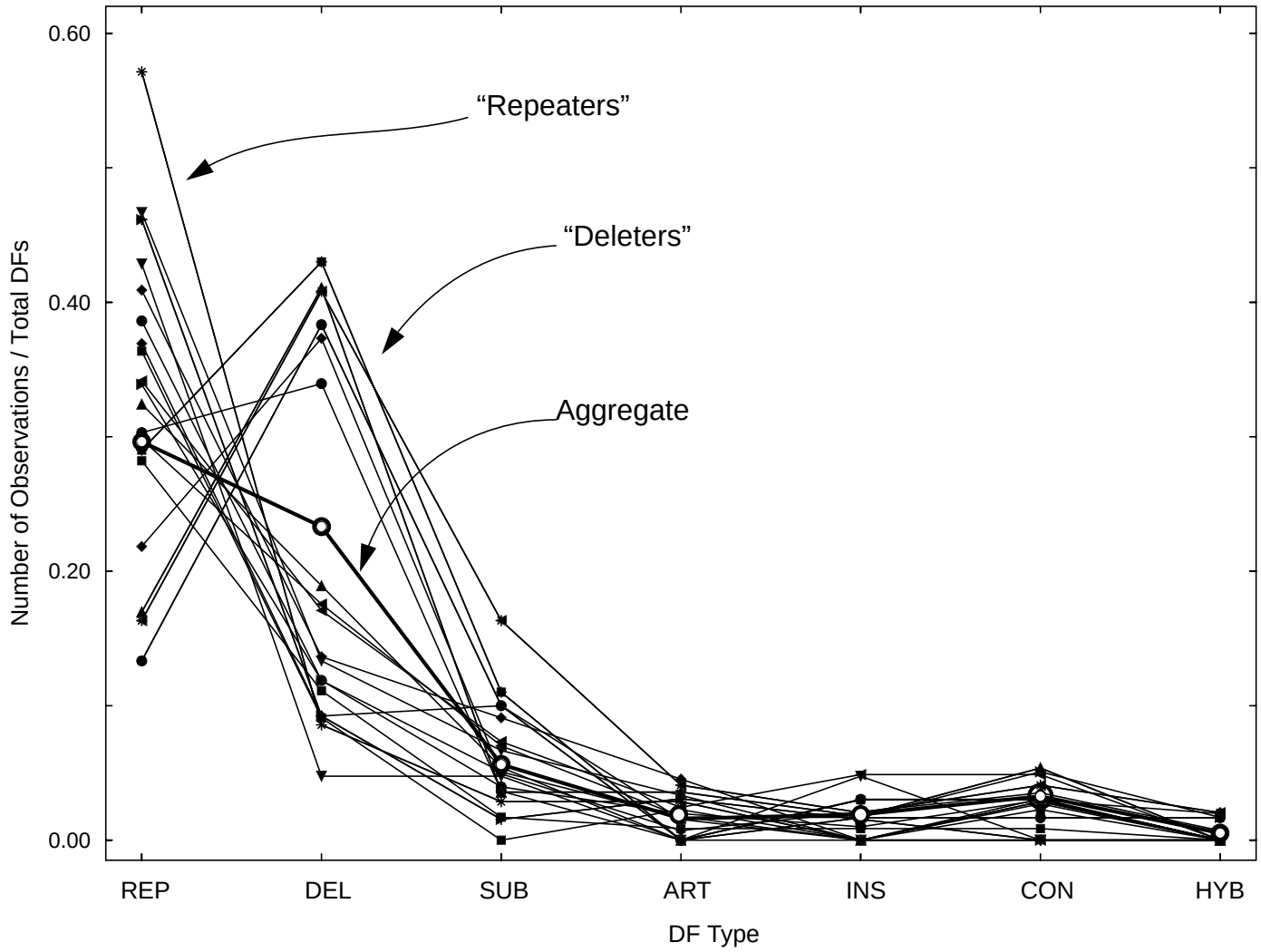


Figure 25. Rate of DF Types (per DF) by Speaker. (Lines connect points for an individual speaker.)

6.2.2.2 Internal consistency

This section asks whether speakers are internally consistent in making either more REPs or more DELs. This question was addressed in a simple test by comparing, for each speaker, the relative rates of REP and DEL in the speaker's two conversations (see Chapter 3). Results showed that 23/30 speakers were consistent over both conversations in producing either a higher rate of REPs or a higher rate of DELs. This is significantly more speakers than predicted by chance, since there is a 50% probability of a "same" outcome in both conversations, $z=2.92$, $p<.05$.

Thus, the characterization of a speaker as a repeater or deleter in Section 6.2.2.1 appears to reflect a reliable tendency for that speaker to produce higher rates of the relevant type in individual conversations. In future work, much could be learned from analyses of the consistency of type distributions within speaker, particularly if aspects of interest (such as the domain) were independently varied. Such analyses, however, require a large amount of data per speaker, especially in order to compare results for the classes that occur at consistently low rates. These findings also suggest that in speech applications, modeling of individual differences in DF type distributions should provide an advantage over modeling an aggregate distribution that does not describe the behavior of any particular speaker.

6.2.2.3 Associated features

This section investigates speaker characteristics that correlate with the tendency for a speaker to be a repeater or a deleter. The first characteristic examined is the speaker's overall rate of DFs (DFs per efficient words). Results are shown in Figure 26. The height of each bar reflects the mean of the speaker-specific DF-rate measures in the condition (i.e. each speaker was weighted equally). Error bars indicate standard deviations.

There was no significant difference in overall DF rates between the two groups, $t=.19$, $p>.05$. Both groups had about the same rate as found for the SWBD corpus overall in Chapter 5 (about 5.5 DFs per 100 words). The tendency to be a repeater or deleter also did not correlate with a number of sentence-related measures, including the speaker's rate of disfluent sentences, mean sentence length for fluent sentences, and mean sentence length for disfluent sentences.

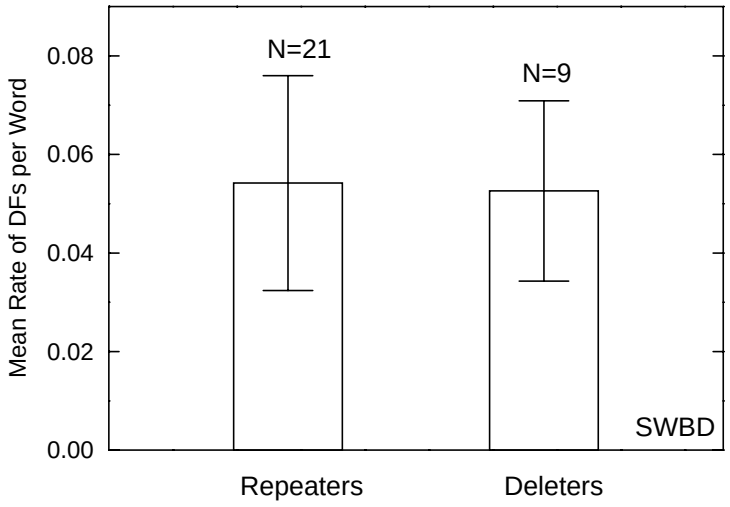


Figure 26. Rate of DFs (per Word) for Repeaters and Deleters

However, a difference was found for speech rate. Speech rate was computed as described in Chapter 4; results for 29 of the 30 speakers were used, because the speech rate measure was unreliable for one speaker as noted in that chapter. This speaker was one of the nine deleters, leaving that set with eight speakers. The speech rates for the 29 speakers ranged from 2.84 to 4.14 words/second, with a mean of 3.39 words/second and a standard deviation of 0.29 words/second.

Figure 27 shows the mean speech rate for repeaters and for deleters. Error bars indicate standard deviations. The mean speech rate for repeaters was 3.29 words/second; for deleters it was 3.64 words/second. Despite the size of the standard deviations, the difference in mean rate between the two groups is significant, (t=3.42, p<.01). Rates of REP and DEL were also significantly correlated with speech rate. The correlation between REP and speech rate was negative, r^2=-.4920, p<.01; that between DEL and speech rate was positive, r^2=.5273, p<.01. Thus, the deleters in these data were reliably faster speakers (on this measure of speech rate) than the repeaters.

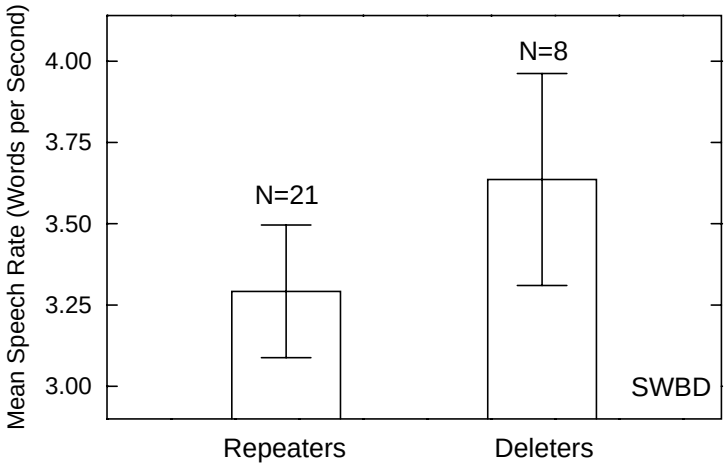


Figure 27. Speech Rate for Repeaters and Deleters

We must ask, however, whether this difference could be accounted for by a difference in the number of deleted words in these two types, since the speech rate measure counted total (rather than efficient) words. In fact, DELs do have more words deleted than REPs (see Section 6.3.1 for values of model parameters for deletion length in these two types). In the SWBD corpus, the average number of deleted words in REPs was 1.27, while for DELs this was 2.12. However, this difference alone cannot explain the difference in speech rates between the two groups. Even in the extreme case of a speaker producing only REPs or only DELs, given the overall DF rate (.06 per word) and proportion of DFs accounted for by these two types (about .54), the difference attributable to the longer deletion length for DEL could account for only about 20% of the observed difference in speech rate between the two groups.

Thus: 1) repeaters and deleters have the same rate of DF per word; 2) repeaters and deleters have similar characteristics for sentence features such as mean sentence length; and 3) deleters produce more words per unit time than repeaters, even after adjusting for the higher average number of words deleted in DELs than in REPs.

These results raise interesting possible interpretations of cross-speaker differences in rates of REP and DEL.[5] One possibility is that REP and DEL reflect different underlying problems in speech production. Under this view, which is consistent with suggestions in past work in psycholinguistics (see Chapter 2), REPs reflect hesitation, while DELs reflect some type of error. The inverse relationship between REP and DEL, and its relationship to speech rate, is explained by assuming faster speakers get ahead of themselves and therefore make more errors, whereas slower speakers take more time to plan, increasing hesitations but reducing errors.

An alternative possibility is that REP and DEL do not necessarily reflect different underlying problems. Rather, both may be employed in situations where the speaker does not yet have speech planned. This view is consistent with past work in conversation analysis, which suggests that a variety of DF types can be used as a provisional start, as well as at turn beginning to secure a listener's gaze or attention (see Chapter 2). Cross-speaker differences in the production of REP and DEL would in this case correspond to individual differences in preference for one form or the other in coping with the same underlying problem.

The first view is more readily related to the speech rate difference, since it seems reasonable that speech with fewer hesitations should be more prone to error. Under the second view, there is no basis upon which to relate speech rate to preference for a particular hesitation form. However, the second view fits better with the results in Section 6.2.1, where per-word rate comparisons across corpora showed DEL to group with FP and REP, rather than with a group containing true errors (e.g., ART and SUB). An important goal for future work is to evaluate these alternative possibilities, since an understanding of what gives rise to the different DF types should be an integral part of a unified theory.

Interestingly, these results may explain an informally observed perceptual effect. It has been reported to the author that faster speakers are more difficult to understand and seem to make more DFs than slower speakers. The present results showed that overall, faster speakers (deleters) made the same number DFs per word as slower speakers (repeaters)--although it is true that the faster speakers made more DFs *per unit time* since they produced more speech per unit time. The perceived effect may be attributable to the proportions of DF *types* produced. Evidence from Fox

---

5. Note that this discussion applies only to differences in REP/DEL ratios *across* speakers. It says nothing about the use of one form over the other for particular contexts *within* speakers.

Tree (1993) suggests that DELs are relatively more difficult for listeners to process than REPs, as measured by reaction time to a target phoneme following the DF. Therefore, it is possible that faster speakers are harder to understand not because they make more DFs overall, but rather because they produce a greater proportion of the types of DFs that impair comprehension.

These results are also potentially relevant to speech applications. As mentioned earlier, modeling speakers as either repeaters or deleters could provide an advantage over using average rates for these frequent DF types, since average rates do not reflect distributions for either kind of speaker. Since speech rate is correlated with type distributions, and is a measure that could be estimated automatically, it could be used in making hypotheses about whether a particular speaker is more likely to be a repeater or a deleter.

6.2.2.4 Rate of FPs and associated features

As noted in the previous section, the speaker-specific type distributions shown in Figure 25 did not include FPs. This section examines the rate of FPs by speaker, and explores speaker characteristics with which this rate is associated.

Figure 28 shows rates of REP and DEL from Figure 25, but additionally including rates of FPs from the same speakers. Lines connect points for individual speakers. The rate measure is the number of cases of the DF type divided by the total DFs produced by the speaker. As shown, most speakers made relatively high rates of FPs overall, consistent with the high rate of FPs for the aggregate data in Figure 22.

A question of interest was which features of the speaker could predict the rate of FPs. FP rates showed no significant correlation with the rate of REPs, as is suggested in Figure 28. FP rates did, however, show a significant inverse correlation with the rates of two DF types: DEL (r^2=-.6391, p<.01) and SUB (r^2=-.5606, p<.01). The significant inverse correlation of DEL with FP, but lack of significant correlation of REP with FP is a concern, since DEL and REP were found to be significantly correlated (Section 6.2.2.1). However, this result reflects the tight bimodal distribution of individual speaker values for DEL; both FP and REP show much wider spreads. Therefore, error variance is lower for comparisons involving DEL. The significant inverse correlation between FP and SUB suggests that speakers who produce higher rates of FPs may be

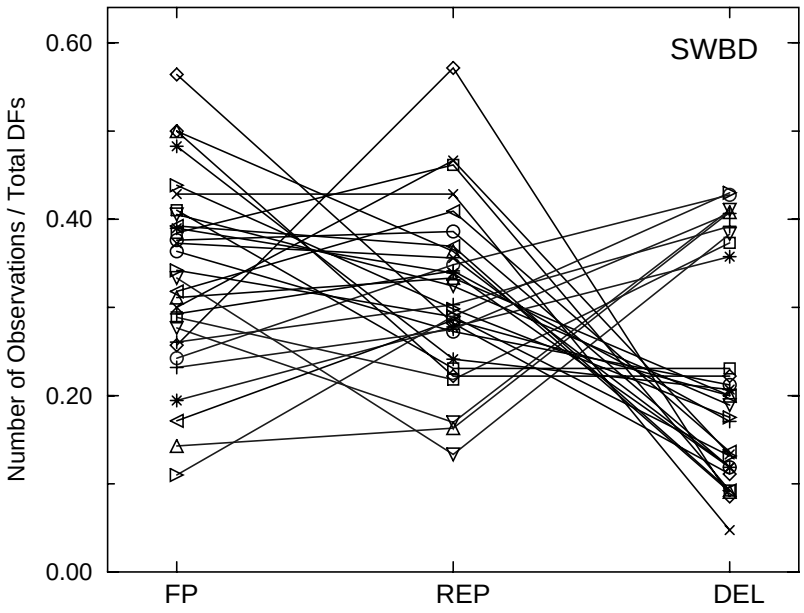


Figure 28. Rate of FPs, REPs, and DELs, by Speaker. (Lines connect points for an individual speaker.)

lowering their production of errors. However, we cannot infer a causal relationship from these data; the inverse correlation between FP and SUB could reflect effects of a latent variable.

FP rate did *not* correlate with a number of sentence production measures, including a speaker's: 1) rate of disfluent sentences; 2) mean fluent-sentence length; 3) mean disfluent-sentence length in total words; or 4) mean disfluent-sentence length in efficient words. In addition, despite the significant correlation between FP rate and DEL rate, and the association between deleters and speech rate described in the previous section, FP rate was not significantly correlated with speech rate.

FP rate did, however, correlate with gender. Men produced higher rates of FPs than women, as shown in Figure 29. This correlation was marginally significant when rates were computed as a percentage of the total DFs, t=1.96, p<.05, one-tailed. However, when computed as the rate of FPs per word, as shown in the figure, internal variability (particularly for the male speakers) was reduced, and the difference was significant by a comfortable margin, t=3.70, p<.05. It is worth noting that the high standard deviations for the men did not reflect a distribution skewed by a few very high values; the distribution appeared to be roughly normal, as did the distribution

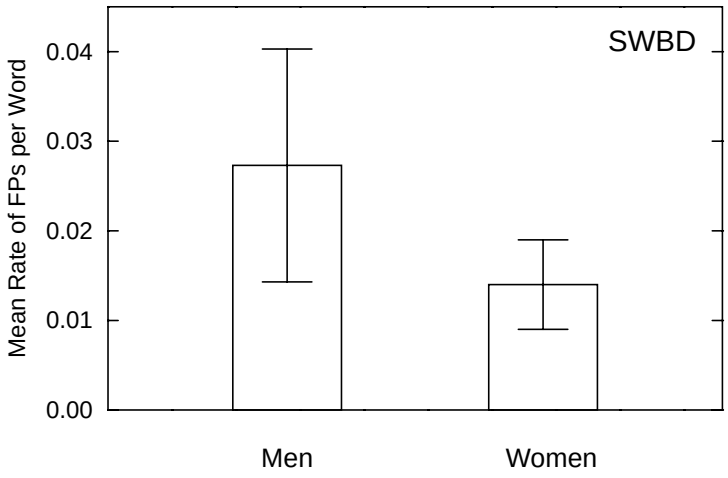


Figure 29. Rate of FPs by Gender

for the women. Thus it is not the case that the higher rates for the men were attributable to only a few individual speakers.

These results reveal an interesting difference between FPs and the other types of DFs included in this study. FPs in these data are associated with a sociolinguistic variable, whereas rates of the other DF types are either fairly similar over speakers, or are associated with characteristics of the speech produced. The higher rate of filled pauses for men is interesting in light of the view that filled pauses may serve to "hold the floor" (see Chapter 2). An inference is then that men may tend to control the floor to a greater extent than women. (Note that this does not imply that men spend more time speaking in conversation).

However, while this is an interesting issue for future work, it would be premature to draw a direct connection between gender and floor-holding. First, as mentioned in Chapter 3, gender in the SWBD corpus was correlated with other sociolinguistic variables, for example education level and occupation. Second, we do not know whether filled pauses actually function as floor-holders, since experiments designed to test this hypothesis have shown mixed results (see Chapter 2). Third, inferences about gender differences should take into account the gender of the listener as well as that of the speaker; this was not addressed in the present analysis.

6.2.3 Rate by position

We know from Chapter 5 that the rate of initial DFs is higher in all corpora than the rate of medial DFs when rates are computed over all DF types. The question addressed in this section is whether relative rates of initial and medial DFs are associated with DF type. The null hypothesis corresponds to a uniform ratio of initial DFs to medial DFs across types within a corpus.

Figure 30 shows the frequency of total DFs, and the frequency DFs by position (initial or medial) for each corpus. Comparisons of the type distributions for initial and medial DFs in each corpus are presented in Table 19. As shown, the Chi square coefficient is significant for all

Table 19: Comparison of Type Distributions for Initial and Medial DFs within Corpora

Comparison	Degrees of Freedom	χ^2 Coefficient	*p*	Cramér Coefficient
ATIS	1,7	127.05	<.0001	.295
AMEX	1,7	43.93	<.0001	.272
SWBD	1,7	207.18	<.0001	.327

comparisons, indicating an interaction between DF position and type distribution within each corpus. Furthermore, although there are differences in the shape of the initial-DF and medial-DF type distributions across corpora, it is noteworthy that corpora are fairly similar in the *degree* of association between position and type (as shown by the Cramér coefficient). Although at present it is not clear how to interpret this observation, such second-order comparisons are important to consider in future work.

When rate is computed as the number of observations divided by the number of potential sites, however, quite a different picture emerges. From Figure 31 we see that in all corpora, FP, REP, and DEL are clearly more likely in initial than in medial position:

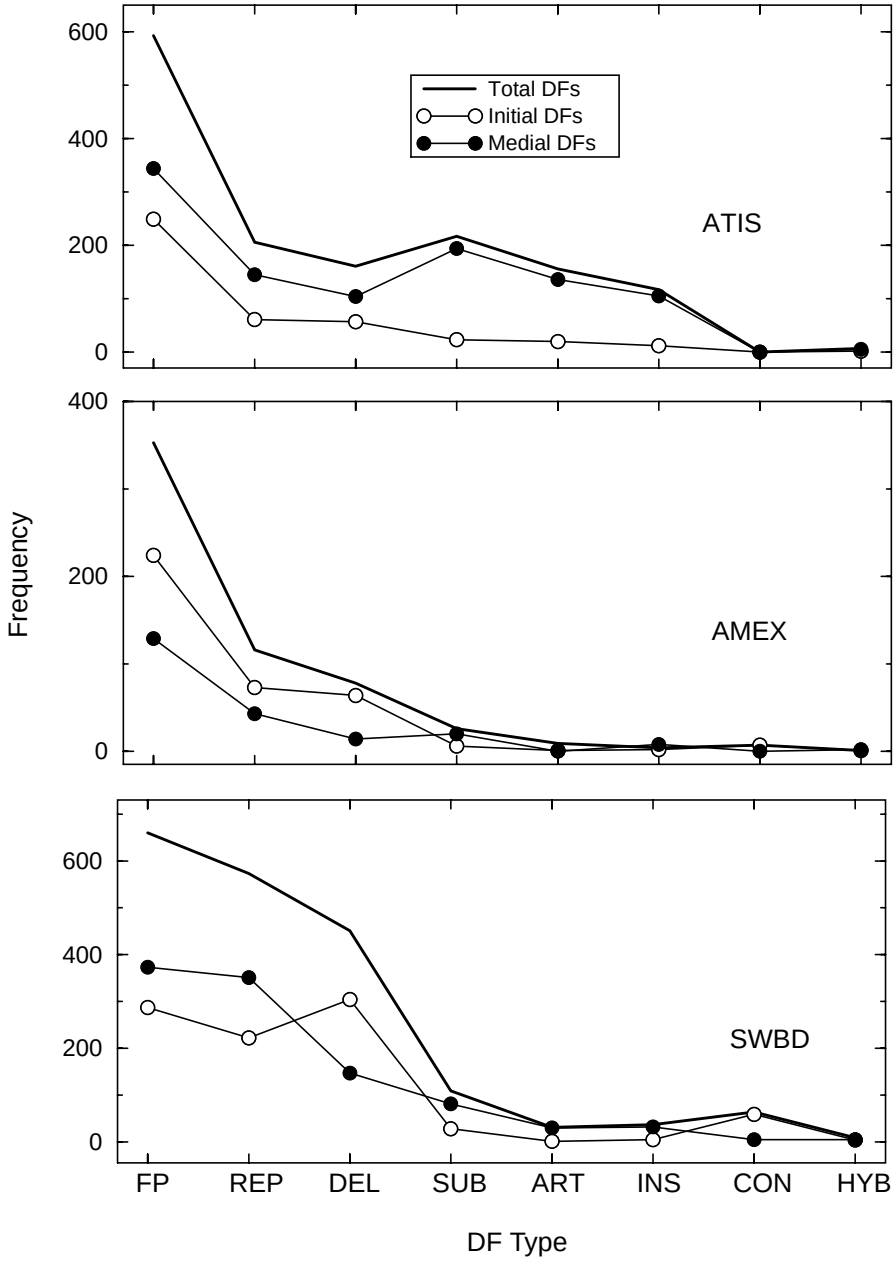


Figure 30. Rate of Types (per DF) by Position

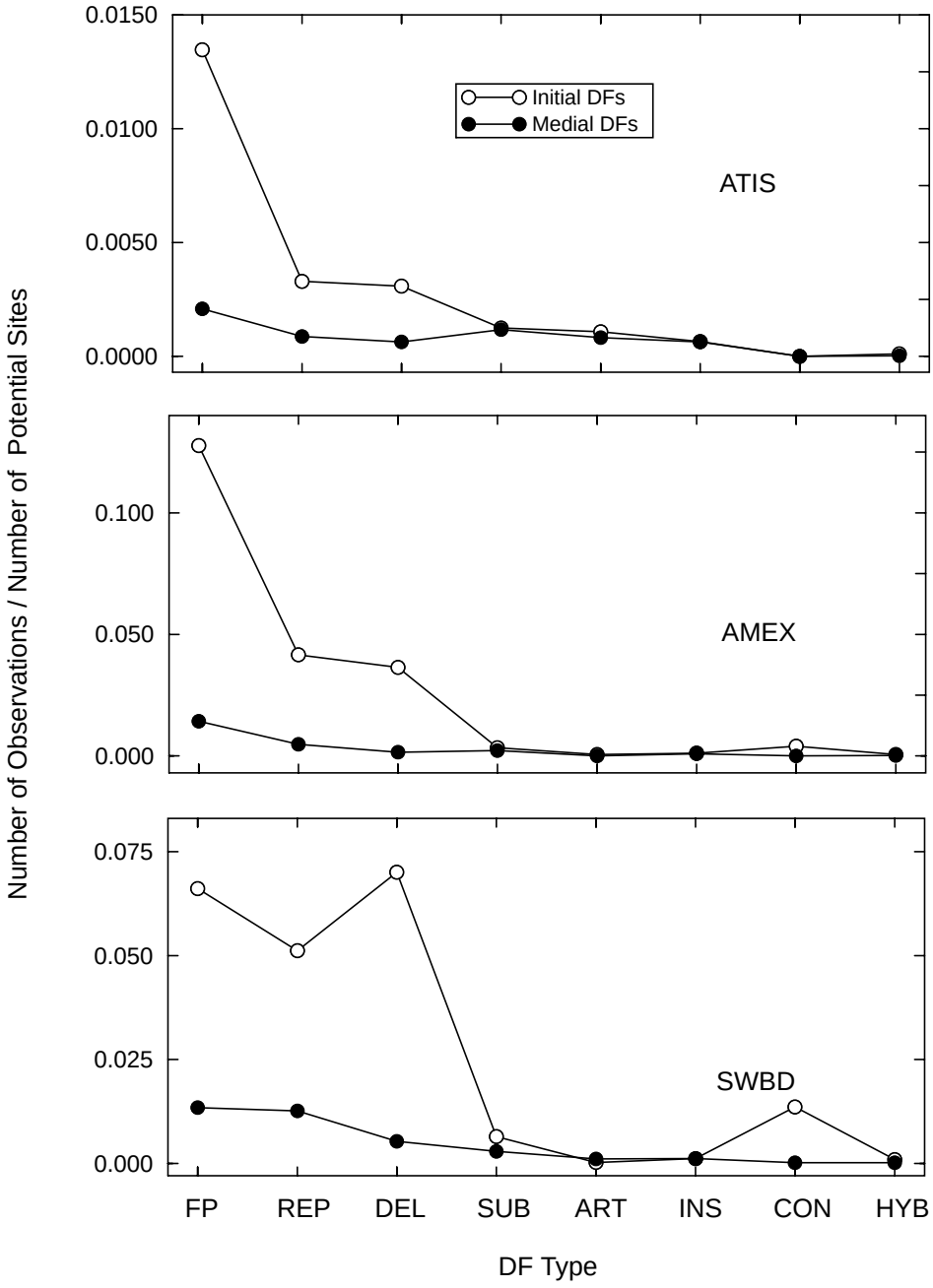


Figure 31. Rate of Types (per Word) by Position

However, for the rest of the types, rates of initial and medial DFs appear to be similar[6] (although it is possible that low sample size for these types obscures differences). This suggests an association between initial position and the types FP, REP, and DEL. The association is consistent with two previous findings: 1) the overall higher rate of initial than medial DFs (Chapter 5); and 2) the high rate of the types FP, REP, and DEL (Section 6.2.1). Recall also from Section 6.2.1 that rates of these three DF types were highest in the two human-human corpora. Taken together, these results suggest that FP, REP, and DEL may serve a discourse function, such as the coordination of turn exchanges and/or the maintaining of the floor between utterances within a turn

It can also be seen from Figure 31 that AMEX groups with ATIS in terms of the shape of the distribution of initial DFs. These two corpora show a very high rate of FPs in initial position, and lower and roughly equal rates of REP and DEL. SWBD, on the other hand, shows a different pattern, with high rates for all three types. One possible interpretation is that FPs are used when the beginning of an utterance is not yet planned, while REP and DEL can be used when the beginning or approximate content of the utterance is planned. Task-oriented dialog may have relatively more FPs because it is more content-oriented and requires precision; in free conversation REP and DEL may be used frequently because there is less of a premium on precise content. However, it is not possible to evaluate these possible explanations in the present work; explaining the relationship between speech task and the rate of particular types of DFs remains an interesting area for future study.

These results also have implications for speech applications. First, a DF in initial position is much more likely to be a FP, REP or DEL than any of the other DF types. Second, the relative likelihoods of initial FP, initial REP and initial DEL depend on the domain. For example, predictions of rates of REP and DEL based on ATIS will underestimate these rates for SWBD.

6.2.4 Rate of FP forms by position and speaker

As mentioned in Chapter 2, Clark and colleagues (e.g., Smith & Clark, 1993) have found that the FP forms “uh” and “um” differ with respect to the length of following pauses. Long pauses

6. An exception is the rate of CON in the SWBD corpus. However, by definition this type occurs only in initial position (see Chapter 4).

are more likely to be preceded by “um” than by “uh”; use of one FP form over the other could therefore inform a listener about the seriousness of an upcoming problem.

In the present work, it was found that the forms “um” and “uh” also show systematic differences in sentence position, both across and within speakers. Table 20 shows data for “um” and “uh” across speakers, by position and corpus. Cells show observed counts, expected values, and residuals, respectively.

Table 20: Rate of FPs by Form and Position

ATIS	“uh”	“um”	Row Total
Initial	137 (156.3) -19.3	106 (86.7) 19.3	243 41.5%
Medial	240 (220.7) 19.3	103 (122.3) -19.3	343 58.5%
Column Total	377 64.3%	209 37.5%	586 100.0%

AMEX	“uh”	“um”	Row Total
Initial	136 (154.3) -18.3	82 (63.7) 18.3	218 64.3%
Medial	104 (85.7) 18.3	17 (35.5) -18.3	121 35.7%
Column Total	240 70.8%	99 29.2%	339 100.0%

Table 20: Rate of FPs by Form and Position

SWBD	"uh"	"um"	Row Total
Initial	162 (187.0) -25.0	115 (90.0) 25.0	277 43.5%
Medial	268 (243.0) 25.0	92 (117.0) -25.0	360 56.5%
Column Total	430 67.5%	207 32.5%	637 100.0%

Chi square tests for association between FP form and position within each corpus yield significant associations for all three corpora, as shown in Table 21. The degree of association, however, is fairly small. Thus while the association is reliable, the distributions of "uh" and "um" by position largely overlap.

Table 21: Tests of Association Between Filled-Pause Form and Position

Corpus	Degrees of Freedom	χ^2 Coefficient	p	ϕ Coefficient
ATIS	1	11.45	<.001	.140
AMEX	1	20.90	<.0001	.248
SWBD	1	18.18	<.0001	.169

As indicated in Table 20, "um" tends to occur in initial position (and "uh" in medial position) in all corpora. A possible interpretation of this result is that "um" and "uh" reflect different underlying problems in production. The form "um" may be used relatively more often during planning of larger units, while "uh" may be relatively more likely to reflect local lexical decision-making. A goal of future work is to determine whether these results have any connection to those of Clark and colleagues, and to gain a better overall understanding of the differential use of these forms by both speakers and addressees.

In addition to the significant associations within each corpus, it is interesting to note that in all three corpora "uh" was overall more frequent than "um."[7] Furthermore, a striking finding is that when the 2x2 tables of data by FP form and position are collapsed to 1x4 tables and compared across corpora, ATIS and SWBD have virtually identical percentage values in the four cells, as shown in Figure 32 (points are connected for each corpus for ease of inspection).

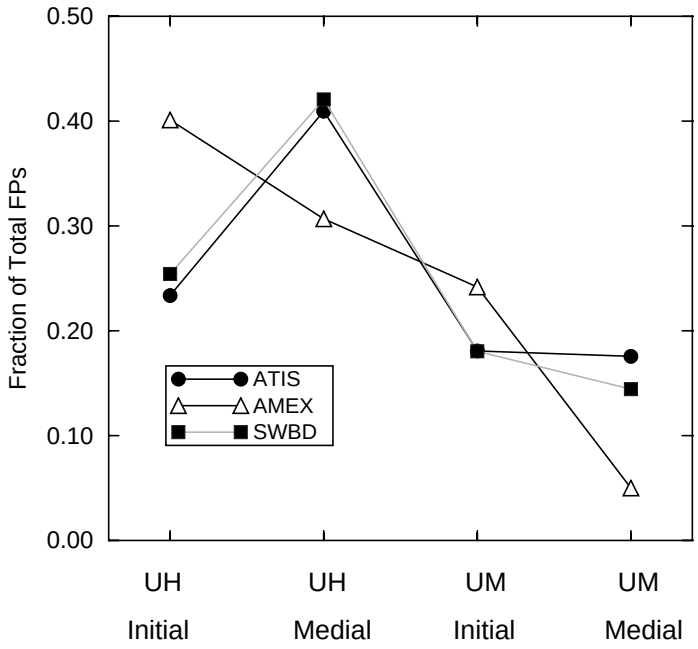


Figure 32. Rate of FP Forms by Position

As shown, ATIS and SWBD values are quite close, while those for AMEX differ considerably. Chi square tests for each of the 2-way corpus comparisons of trends in Figure 32 are shown in Table 22. As indicated, we cannot reject the null hypothesis that filled-pause forms in initial and in medial position are drawn from the same distribution for ATIS and SWBD, while we can reject this hypothesis for comparisons between AMEX and each of the other two corpora.

[7.] Lickley (1994), however, found "um" was more common than "uh" for six British English speakers in informal conversation. This may indicate a dialectal difference between British and American English in usage of the two filled-pause forms.

Table 22: Comparisons of Distribution of Filled-Pause Forms by Position

Comparison	Degrees of Freedom	χ^2 Coefficient	p
ATIS vs. AMEX	3	56.54	<.0001
AMEX vs. SWBD	3	44.90	<.0001
ATIS vs. SWBD	3	2.50	.48, n.s.

Interestingly, the higher relative rate of "uh" in medial position (and of "um" in initial position) was also found to hold within speakers. Figure 33 shows data from the 30 SWBD speakers. For each speaker, two percentages were computed. One percentage was the number of initial "uh"s divided by total initial FPs. The other percentage was the total medial "uh"s divided by total medial FPs.

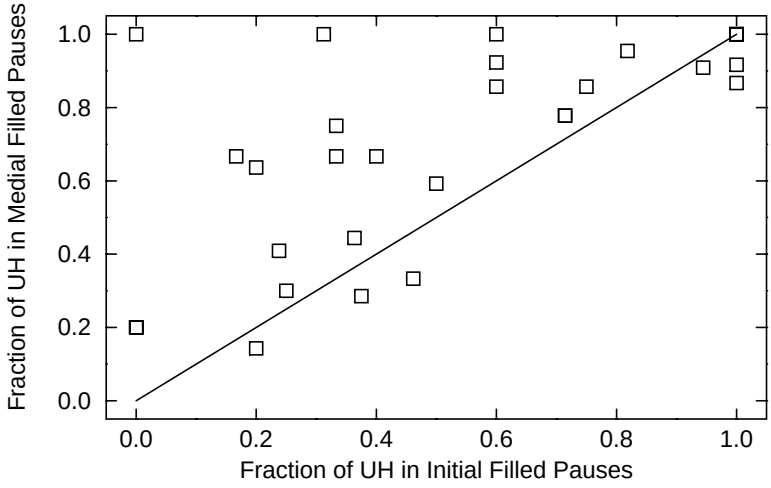


Figure 33. Rate of FP Forms by Position, by Speaker

As shown, most speakers used both forms in both positions, but there is wide variability in relative rates of usage of the two forms in each position. For example, the speaker at the top right corner used "uh" for all FPs--i.e. for 100% of initial FPs and for 100% of medial FPs. The speaker at about (0, 0.2) used "um" for 100% of initial FPs and for 80% of medial FPs. The speaker at (0, 1) used "um" for all initial FPs and "uh" for all medial FPs. Thus, speakers differ both in preference for a FP form, and in degree to which their preference is associated with FP position.

A clear observation from Figure 33, however, is that points tend to fall above the equivalence line. The number of points above the line (23/29; one point falls on the line) is significantly different than that expected under a lack of association between filled-pause form and FP position, $p<.01$ by a Sign test. Since the space above the equivalence line corresponds to an association between "uh" and medial FPs (or "um" and initial FPs), we can infer from these results that the association between FP form and position found over speakers in the previous analysis also obtains for individual speakers in the SWBD corpus.

6.2.5 Section summary

In all corpora, the distribution of DF types is different from uniform. The corpora differ, however, in rates of specific types. Comparisons of type distributions across corpora show roughly equal per-word rates for SUB, ART, and INS, but much higher rates of FP, REP, and DEL in the human-human corpora. The rate of HYB is extremely low in all corpora.

Individual speakers in the SWBD corpus fall into two groups based on relative rates of the frequent types REP and DEL. Some speakers (repeaters) produce a much higher rate of REP than DEL; others (deleters) produce a much higher rate of DEL than REP. Repeaters do not differ from deleters in overall rate of disfluency, nor on a number of sentence-related measures. The groups differ, however, in speech rate: deleters are overall faster speakers than repeaters. Although the rate of FPs by speaker is inversely correlated with the rate of DEL, it is not correlated with a number of sentence-related measures, nor with speech rate. The rate of FPs is, however, correlated with gender: men produce higher rates of FPs than women.

Distributions of DF types by DF position show that in all corpora, the ratio of initial to medial DFs is not uniform over types. Per-word rates by position indicate that the three most common DF types (FP, REP, and DEL) are much more likely to occur in initial position than in medial position. The remaining types appear to be roughly equally likely in initial and medial positions.

The FP forms “um” and “uh” differ in distribution, both across and within speakers. “Um” occurs relatively more often in initial than in medial position. Rates of the two FP forms by position are nearly identical for ATIS and SWBD, but differ for AMEX because AMEX has an overall higher rate of initial DFs.

6.3 Rate of Pattern Features in Basic DFs by Type

This section examines the relationship between DF type and pattern features. The pattern features examined include deletion length, the presence of fragments, and the presence of words in the interregnum. In addition, a study of SUB DFs examines cooccurrences of pattern features, including word fragments, editing phrases, and retracing.

6.3.1 Rate of DFs with *k* deleted words

In Chapter 5 we saw that the frequency of DFs having *k* deleted words fell off exponentially with *k*, and that the same rate of decay was obtained for all three corpora. In this section, deletion length is examined for individual DF types.

Figure 34 shows the frequency (on a log scale) of DFs of particular types by deletion length. Trends are not shown for cases with very low sample size (e.g., CON and HYB, and cases in which there were too few examples of a type in a particular corpus). For ease of comparison, all trends are plotted using the same range of *k* and log(y). Lines are linear fits in the log frequency space; slopes for these fits are indicated next to each trend.

With a few exceptions (e.g. the DEL trend for AMEX and the SUB trends for all corpora) trends are roughly linear in the log frequency space. Therefore the distribution of deletion lengths for each of the types can be modeled separately using the simple function introduced in Chapter 5, where *q* is now a type-specific parameter:

$$log(y) = log(C) + k * log(q)$$

or, in linear space,

$$C * q^{k}$$

As noted in Chapter 5, this is a single-parameter model, since *C* is a deterministic function of *q*.

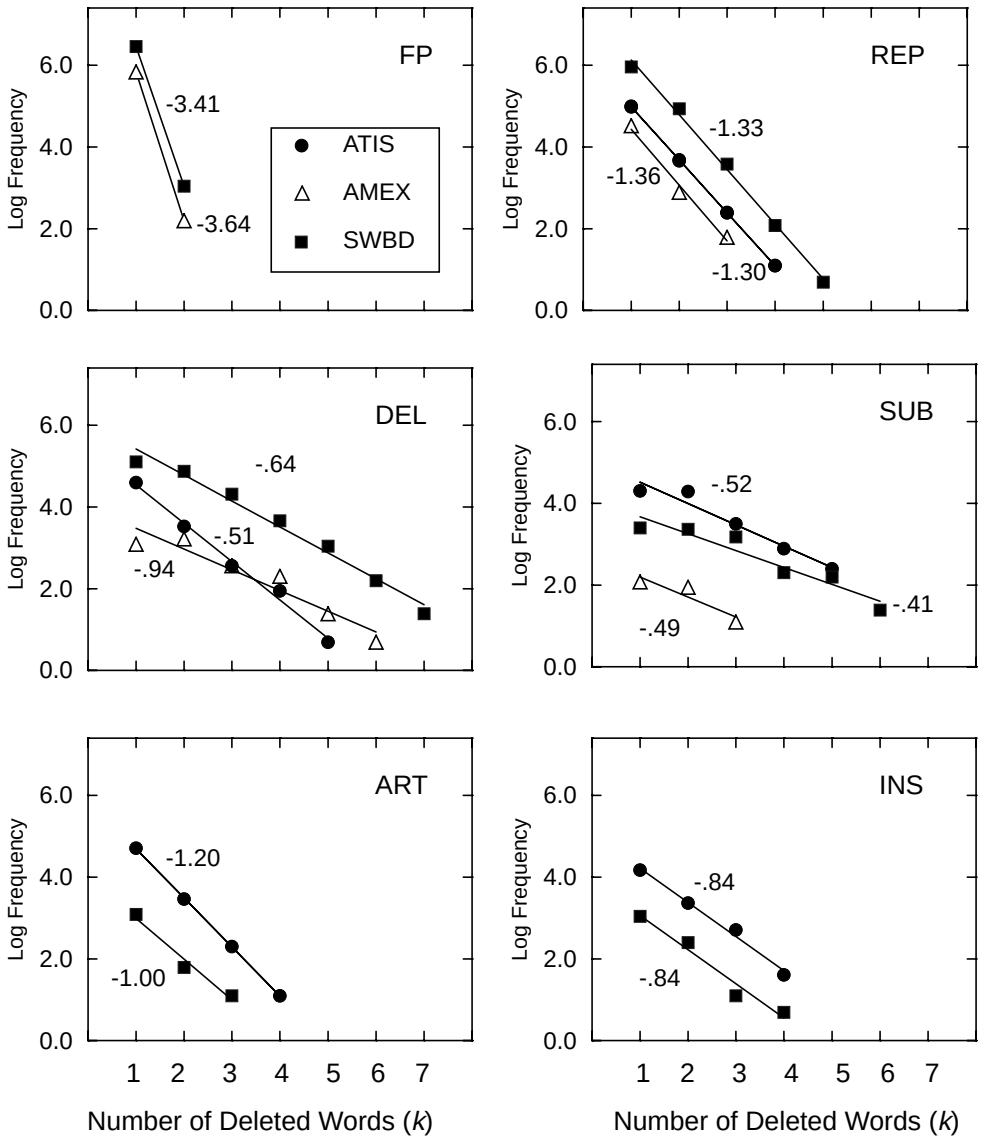


Figure 34. Value of *q* in Deletion Length Model by DF Type

By inspection, the results in Figure 34 suggest that the aggregate trend for deletion length in Chapter 5 reflects a summation of (roughly) exponential trends for each of the different DF types. As pointed out in Chapter 5, the interpretation of the exponential decay rate is unclear, since deletion length represents different phenomena for different types of DFs. Thus in future work,

although there may be different explanations for decay rate for different DF types, it will be important to offer an interpretation for why these rates appear to follow the same basic function.

We turn now to the value of the parameter *q* for each DF type. A notable observation from inspection of the slope values indicated in Figure 34 is that in most cases, values for the same type in different corpora are quite close, while values for different types vary considerably. (An exception, however, is the graph for DEL, which shows three rather different slopes. Given the large amount of total data and data points for type DEL, this is likely to be a robust effect and deserves attention in future work.) Statistical comparison of these slopes is beyond the scope of the present work. However, the results suggest that the domain-invariant-*q* found in Chapter 5 may also be domain-invariant when broken down by type (perhaps with the exception of the type DEL).

For cognitive models, it will be important to explain both why deletion length is different for different types, and why deletion length appears to be generally unrelated to speech domain. In speech applications, this information could be useful in DF correction algorithms by adjusting the probabilities of various deletion lengths based on hypotheses about the DF type. These adjustments are likely to have similar values across speech domains. Results also indicate that the type classification system outlined in Chapter 4 makes meaningful type distinctions that apply generally across speech domains.

6.3.2 Rate of fragments

This section examines the rate of fragments by DF type. Figure 35 shows the probability of a fragment for each DF type. Error bars indicate the standard deviations for the binomial distribution. FPs are not included, since they were never marked as fragments.[8] The remaining missing values correspond to cases of low sample size for the type in the particular corpus (i.e. not to low rates of fragments).

A number of observations can be noted from Figure 35. First, as we saw in Chapter 5, the rate of fragments in ATIS is much higher overall than in either of the other two corpora. Here we see that the high fragment rate in ATIS is not restricted to a particular DF type; rather it holds over the majority of types (including REP, DEL, SUB, and ART). The significance of these

8. It was not possible to determine whether a filled pause was prematurely cut off.

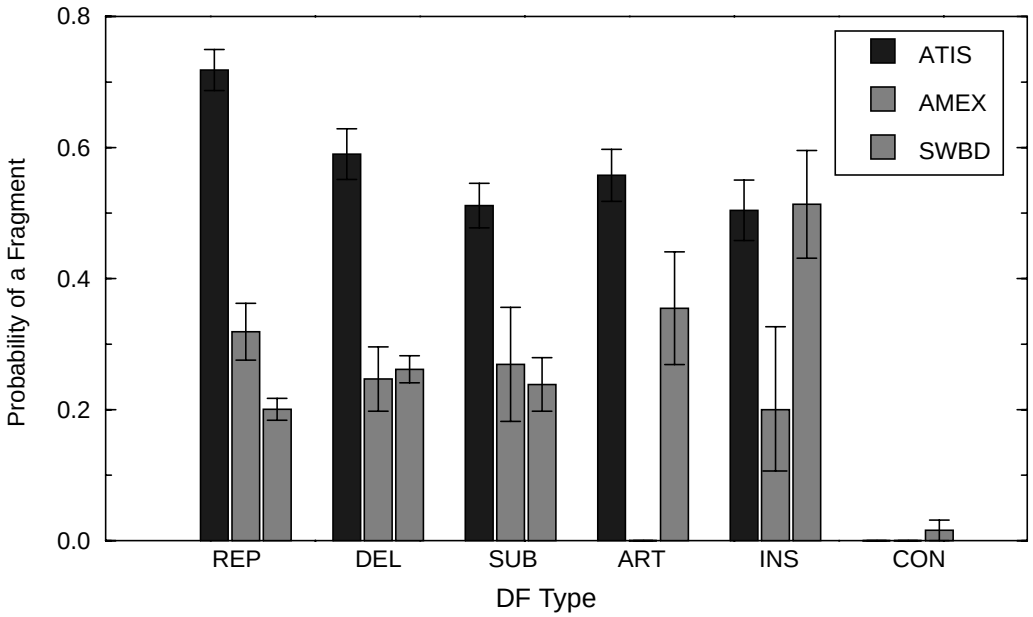


Figure 35. Rate of Fragments by DF Type

comparisons can be inferred from the error bars. It is not clear why ATIS has such a high rate of fragments both overall and by type; this is left as a question for future work.

Second, the patterns of relative rates of fragments are different for the different corpora. In ATIS, for example, the type showing the highest rate of fragments is REP, while in SWBD, REP has the *lowest* rate of fragments. In terms of absolute rates, AMEX has rates similar to ATIS for some types, but not for other types. It is not clear how to interpret these complex results; they may indicate a three-way interaction between presence of a fragment, DF type, and domain.

Third, in SWBD the highest rate of fragments is found for the type INS. This finding is interesting because INS, although not exactly equal to Levelt's appropriateness repair (Levelt, 1983; see Chapter 2), generally involves added information that *does not contradict* previous information. Yet in SWBD these types are interrupted mid-word *most* often, a result inconsistent with previous work showing DFs involving *error* (e.g., ART and SUB) to be more likely to be interrupted mid-word than DFs not involving error (see Chapter 2). This is puzzling and deserves attention in future work.

Fourth, the rate of fragments is quite low in CON DFs. This has an important implication for the Type Classification Algorithm presented in Chapter 4. In the TCA, repetitions of coordinating conjunctions between sentences were separated out from within-sentence repetitions (REPs) in the labeling system, based on suggestions in previous work (see Chapter 2). The difference in fragment rates between CON and REP observed here (1.5% for CON versus 20% for REP) suggests that it is indeed useful to distinguish these two types in the classification system.

Fifth, note that in absolute terms, there is a high rate of fragments for ART DFs in ATIS. This is consistent with past findings (see Chapter 2) that articulation errors tend to be cut off quickly. Of the total fragments in ATIS (N=502), 87 or 17.3% corresponded to words that were misarticulated (e.g. “I'd like to fr- fly from ...."). This point should be taken into consideration when designing automatic approaches to processing fragments. In particular, one possible approach to processing fragments is to allow a recognizer to iterate over the first phones in a word. These results suggest, however, that over 17% of the time in ATIS, the phones in the fragment will not be good renditions of the phones in the onset of the correctly articulated word.

6.3.3 Cooccurrence of pattern features in SUBs

This section investigates pattern features that have been posited in past work to help listeners in processing DFs (see Chapter 2). These include features that mark the presence of a DF, as well as those that could help a listener link up a repair to the original utterance. Three such features labeled in this work include: 1) the cut-off of speech mid-word (the presence of a word fragment at the IP); 2) the presence of an editing term (e.g. “sorry”) in the IM; 3) the retracing of one or more words in the RR. The question in this section is whether these features tend to cooccur in DFs.

The analysis focuses on SUB DFs. This is because: 1) for REP and INS, retracing is always present; 2) for FP, retracing is always absent; 3) for ART, word fragmentation is confounded with misarticulation (as discussed in the previous section); and 4) frequencies of CON and of HYB are too low for analysis.

Data are from the ATIS corpus, since this corpus had the largest number of SUBs (N=217). It should be noted that the SUBs in ATIS are more homogeneous (e.g., replaced city names, dates, and airlines) than those in the other corpora. This was an advantage in the present study, since it reduced variability related to semantic and syntactic variables. The 217 cases came from 144 different speakers, with at most five examples from a particular speaker. Only the sentence-medial SUBs (N=194) were used, because the small percentage of sentence-initial SUBs (N=23) were initial by virtue of having retraced to the beginning of the sentence, thereby confounding the "retrace" feature.

The 194 cases are arranged in Table 23:

Table 23: Three-Way Contingency Table for Presence of Pattern Features in SUB DFs

Retracing:	Fragment	No Fragment
Editing	5	19
No Editing	38	41

No Retracing:	Fragment	No Fragment
Editing	5	8
No Editing	55	23

In order to examine associations among the three categorical features in the table, hierarchical loglinear modeling was used (Goodman, 1970; Fienberg, 1991). In these models, which are stated in terms of sums of effects and interactions of various orders, the existence of an interaction of a certain order implies the nonzero presence of all lower-order interactions and effects involving the same variables. To obtain a parsimonious model of the data, one selectively removes terms from the model, starting with the "saturated" model (one containing all terms). Terms which can be removed without significantly affecting the fit are discarded; terms which affect the fit are retained. The effect of removing a term is evaluated by using the difference in Chi

square values between the original model and the model with the term removed, to test the null hypothesis that the removed term is actually zero.

Table 24 shows the models tested, and their associated Chi square values. The likelihood ratio Chi square is used to evaluate the statistical hypotheses. The Pearson goodness-of-fit Chi square is also shown, because for data with low cell counts these two statistics diverge, indicating problems with the application of the model. As can be seen, for these data the agreement between these two statistics is quite good, indicating the appropriateness of models based on the Chi square statistic for these data.

Table 24: Loglinear Analysis of Associations Among Pattern Features in SUB DFs

Term Removed	Degrees of Freedom	Pearson χ^2	Likelihood Ratio χ^2	Difference	Probability at $\alpha = .05$
None (saturated model)	0	0	0	0	1
Editing x Fragment x Retrace	1	.01	.01	.01	n.s.
Editing x Fragment	2	10.72	10.82	10.81	$p<.05$
Fragment x Retrace	2	9.49	9.54	9.53	$p<.05$
Editing x Retrace	2	.65	.65	.64	n.s.

As shown, the 3rd-order term and the term “Editing x Retrace” can be dropped from the model without significantly affecting the fit of the saturated model. However, the remaining two 2nd-order terms cannot be removed without significantly affecting the fit. Therefore, the most parsimonious model of these data contains the terms “Editing x Fragment” and “Fragment x Retrace.” As can be deduced from the contingency table, both significant interactions are inverse relationships: editing and retracing are both *less* likely after a fragment.

One interpretation of these results is that they reflect trading relationships in the manner in which speakers mark DFs. A speaker who has stopped mid-word may be less likely to use an editing term and less likely to back up when effecting the repair, because the fragment makes the repair salient enough for a listener to detect. An alternative, though not mutually exclusive possibility is that results reflect inverse correlations in features associated with the *time course* of repair for the *speaker.* Retracing and editing could be used as hesitation forms by a speaker not yet

ready with a repair, whereas cutting off mid-word could occur in cases in which the speaker is ready with the repair before completing the error. This possibility could be investigated by examining the temporal characteristics of these DFs.

6.3.4 Section summary

The value of the parameter in the model of deletion length presented in Chapter 5 appears to be similar across domains when examined by DF type. The parameter is similar for the different domains within a DF type, but varies across DF types.

The probability of a fragment is higher in ATIS than in the other corpora, across most DF types. The relationship between corpus, fragment and type appears to be complex. Fragments are unexpectedly high for INS, given that this type of DF typically does not involve correction of error. Nearly 20% of the fragments occurring in REPs in ATIS correspond to onsets of misarticulated words.

Loglinear analysis of the pattern features “retracing”, “fragment” and “editing phrase” in SUB DFs in the ATIS corpus shows significant inverse associations for the second-order terms “editing phrase x fragment” and “fragment x retrace.”

6.4 Rate and Composition of Complex DFs

As described in detail in Chapter 4, complex (or overlapping) DFs are represented by the PLS as complexes of basic DFs in a hierarchical representation.[9] Basic DFs in a complex structure are referred to as “component DFs.” Complex DFs can also be described as cases in which a DF occurs within the repair region of another DF, causing one or more words to play a role in more than one DF. For example, in the following utterance:

```
he -- she -- she went
```

the first “she” plays a role in two DFs: “he--she” and “she--she.”

9. As mentioned in Chapter 4, the hierarchical representation is strictly a formalism; it is not at this point intended as a theoretical model for these cases.

As can be deduced from Table 6 in Chapter 4, about 7-10% of DFs occur overlapping with another DF, across corpora. Since these counts are total component DFs within the complex DFs, and each complex must have at least two members, the total number of complex DF structures is less than half this rate. The analyses in this section use the SWBD data, because of its high number of complex DFs. Although ATIS also has a reasonable number of complex DFs, interestingly, over 30% of the complex DFs in ATIS were degenerate (unlabelable; see Chapter 4). This was mainly due to cases in which the subject stopped speaking and started a new utterance after making two DFs in close proximity.

Section 6.4.1 examines the distribution of *m*-component DFs in these complexes. Section 6.4.2 compares the type distributions of the component DFs to those for basic DFs. Section 6.4.3 asks whether component types combine independently. Section 6.4.4 asks whether complex DFs occur at rates predicted by the occurrence of basic DFs.

6.4.1 Rate of *m*-component complex DFs

This section describes a model for predicting rates of complex DFs having *m* component DFs. Of the 266 component DFs in SWBD, 13 were degenerate. Of the remaining 253, 212 were in 106 2-member complexes, 33 were in 11 3-member complexes and 8 were in 2 4-member complexes. These frequencies lie roughly on a straight line in log frequency space, as shown in Figure 36.

Furthermore, when the appropriate value for basic (noncomplex) DFs[10] is plotted, it also lies roughly on the line, as indicated in the figure. The fact that values for $m>1$ lie on a line suggests that the likelihood of making a DF in the repair region of a previous DF is independent of the number of members in the complex DF. The fact that the point for $m=1$ (or basic DFs) also lies on this line suggests that there is nothing special about a complex DF in terms of its tendency to be interrupted: a basic DF is just as likely to be interrupted within its repair region as a DF that is already complex.

[10.] The appropriate value for basic DFs is the number of basic DFs that could potentially have become complex DFs by being interrupted during the repair phase. As discussed in Chapter 4, DFs having no words in the repair region cannot be the lower member of a complex DF. This includes the types FP and DEL. Therefore, the point plotted for $m=1$ does not include these types.

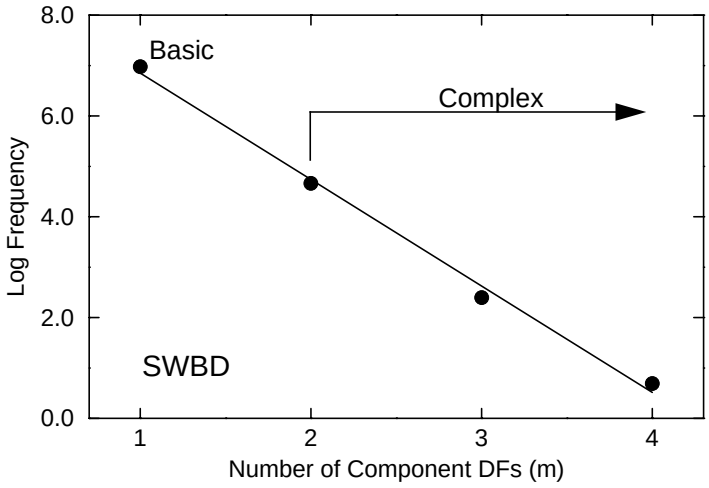


Figure 36. Rate of DFs with *m* Component DFs

The rate of complex DFs having *m* component DFs can thus be modeled by the function:

$$C * v^{m}$$

A linear fit in log-frequency space yields the values:

$$8.965 - 2.113 * m$$

Therefore, *v* is equal to $e^{(-2.113)}$, or about 0.12. This rate corresponds to the probability of making a DF while in the repair region of a previous DF; it is discussed further in Section 6.4.4 below.

6.4.2 Rate of component-DF types

This section asks whether the DF types that occur in complex DFs are drawn from the same distribution as the types that occur in basic (noncomplex) DFs. The question is addressed by

comparing the distribution of types found in the lower member of two-member DFs,[11] and in the upper member of these DFs--to the distribution of basic DFs.

The possible combinations of these lower and upper types are shown in Table 25, with simple examples and pattern representations. Note that not all eight types appear. Types with very low frequency (ART and HYB) play no role in this analysis; CON was for this study grouped with REP, since it would otherwise be too infrequent to include. As explained in Chapter 4, the lower member of a complex DF is never FP or DEL, since these types have no words in the repair region. The upper member is never FP, since FPs have no words in the reparandum. This leaves the types: REP, INS, and SUB, as possibilities for the lower member. with DEL an additional possibility for the upper member.

Table 25: Type Combinations in Two-Member Complex DFs

Upper DF: \ Lower DF:	REP	SUB	INS
REP	he -- he -- he went [R[r.r].r]	he -- she -- she went [R[s.s].r]	a toy -- new toy -- new toy [RR[r.ir].rr]
DEL	he -- he -- it rained [D[r.r].]	he -- she it rained [D[s.s].]	a toy -- a new toy -- it rained [DDD[rr.rir].]
SUB	he -- he -- she went [S[r.r].s]	he -- she -- we went [S[s.s].s]	a toy -- old toy -- new toy [SR[r.ir].sr]
INS	a toy -- new -- new toy [r.I[r.r]r]	a new toy -- old toy -- big old toy [RR[sr.sr].irr]	a toy -- new toy -- big new toy [RR[r.ir].Irr]

In each example, "--" marks the interruption point. In the pattern representations, lowercase symbols match up one-to-one, left to right, with the actual words in the DF; these words

[11.] Use of only the two-member complexes (which account for 80% of the total complex DFs) simplifies analyses.

are underlined in the examples. Uppercase font represents the output of a lower DF; the *letters* in uppercase font indicate *how* the output of the lower DF functions in the upper DF. See Chapter 4 for further description of these conventions.

Table 26 shows the distribution of types that occurred as the lower member of complex DFs, and the distribution of these types that occurred as basic DFs. Cell values are counts, expected values, and residuals, respectively.

Table 26: Comparison of Type Distributions for Lower Member of Complex DFs and Basic DFs

	Lower Member in Complex DF	Basic DF	Row Total
REP	75 (73.4) 1.6	636 (637.6) -1.6	711 81.5%
SUB	10 (12.3) -2.3	109 (106.7) 2.3	119 13.6%
INS	5 (4.3) .7	37 (37.7) -.7	42 4.8%
Column Total	782 89.7%	90 10.3%	872 100.0%

The value of the Chi square coefficient for this table is .63, $p>.05$. The lack of a significant interaction suggests that types occurring as the lower members in a complex DF are drawn from the same distribution as types occurring in basic DFs. Note, however, that the inability to reject the null hypothesis of independence in the table does not constitute direct evidence for independence; rather we can infer only that results are not inconsistent with the hypothesis that types in both conditions are drawn from the same distribution. Since failure to reject the null hypothesis may result from a lack of power, a larger set of data would be particularly helpful for this kind of analysis. The 90 examples used were only minimally sufficient for application of the Chi square test. Note that in order to obtain these 90 two-member non-degenerate complex DFs, it was necessary to have a hand-labeled corpus of over 40,000 words (as well as a high DF rate in the corpus).

A similar test was conducted to compare the distribution of types in upper members of complex DFs to the distribution of types in basic DFs. Since upper members include the type DEL, there are four types for this comparison. Observed counts, expected values, and residuals, respectively are shown in Table 27.

Table 27: Comparison of Type Distributions for Upper Member of Complex DFs and Basic DFs

	Upper Member in Complex DF	Basic DF	Row Total
REP	42 (46.1) -4.1	636 (631.9) 4.1	678 51.2%
DEL	33 (32.9) .1	451 (451.1) -.1	484 36.6%
SUB	12 (8.2) 3.8	109 (112.8) -3.8	121 9.1%
INS	3 (2.7) .3	37 (37.3) -.3	40 3.0%
Column Total	90 6.8%	1233 93.2%	1323 100.0%

The Chi square value for this table is 2.27, $p>.05$. Thus, as in the previous comparison for the lower members, we cannot reject the null hypothesis of independence in the table.

Results in this section are not inconsistent with a hypothesis that both lower and upper members in complex DFs are drawn from the same distribution as types in basic DFs. This empirical finding lends support to the approach developed for representing complex DFs as described in Chapter 4. That is, rather than represent such cases using a flat pattern with multiple interruption points, we can analyze them as complexes of basic DFs.

6.4.3 Compositional independence

As just shown, the DF types in the lower and upper members of complex DFs occur at rates that can be predicted from the distribution of types occurring in basic DFs. This section asks

whether types *combine* in a complex DF *independently*, i.e. whether the cooccurrence of two types in a complex structure is no different than the rate expected if one were to randomly draw a type for each member from the distribution of basic types. Note that while compositional independence is consistent with the outcome of the previous section, it is not the only situation in which the previous outcome could obtain. For example, similar distributions for upper and lower DFs could also result if each type cooccurred in a complex only with another DF of the same type.

Results for the 90 complex DFs are shown in Table 28. Cell values contain counts, expected values, and residuals, respectively. Expected values correspond to values predicted by a model of compositional independence.

Table 28: Type Composition of 2-Member Complex DFs

	REP	SUB	INS	Row Total
REP	34 (35) -1.0	5 (4.7) .3	3 (2.3) .7	42 46.7%
DEL	31 (27.5) 3.5	2 (3.7) -1.7	0 (1.8) -1.8	33 36.7%
SUB	7 (10) -3.0	3 (1.3) 1.7	2 (.7) 1.3	12 13.3%
INS	3 (2.5) .5	0 (.3) -.3	0 (.2) -.2	3 3.3%
Column Total	75 83.3%	10 11.1%	5 5.6%	90 100%

The value of the Chi square statistic for this table is 9.42, p>.05, indicating no basis for rejection of compositional independence. However, because of the low expected values in many of the cells, this is not a reliable result. To address the problem of low cell values, cells with expected values below 4.5 were combined into a single cell, and the Chi square statistic was recomputed for the resulting one-way table. Again, no significant interaction was found, χ^2=9.26, p>.05. Other

methods of regrouping the low-count cells yielded the same (nonsignificant) result. We can infer that there is no basis upon which to reject the null hypothesis that types combine independently in complex DFs. However, as previously noted, lack of significance does not constitute evidence for independence.

These results lend further support to the compositional analysis of complex DFs, since there seems to be nothing special about the distributions of member DFs, nor about the ways in which they combine. In addition, this particular result is consistent with the decision *not* to allow cases such as "he he he" to be analyzed as N-ary branching, but rather to analyze them as binary-branching along with the rest of the complex DFs (see Chapter 4). This is because like all other combinations of types, REP appears to combine with REP at rates predicted by the rate of REP in basic DFs.[12] The fact that cases like "he he he" occur relatively frequently should not be taken as evidence for special treatment in a representation system, since the high rate of the complex is predicted by the high rate of REP in basic DFs.

6.4.4 "Synergy" effect

The previous three sections have shown ways in which complex DFs appear to be similar to basic DFs. Section 6.4.1 found that the probability of interrupting within the repair region of a previous DF did not depend on whether the previous DF was basic or complex. Section 6.4.2 found that the distribution of DF types occurring in the lower and the upper members of complex DFs can be predicted from the distribution of types in basic DFs. Section 6.4.3 found that the types in lower and upper members combine independently. The question raised in this section is whether the probability of producing a DF while in the repair region of a previous DF is the same as the probability of producing a DF elsewhere. The three earlier findings are consistent with a hypothesis that these rates are the same. However, the earlier findings could also obtain under a difference in rates; that is, one rate of disfluency for words in the repair regions of previous DFs, and another rate elsewhere.

In Section 6.4.1 we saw that the probability of making a DF within the repair region of a previous DF is about .12. The per-word rate of disfluency for basic DFs in SWBD found in

---

12. Further analyses of complex DFs composed of REPs are discussed in Shriberg (in preparation (b)).

Chapter 5 was about .06; after adjusting for the removal of FPs (which as noted earlier cannot occur as a member of a complex DF) the rate is about .04. The question is whether the rate for complex DFs, .12, is significantly different from the rate expected if the adjusted overall per-word rate, .04, is applied at each word in the repair region of DFs.

The estimated number of complex DFs under the null hypothesis (no difference in per-word DF rate in repair regions versus elsewhere) is most easily computed by estimating the total number of DFs that could potentially have become part of a complex DF, and subtracting from this value the number of DFs expected *not* to be interrupted during their repair phase. The first value is the total number of DFs that could have potentially become part of a complex DF (by being interrupted during their repair phase), including those which actually did become part of a complex DF. This value is equal to: the number of basic DFs, minus any FPs and DELs (which cannot become part of a complex DF; see above), plus a count for each complex DF (the entire complex, not the component members). For the SWBD corpus, this value came out to 1194.

The second value estimates the number of potentially interruptible DFs that should *not* be interrupted, based on the adjusted per-word *fluency* rate for basic DFs. This is a summation, over each repair-phase length, of the number of observations at a particular length times the per-word fluency rate raised to that length. Since the adjusted per-word fluency rate for SWBD is close to one (.9597), most DFs will "pass through" without a DF in their repair region; however, some small number of DFs will have an interruption on one or more of the words in the repair region. This estimate came out to 1112.88.

The estimated total number of complex DFs under the null hypothesis is thus 1194-1112.88, or 81.12. In actuality, 133 complex DFs occurred. To test whether these values differ significantly, the 95% confidence limits were computed for the total interrupted words in the repair regions of the 1194 DFs. Results showed a range from 64.68 to 99.49 complex DFs. This is the range of total interruptions; note that the total complex DFs can be lower than the lower bound of the range, because DFs can double up within repair regions. We are interested only in the upper end of the range. The 133 observed complex DFs require at least 133 total complex DFs. Since 133 is beyond the upper bound of the 95% confidence limits, we can infer that significantly more complex DFs occurred than were predicted by the rate of disfluency for basic DFs.

This suggests a "synergy" effect: a DF is more likely to occur during the repair phase of a previous DF than elsewhere. Note that the synergy effect is similar to the "cooccurrence" effect discussed in Chapter 5; however, the two effects are definitionally distinct. The synergy effect is defined for DFs in the repair region of another DF. The cooccurrence effect was analyzed for sentences containing no complex DFs, and required DFs only to cooccur in the same sentence. It is not clear whether or not these effects reflect similar underlying factors. As noted in discussing the cooccurrence effect, an interesting area for future work is determining whether the effect reflects causation or correlation. The causation hypothesis attributes a later DF to disruption caused by making the earlier DF. The correlation hypothesis attributes both DFs to some latent variable (for example, difficult semantic content) in the sentence overall. These alternative hypotheses also apply to the synergy effect, and are important to investigate in future work.

6.4.5 Section summary

The probability of making a DF while in the repair phase of a previous DF does not depend on whether the previous DF is basic or complex, nor on how many members make up the previous DF if it is complex. The component DFs in the lower and the upper members of two-member complex DFs appear to be drawn from the same distribution of DF types as basic DFs. The lower and upper DF types appear to combine independently. The probability of making a DF while in the repair region of a previous DF is higher than predicted by the rate of basic DFs overall ("synergy effect").

6.5 Acoustic Properties of Simplest Types

This section examines acoustic properties of the two simplest DF types: FP and REP. These types are simple in the sense that they do not involve changes in the syntax or semantics of an utterance. They are also the two most prevalent types in all corpora (see Figure 22). These factors make FP and REP good starting points for pretheoretical acoustic studies.

Section 6.5.1 examines the duration of phones in FPs, comparing them to the same phones in other contexts, as well as noting the manner in which they scale with increased duration. Sections 6.5.2-6.5.4 compare the duration and F0 characteristics of words in REPs to those of filled pauses.

6.5.1 Duration of phones in FPs

This section investigates the duration of filled-pause vowels in the ATIS corpus. In English, the vowel in the FPs “um” and “uh” is typically close to schwa; however it can also carry stress ([^]), or occur further back and lower in the vowel space ([*a*]). In automatic speech recognition, FPs are sometimes misrecognized as “a,” or as parts of other words containing the relevant vowels (Butzberger, Murveit, Shriberg, & Price, 1992; see Chapter 2). Since FPs have been observed to show quite long durations (e.g., O'Shaughnessy, 1992; Shriberg & Lickley, 1993), a question is how helpful duration could be in distinguishing the vowels in FPs from the same vowels elsewhere.

Durations for the vocalic portion of 700 FPs and for 40,000 instances of the filled-pause vowels elsewhere, including in the determiner “a,” were obtained automatically using forced alignment recognition (as explained in Chapter 4).

Figure 37 shows normalized distributions for: 1) the filled-pause vowels; 2) vowels in the word “a”; 3) schwa and [*a*] in any words; and 4) [^] in any words. As shown, vowels in FPs have much longer durations than the same vowels in other contexts. Duration, then is a simple cue that could be used by speech recognition systems in discriminating vowels in FPs from the same vowels elsewhere. This information also has implications for duration modeling. FPs should probably be omitted in duration modeling in order not to skew the distributions for filled-pause vowels toward higher values.

Interestingly, a second observation suggests that the durations of phones in FPs do not scale like those of other words. Figure 38 shows the duration of the nasal ([m]) in 238 tokens of “um,” plotted against the duration of the vowel in the same token. As shown, there is considerable variability in the durations of each of these phones. However, there is no significant correlation between the duration of the vowel and that of the nasal within a token, r^2=.10, p>.05. In some

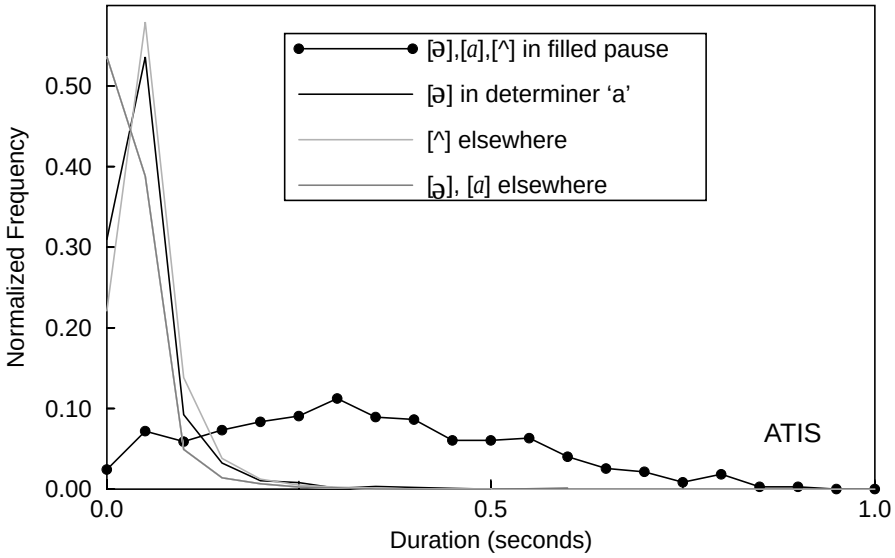


Figure 37. Duration of Vowels in FPs and of Same Vowels Elsewhere

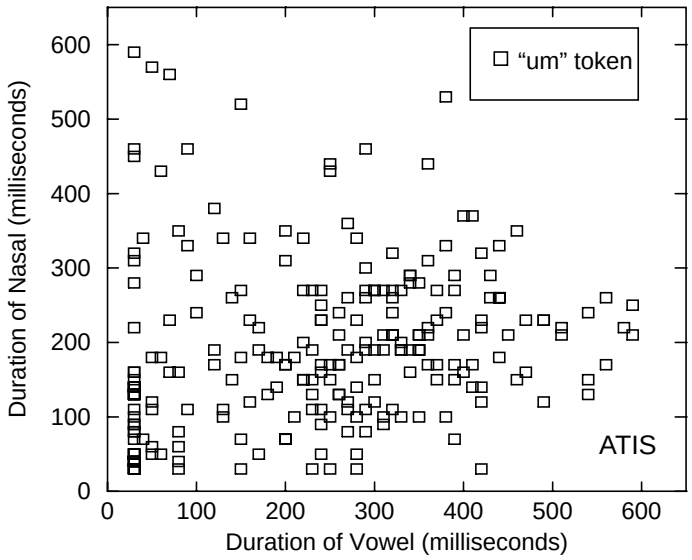


Figure 38. Duration of Vowel and Nasal in "um"

tokens the vowel is sustained; in others the nasal is sustained; in still others both phones are sustained.

A question for future work is whether the data in Figure 38 reflect variation in the relative scaling of the two phones within speakers, or idiosyncratic tendencies of speakers. This question could not be investigated in the present study since the 238 examples represented 70 different speakers.

6.5.2 Duration of words in REPs

This section examines the duration of words in single-word REPs, for example:

```
Ford is   is  focusing on quality.
     R1   R2
```

In this and the following sections, evidence is presented which suggests that it is not the repeated instance (R2), but rather the first instance (R1) of the repeated word that should be viewed as the disfluent member in the REP. The evidence is in the form of duration and F0 characteristics of R1 that show it to be similar to filled pauses such as “uh” and “um” (as described by Shriberg & Lickley, 1993).

Waveform analysis was conducted for multiple REPs from a single speaker in the SWBD corpus (male, age 34) who participated in a large number of conversations. A single speaker was used because the analyses required highly controlled examples. For purposes of comparison to previous work on filled pauses, the selected cases were unaccented, single-word REPs with no intervening words (i.e. with no word in the IM). In addition, the REPs chosen occurred between two pitch-accented syllables within the same clause. Cases in which R2 was followed by another DF (including a silent hesitation pause) were excluded. The durations of R1, R2, and any intervening pause were hand-marked with the aid of spectrograms.

The first question was whether there was a consistent difference in duration between R1 and R2. As noted in Chapter 2, O'Shaughnessy (1994) studied REPs in the ATIS corpus and reported that R1 and R2 were either of similar duration, or, if there was a pause between the repeated words, R2 showed shortening when R1 showed prepausal lengthening.

The analysis of duration used 66 examples, the majority of which occurred as determiners, prepositions, verbs, auxiliaries, and possessives.[13] For example ('*' precedes the vowel in an accented syllable):

```
*only the the g*irls

w*orld of of electr*onics

t*axes are are k*ind of

c*oncept is is c*arried

m*en have have m*anaged

b*oth your your sec*urity
```

Figure 39 shows the duration of R2 plotted against that of R1, as well as the equivalence line y=x. A Wilcoxon signed ranks test for large samples[14] showed that R1 was reliably longer than R2, $z=7.09$, $p<.0001$. This is consistent with O'Shaughnessy's finding for REPs in the ATIS corpus (O'Shaughnessy, 1994).

We cannot deduce from this result, however, whether R1 is lengthened or R2 shortened, or whether both effects occur. The lengthening of R1 could be interpreted as hesitation. The shortening of R2 could be interpreted as a reduction effect similar to that observed for previously mentioned words in discourse (e.g. Fowler & Housum, 1987). To investigate these alternatives, the durations of R1 and R2 were compared with durations for the same word produced in fluent contexts by the same speaker.

---

13. The majority of REPs involved function words, consistent with results in other studies (e.g., Maclay & Osgood, 1959; Lickley, 1994). However, Shriberg (in preparation (b)) offers an alternative account of repetition frequency that predicts this outcome without referring to the content-word/function-word distinction.

14. Since there is great variability in the duration of R1 (see Figure 39), a nonparametric test is used, because a few very large values of R1 could skew results in favor of the hypothesis that R1 is longer than R2.

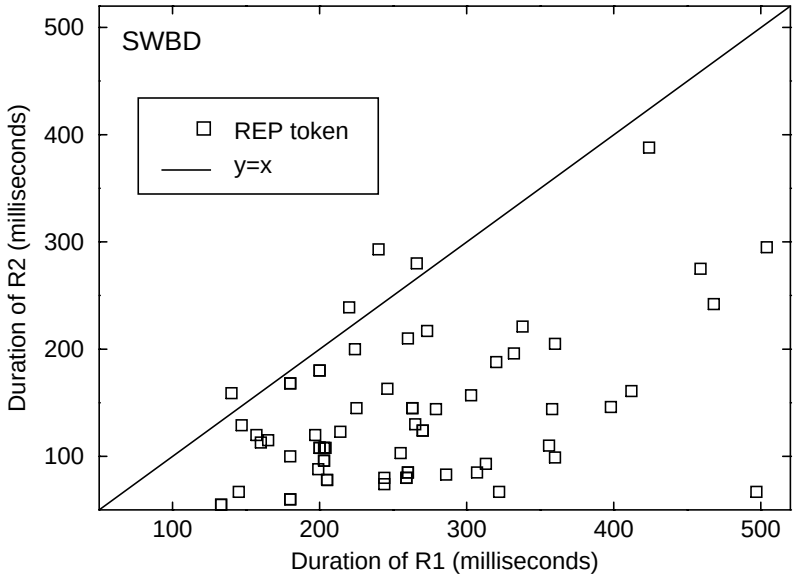


Figure 39. Duration of R1 and R2 in Single-Word REPs

To minimize variability, the analysis was restricted to the study of a single word. The word chosen was “the” because: 1) it is typically unaccented; 2) it occurs frequently and is repeated frequently; and 3) it is often found in clause-internal position, unlike other very frequently repeated words such as “I.”

The speaker described above participated in 33 conversations (about 29,000 total words). Within this set of conversations, 17 instances of “the the” were found that met the requirements described above (note that this is an advantage of using large data sets). For each case, the duration of R1 and of R2 was hand-measured. To serve as comparison tokens, the durations of 40 instances of “the” in unrepeated contexts were also recorded. The unrepeated instances were also restricted to occur clause-internally and to show no following DFs. Instances in which “the” was pronounced with a tense vowel (“thee”)[15] were excluded because of the confound of the change in vowel quality with duration.

---

15. The form “thee” has been associated with upcoming trouble (see Fox Tree & Clark, 1994). The tense form appeared to occur more often for R1 then for R2 in repetitions (ignoring cases of “thee” attributable to a following vowel-initial word.) This fits with the notion of R1 as a hesitation form, as suggested in this and the following sections.

Figure 40 shows the mean duration and standard deviations for R1 and R2 in the 17 repeated instances, along with the durations for the 40 unrepeated instances.

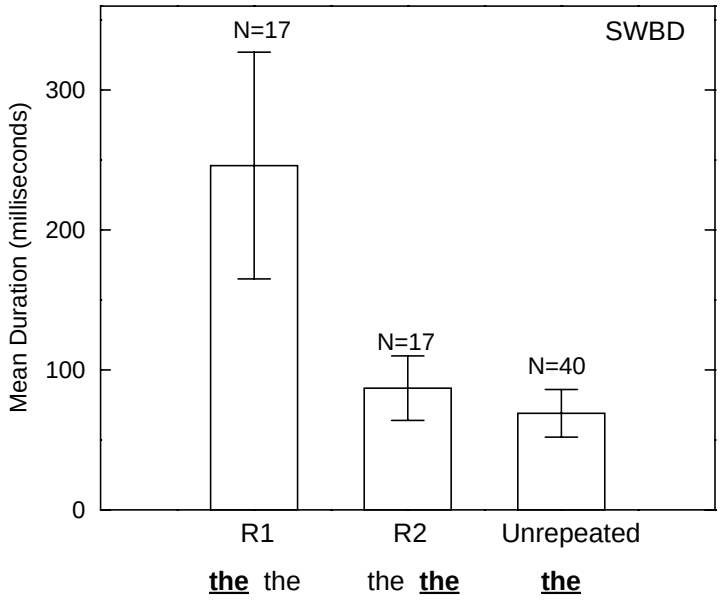


Figure 40. Duration of R1, R2, and Unrepeated Tokens of "the"

Despite the small sample size for repeated tokens, results show that the duration of the unrepeated instances is much closer to the duration of R2 than to R1. As can be inferred from the error bars, R1 tokens are significantly longer than unrepeated tokens; this result is consistent with the hypothesis that R1 is lengthened, as discussed above.

Also, importantly, the mean duration of R2 tokens is *greater* than that of unrepeated tokens. Although it may not be obvious from Figure 40, which is scaled for the much larger R1 values, the difference is significant. The small-sample 95% confidence interval for the difference in mean duration between R2 tokens and unrepeated tokens is 5.33-25.71 milliseconds. Therefore, we can say with 95% confidence that the mean duration of R2 (in the population) exceeds that of unrepeated instances (in the population) by a value within this range. This result is inconsistent with the shortening hypothesis, since R2 is slightly but reliably *longer* than unrepeated tokens. The increased duration of R2 tokens over unrepeated tokens may reflect an effect of restarting after

having slowed down in the hesitation. However, it would be helpful to reevaluate this small effect with a larger data set.

Thus, results show that the difference in duration between R1 and R2 observed in the previous section is attributable to a considerable lengthening of R1, and that there is no shortening of R2. This suggests that speakers may hesitate at R1, and resume fluent tempo at R2. An additional finding in favor of this hypothesis is visible in the previously shown Figure 39, where R1 was plotted against R2 for the 66-word set of REPs. This set contained different word-pairs (e.g. “the the,” “of of”). One would therefore expect some correlation between the durations of R1 and R2 for the same tokens. However, there was no significant correlation between R1 and R2 (r^2=.45, p>.05). The lack of correlation is consistent with a hypothesis that the duration of R1 is determined by the extent to which the speaker needs to pause. In future work, it will be important to determine whether these results hold for additional speakers, as well as for different grammatical and prosodic contexts.

6.5.3 F0 relationships in FPs

This section and the following section examine the F0 patterns in FPs and REPs, respectively. The present section briefly describes a model of filled-pause intonation proposed in previous work. The following section applies the method used in the filled-pause study to the analysis of F0 patterns in REPs.

As noted in Chapter 2, previous work on FPs in the ATIS corpus found that FPs show a gradual, roughly linear F0 fall (O'Shaughnessy, 1992; Shriberg, 1991), and tend to be lower in F0 than surrounding words (Shriberg, 1991). In addition, the F0 of FPs occurring within a clause was found to be related to the F0 of the surrounding speech. Shriberg and Lickley (1993) analyzed clause-internal filled pauses from two speech corpora: ATIS, and a corpus of British English conversational speech (described in Lickley, 1994). The authors found that F0 values at the beginning and end of the filled pause, and at the closest preceding and following F0 peaks within the same clause (peaks typically corresponded to accented syllables) showed regular relationships.

Figure 41 shows data for a female speaker (from the set of American English speakers in the ATIS data). Lines connect points for a specific filled pause. The four F0 measurements are plotted at equally-spaced intervals; therefore the actual temporal intervals between these points (which varied greatly) are not represented in the figure. The solid heavy line indicates the speaker's estimated “baseline” F0. This value was intended to reflect the lowest F0 the speaker could produce (excluding F0 values in regions of vocal fry). Baseline F0 was estimated by measuring F0 at the end of sentence-final F0 falls.

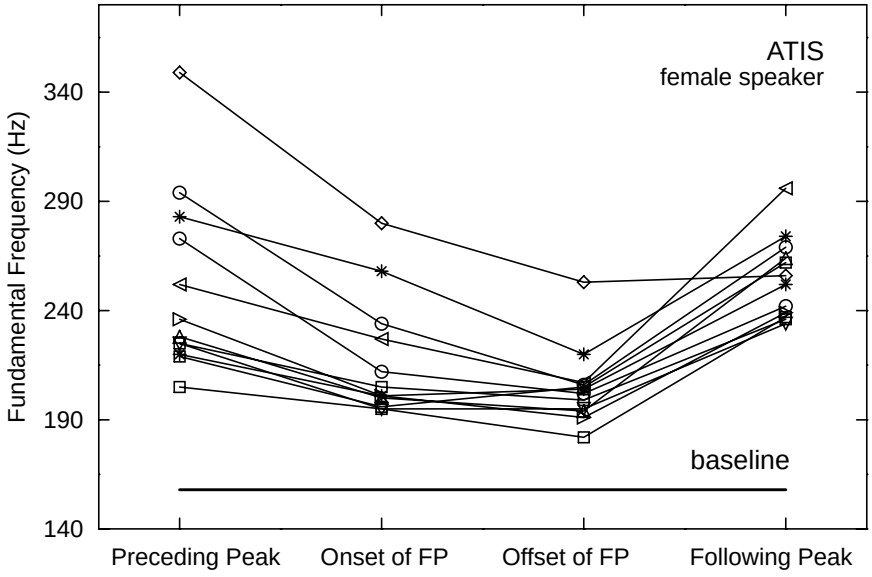


Figure 41. F0 of Clause-Internal FPs and Surrounding Peaks. (Lines connect points for an individual FP token.) From “Intonation of Clause-Internal Filled Pauses” by E.E. Shriberg and R.J. Lickley, 1993, *Phonetica*, *50*(3), p. 174. Copyright 1993 by S. Karger.

The authors proposed a model to predict the starting F0 of the filled pause from the preceding peak F0. (Note that there appear to be many additional relationships that could be modeled). In modeling the relationship, two observations were considered. First, the F0 of filled pauses did not fall below the F0 of the speaker's estimated baseline; this suggests that filled-pause F0 should not be modeled as a subtractive function of peak F0. Second, there appeared to be a compressive effect for points closer to the baseline, i.e. a shallower drop from the peak to the filled

pause for peaks closer to the baseline. This suggests that filled-pause F0 should not be modeled as a multiplicative function of peak F0.

Considering these observations, an additive-multiplicative model was proposed to predict the onset F0 of the FP from the preceding peak F0:

$$F0_{\text{filled-pause}} = r\ (F0_{\text{peak}} - F0_{\text{baseline}}) + F0_{\text{baseline}}$$

This is a single-parameter model, since the coefficients of peak and baseline are both determined by *r*. In addition, the constant term in the model is not an arbitrary value, but rather corresponds to a meaningful value that can be estimated empirically (the speaker's baseline F0).

The starting F0 of the FP was thus modeled as a function of the previous peak (representing the local pitch range), the speaker's baseline F0 (representing the speaker) and the scaling parameter *r*, which expresses the proportion of the distance between peak and baseline at which the FP begins. Over speakers, *r* came out to .62; therefore FPs begin at an F0 value slightly above a value that is halfway between the preceding peak F0 and the baseline F0.[16] This single-parameter model provided a better fit than alternative models having an additional free parameter.

In addition, results indicated that the intonation of FPs may be independent of temporal variables. As mentioned above, across examples there was great variation in time intervals between the four measured F0 values. Yet, these values showed regular relationships when points were plotted at regular intervals as in Figure 41. Focused analyses showed no significant correlation between the time from the preceding peak to the FP and the F0 fall in this region. Similarly, there was no correlation between the duration of the FP and the total fall in F0 over the course of the FP. These findings suggest that speakers may preserve intonational relationships under changes in duration necessitated by the need to pause.

---

16. It is possible that FPs start closer to halfway between the peak and baseline, because the reference F0 used to scale pitch over the course of an utterance may be higher than the F0 observed after final lowering (e.g., Pierrehumbert, 1980). In fact, prediction error was minimized by using a value for baseline F0 that was roughly 10% higher than that estimated empirically.

6.5.4 Fundamental frequency relationships in REPs

In this section, the methodology used in the Shriberg and Lickley study is applied to the analysis of F0 relationships in the set of clause-internal, single-word REPs introduced earlier. Recall that this set contained different repeated words, from a single male speaker. Of the 66 REPs, 24 were chosen for analysis. These were cases having high-quality pitch tracks (see Chapter 4) and lacking segmental effects that would obscure the measurement of F0 in the relevant locations.

F0 was hand-measured at six points: at the beginning and end of R1 and R2, and at the maximum F0 point in the closest preceding and following F0 peaks. As in the filled-pause study, the speaker's baseline F0 was estimated from measurements of F0 after final lowering; the estimated baseline F0 for this speaker was 70 Hz. The six points for each of the 24 examples are plotted at regularly-spaced intervals in Figure 42. The estimated baseline is also indicated.

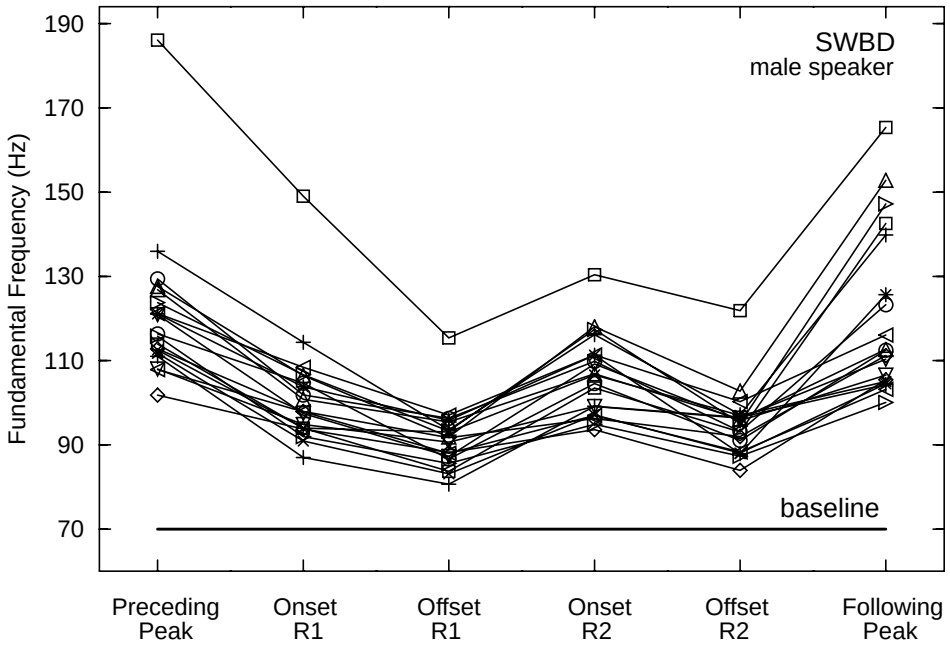


Figure 42. F0 of Clause-Internal REPs and Surrounding Peaks. (Lines connect points for an individual REP token.)

As shown, the relationships among these six F0 points are quite regular. This is particularly striking given the wide variation in temporal relationships among the points at which F0 was measured. The temporal variation included variation in the time between the peaks and the repeated words, variation in the duration of R1 and R2 (as seen in Figure 39), and differences in the length of the unfilled pause (if any) between the repeated words.

In addition, although it cannot be seen from Figure 42, R1 tended to show an F0 trajectory much like that observed for most filled pauses, i.e. a gradual, roughly linear fall. This combination of increased duration and slowly falling F0 thus seems to characterize hesitation syllables more generally; it is seen for filled pauses, R1 in REPs, as well as for syllables preceding hesitation pauses in an otherwise fluent region (e.g., Guaïtella, 1993 [for French]; Ostendorf, Price, & Shattuck-Hufnagel, 1994).

A curious, unexpected finding visible in Figure 42 is that R2 showed an F0 fall, much like R1. Because R2 was always an unaccented syllable, it was expected that the F0 of R2 would interpolate to the following F0 peak, which corresponded to an accented syllable as explained earlier. In future work, examination of additional speakers and inspection of the F0 of unrepeated instances could help discern whether the fall on R2 represents a departure from the F0 pattern expected for fluent regions.

Also observable from Figure 42 is that there is a slight but consistent increase in the F0 of R2 relative to the F0 of R1. The magnitude of the increase is too small (indeed, it was imperceptible from listening alone) to be interpreted as a functional resetting of the pitch range. However, it may be too large to be attributed to a physiological effect such as an increase in subglottal pressure. Preliminary inspection of additional data indicated that this small increase in F0 in the repair region of DFs also appeared for other speakers, as well as for other DF types. This finding may be related to an effect reported by Nakatani and Hirschberg (1994) for DFs in the ATIS corpus, and is worthy of attention in future work.

Two of the relationships shown in Figure 42 were modeled. These included: 1) the relationship between the preceding peak F0 and the onset F0 of R1; and 2) the relationship between the onset F0 of R1 and the offset F0 of R1. For simplicity, the F0 of the temporally earlier event (the predictor F0) is referred to as “A,” and the F0 of the temporally later event (the predicted F0) is referred to as “y.”

In each case, prediction error was computed for four different single-parameter models, as shown in Table 29. The additive-multiplicative model is the same as that used in the scaling of filled pauses (see the previous section); the value 70 in the equation corresponds to the speaker's empirically estimated baseline F0. The remaining entries in Table 29 correspond to models in which y is predicted only from A, using three different functions (additive, multiplicative, and logarithmic). The logarithmic model is equivalent in prediction error to a traditional model used in F0 scaling, the semitone model (e.g., ’t Hart, Collier, & Cohen, 1990).

Table 29: Standard Deviations for Models of F0 Relationships in REPs

Model	Additive-Multiplicative $y = r * (A - 70) + 70$	Additive $y = r + A$	Multiplicative $y = r * A$	Logarithmic $y = r * \log(A)$
Peak to Onset R1	4.95	7.27	5.43	10.68
Onset R1 to Offset R1	4.68	7.49	6.49	6.10

As shown, both relationships are best fit by the additive-multiplicative model. This is a point of comparison between the scaling of filled pauses and the scaling of R1 in REPs. In addition, the fit for the first relationship, which corresponds to the relationship fit in the filled-pause study (i.e. the drop from the preceding peak to the beginning of the hesitation form) showed an *r* value of .6247. This is the same value as obtained for *r* across speakers in the filled-pause study (.62), although it is likely to be coincidental that these results are so similar.

Thus, there are many similarities between FPs and R1 in clause-internal unaccented REPs, including: 1) significant duration lengthening; 2) F0 shape; 3) invariant F0 relationships under variations in temporal characteristics; 4) form of the best-fitting F0 scaling model; and 5) value of the scaling parameter in the F0 model. These similarities have methodological and theoretical implications. While in some analyses (e.g., the analysis of deletion length in Chapter 5) FPs were seen to group separately from all other types of DFs, in terms of acoustic characteristics these forms appear to be related. Thus, there are alternative meaningful ways to group DF classes. It is

appropriate, then, that the system developed for DF classification avoided predetermined hierarchical grouping of DF types (see Chapters 2 and 4).

The acoustic characteristics of FPs and R1 in REPs have consequences for the automatic recognition of hesitation syllables. Results suggest that the detection of hesitations could be aided by searching for gradually falling F0 patterns in regions in which there is little spectral change over a rather large time window (on the order of over 300 milliseconds). Additional cues could be provided by F0 relationships between the starting point of the hesitation and the closest preceding F0 peak.

Cues to hesitation syllables could be used in automatic systems in a number of ways. First, cues to the presence of a hesitation form could help prevent filled pauses from being misrecognized as other words. Second, the cues could aid in discriminating disfluent repetitions from *fluent* repetitions. This is particularly important for syntactically and semantically ambiguous cases (e.g., “flight one one five nine” could correspond to “flight 159” or “flight 1159.”) Third, F0 cues could aid systems using prosodic modeling in discriminating hesitation lengthenings from lengthenings associated with accentuation or prosodic phrasing. This is important because the different types of lengthened syllables carry different information; for example, accented syllables typically have high semantic value whereas hesitations typically occur on less important words. And fourth, recognition of hesitation syllables could help prevent the premature cutoff of speakers for applications using automatic endpointing techniques. The hesitation syllable could serve as a warning that a following pause should not be interpreted as the end of the speaker’s turn; the endpointing threshold could be increased in such locations.

6.5.5 Section summary

Vowels in FPs in the ATIS corpus have much longer durations than the same vowels elsewhere in that corpus. There appears to be no relationship between the duration of the vowel and that of the nasal in “um”; for some tokens the vowel is lengthened, for others the nasal is lengthened, and for still others both are lengthened. In single-word REPs, the first instance (R1) is longer in duration than the second instance (R2). This appears to reflect a lengthening of R1, since R2 is still longer than the duration expected for the same word in fluent contexts. The durations of

R1 and R2 are not correlated. F0 relationships involving R1 in REPs are better described by an additive-multiplicative model than by a linear, multiplicative, or logarithmic model.

6.6 Chapter Summary and Discussion

6.6.1 Summary

The four analysis sections showed regular trends in DF types along a number of different dimensions. The rate of occurrence of DFs by type showed the three corpora to differ in absolute rates of types, but to be similar in: 1) degree of departure from a uniform distribution of types; 2) association between type and position; and 3) association between filled-pause form and position. Certain DF types (SUB, ART, INS) showed similar per-word rates across domains; other types (FP, REP, DEL) showed much higher per-word rates in the human-human dialogs. Examination of type distributions for individual speakers in the SWBD corpus suggested that there are two types of speakers: repeaters and deleters. The deleters produced more words per second than the repeaters, but the groups did not differ in number of DFs per word. The rate of FPs for these speakers did not correlate with the repeater/deleter distinction nor with speech rate, but did correlate with gender: men used significantly more FPs than women.

Analyses of pattern features by DF type showed that the value of the parameter in the model of deletion length in Chapter 5 differs for different DF types, but is similar across domains for all DF types except DEL. Analysis of fragments by type showed that the high rate of fragments across types observed for ATIS in Chapter 5 reflects a high rate for each of the DF types. Loglinear analysis of the cooccurrence of fragments, editing phrases and retracing in SUB DFs in the SWBD corpus showed significant inverse associations for the second-order terms “editing phrase by fragment” and “fragment by retrace.”

Examination of complex DFs in the SWBD corpus indicated that the probability of producing a DF while in the repair phase of a previous DF is independent of the number of components in the previous DF. Types in the upper and lower members of two-component complex DFs appeared to be drawn from the same distribution as basic DFs, and to combine independently. The probability of a DF was found to be greater during the repair phase of a previous DF than elsewhere.

Acoustic analyses of FPs in the ATIS corpus showed vowels in FPs to be longer than the same vowels elsewhere, and showed no correlation between the duration of the vowel and the nasal in “um.” Analyses of REPs from a single speaker in the SWBD corpus suggested that the first instance of a repeated word (R1) bears similarity to a FP in terms of duration, F0 shape, F0 invariance under temporal variation, and F0 scaling characteristics.

6.6.2 Discussion

Comparison of results across the major chapter sections, as well as across the sections in Chapter 5, suggest a number of themes for future work. These include: 1) interpretation of domain differences; 2) interpretation of domain universals; 3) interpretation of the preponderance of independence models; 4) evaluation of the appropriateness of the type classification system; 5) flexibility in the grouping of classes; and 6) use of a large speech database.

6.6.2.1 Domain differences

As in Chapter 5, ATIS was found to differ from the other two corpora in a number of analyses. However, a notable difference between this chapter and Chapter 5 is that in Chapter 5, AMEX and SWBD showed nearly identical trends on many measures, whereas in this chapter a number of differences were seen between these corpora. For example, AMEX and SWBD differed in the distribution of types, the distribution of types by position, the distribution of filled-pause forms by position, and the rate of fragments by type. Thus, these corpora appear to be similar at a very general level, but to differ upon closer examination. This result suggests that both levels of analysis--across DF type, and by DF type--are important in describing the relationship between DF production and speech domain.

6.6.2.2 Domain universals

Despite the preponderance of domain differences observed when DFs were broken down by type, selected analyses revealed potential domain universals. One such case was the finding of similar per-word rates for the DF types SUB, ART, and INS across corpora. A second case was seen in the analysis of deletion length by type, where the parameter q in the deletion length model was found to vary across DF types, but show similar values for particular types across domains.

6.6.2.3 Independence models

A notable observation across this chapter as well as across results in Chapter 5 is the form of the proposed models: nearly all of the models are independence models. This result speaks to the homogeneity of DFs--quite a surprising result given the common view of DFs as irregular events. Thus, while traditional linguistic factors such as syntax and semantics will undoubtedly be crucial in further characterizing and understanding DF production, it is also important to recognize that regularities can be found using simple features and extremely parsimonious models.

6.6.2.4 Type classifications

As discussed in Chapter 4, the Type Classification Algorithm (TCA) represents only one of many possible methods for collapsing the large set of patterns produced by the Pattern Labeling System (PLS) into a smaller set of classes. Results from a number of analyses in this chapter indicate that the particular algorithm chosen for this work provides a meaningful set of type distinctions.

One such result is the TCA produces systematic type profiles for individual speakers in the SWBD corpus, where speakers fell into one or the other of two groups (repeaters or deleters). A second piece of evidence is that the parameter in the model of deletion length differs across DF types, but is relatively invariant by type across domains. Third, the TCA provides a link between complex and basic DFs, since the rates of types occurring as components of complex DFs can be predicted by the rates of types occurring as basic DFs.

6.6.2.5 Flexibility in class groupings

As mentioned in Chapters 2 and 4, a problem noted in some previous systems was the grouping of DF classes based either on surface similarity, or on presumed underlying cause of the DFs. For example, some systems grouped filled pauses and repetitions as a class separate from other types of DFs involving changed words.

In this work, predetermined groupings were avoided in order not to obscure regularities in the data. Results suggest that this was an appropriate approach, because DFs were found to group in different ways depending on the analysis. For example, the analysis of rates of types by speaker showed that FP production was associated with gender, while production of other common DF

types was associated with speech rate. Similarly, the analysis of deletion length for DFs overall (in Chapter 5) showed a cleaner distribution of deletion lengths when data for FPs were removed. The analysis of per-word DF rates across corpora, on the other hand, indicated that FPs could be meaningfully grouped with REP and DEL, in contrast to a group containing SUB, ART, and INS. Finally, FP and REP could be reasonably grouped based on acoustic analyses of duration and F0.

6.6.2.6 Large datasets

Finally, as also indicated in Chapter 5, it was found to be crucial to use a large set of speech data. This was especially important for analyses involving breakdowns by type, because certain types are inherently rare. A large dataset was also particularly important for the analyses of complex DFs and of acoustic properties of DFs. For example, in order to obtain 90 appropriate two-member complex DFs in the SWBD corpus, it was necessary to start with a set of data containing over 40,000 words. Similarly, in order to obtain 17 clearly-recorded cases of repeated "the" from a single speaker (and in the appropriate prosodic context), it was necessary to search a set of data containing nearly 30,000 words.

Chapter 7: Conclusion

7.1 Summary

The study of disfluencies will gain importance as research in psycholinguistics as well as applied fields expands outward from the study of read or ideal speech to the study of natural, spontaneous speech. Although DFs are commonly viewed as irregular events, the goal of this thesis, as set forth in Chapter 1, was to demonstrate that DFs show regularities in a number of dimensions.

Chapter 2 surveyed past work on DF classification systems, and reviewed a variety of descriptive studies. While past research served as an important starting point for the thesis, four requirements for pretheoretical research on DFs remained to be addressed: 1) the development of a theory-neutral classification system; 2) the gathering of robust trends; 3) the examination of relationships among features; and 4) the attainment of predictive power by direct modeling of trends. These requirements guided the methodology and analyses of the thesis.

Given the pretheoretical stage of DF research, the approach adopted was strongly data-driven. An important component of this approach was to use a large amount of speech data. Chapter 3 described the selection of data from three corpora representing different styles of spontaneous speech: task-oriented human-computer dialog, human-human free conversation, and a third, comparison corpus of task-oriented human-human dialog.

In keeping with the data-driven approach, the effort was to identify systematic variation in observable “features” of the data. The study of DFs was expected to cut across areas relevant to the study of fluent speech; therefore, in Chapter 4, a range of features was defined for labeling. A methodology was developed for the annotation of features of the speech domain, the speaker, the sentence in which a DF occurs, and acoustic properties of the DF. A theory-neutral system was developed to represent pattern features of DFs in a single representation, and an algorithm was provided for automatic type classification based on this representation. The annotated data was organized automatically into a database.

Chapters 5 and 6 presented the results of analyses based on the annotated data. Results showed regular trends in DF rates by sentence length, by DF position, by presence of another DF

in the same sentence, by DF type, by filled-pause form, and by combinations of these features both across and within speakers. Regularities were also found for pattern features of the DF, including the number of deleted words, the rate of word fragments, the rate and type of words in the interregnum (filled pauses, editing phrases, and discourse markers), and the occurrence of retraced words. Additional analyses described regularities involving specific DF types, including the type distributions, rate, and compositionality of component members in complex DFs, and the duration and F0 properties of filled pauses and single-word repetitions. Across analyses, the three different speech styles were compared, and simple parametric models were suggested.

7.2 Contributions

This thesis developed a methodology for analyzing disfluencies, and applied it to three large corpora of spontaneous speech. Results showed that there are, indeed, significant regularities in the distribution and characteristics of disfluencies. Thus, the goal of the thesis is met. These regularities can help constrain cognitive theories of speech communication. They can also be exploited to improve the automatic processing of DFs in speech applications.

7.3 Future Work

The contributions of this thesis represent only a preliminary step, however, toward the longer-term goal of a unified theory of DF production. To that end, much work remains to be done. A number of specific areas for future study were noted throughout the analysis sections in Chapters 5 and 6. More generally, however, important broad areas remain to be addressed.

One area is the modeling of aspects of DFs beyond the simple and limited set of features examined in the present work. This includes exploring regularities in syntactic, semantic, prosodic, discourse-related, pragmatic, and cultural factors in DF production. It also includes examining additional styles of spontaneous speech and additional speakers to test the generalizability of results.

A second goal is to develop a comprehensive understanding of the range of options a speaker has when encountering trouble. Therefore, future work is needed to integrate the study

of the DF types examined in this thesis with the study of other, related phenomena (for example, unfilled pauses and discourse markers.)

Third, much insight about DF production could be gained from cross-linguistic studies. For example, comparison of languages differing in syntax could shed light on the syntactic organization of DFs. Examination of languages with greater grammatical agreement than English would make visible certain repairs that cannot be seen on the surface in English (e.g., “a -- a man” but “une -- un homme”). Languages with lexical tone, in which the F0 relationships across syllables are more highly constrained than in English, could be useful for studying the intonation of DFs.

Fourth, future work could be facilitated by methods to reduce time and effort in annotating data. One approach would be to develop tools to automatically label DFs in large databases. A second approach would be to reduce the overall amount of data needed to discern trends, by reducing variability in the data. This could be achieved by the use of within-subject designs or appropriate elicitation methods.

Finally, the long-term goal of the study of speech disfluencies in psycholinguistics is the development of an explanatory theory that will tie in with a broader cognitive model of speech communication. Similarly, for speech applications, although the issues and methods differ considerably from those in theoretical fields, the long-term goal is an integration of DF detection and correction techniques in an overall architecture for the intelligent automatic processing of spontaneous speech.

Bibliography

Adams, M.R. (1982). Fluency, nonfluency, and stuttering in children. *Journal of Fluency Disorders, 7*, pp. 171-185.

Allen, J.F. & Shubert, L.K. (1991). *The TRAINS project.* Technical Report 91-1. Computer Science Dept., University of Rochester.

Allwood, J., Nivre, J., & Ahlsen, E. (1989). Speech management: On the non-written life of speech. *Gothenburg Papers in Theoretical Linguistics,* Vol. *58*.

Baars, B.J. (Ed.). (1992). *Experimental Slips and Human Error: Exploring the Architecture of Volition.* New York, NY: Plenum Press.

Bear, J., Dowding, J., Shriberg, E., & Price, P. (1993). *A system for labeling self-repairs in speech.* Technical Note 522. SRI International.

Bear, J., Dowding, J. & Shriberg, E. (1992). Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog. *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics,* pp. 56-63. Association for Computational Linguistics.

Beattie, G.W. (1979). Planning units in spontaneous speech: Some evidence from hesitation in speech and speaker gaze direction in conversation. *Linguistics, 17*, pp. 61-78.

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

Blankenship, J. & Kay, C. (1964). Hesitation phenomena in English speech: A study in distribution. *Word, 20*, pp. 360-372.

Bock, K. (1991). A sketchbook of production problems. *Journal of Psycholinguistic Research, 20* (3), pp. 141-160.

Boomer, D. S. (1965). Hesitation and grammatical encoding. *Language and Speech, 8*, pp. 148-158.

Broen, P.A. & Siegel, G.M. (1972). Variations in normal speech disfluencies. *Language and Speech, 15*, pp. 219-231.

Brown, E.K. (1980). Grammatical incoherence. In H. W. Dechert and M. Raupach (Eds.), *Temporal Variables in Speech: Studies in Honour of Frieda Goldman-Eisler.* The Hague: Mouton.

Butcher, A. (1980). Pause and syntactic structure. In H. W. Dechert and M. Raupach (Eds.), *Temporal Variables in Speech: Studies in Honour of Frieda Goldman-Eisler.* The Hague: Mouton.

Butcher, A. (1981). *Aspects of the speech pause: Phonetic correlates and communicative functions.* Doctoral dissertation, University of Kiel.

Butzberger, J., Murveit, H., Shriberg, E., & Price, P. (1992). Spontaneous speech effects in large vocabulary speech recognition applications. *Proceedings of the 1992 DARPA Speech and Natural Language Workshop,* pp. 339-343. Harriman, NY.

Carbonell, J.G. & Hayes, P.J. (1983). Recovery strategies for parsing extragrammatical language. *American Journal of Computational Linguistics, 9* (3-4), pp. 123-146.

Carletta, J., Caley, R.J., & Isard, S.I. (1993). *A collection of self-repairs from the map task corpus.* Technical Report TR-47. Human Communication Research Centre, University of Edinburgh.

Christenfeld, N., Schachter, S. & Bilous, F. (1991). Filled pauses and gestures: It's not coincidence. *Journal of Psycholinguistic Research, 20* (1), pp. 1-10.

Clark, H.H. (1994). Managing problems in speaking. *Speech Communication, 15,* pp. 1-8.

Clark, H.H. & Clark, E.V. (1977). *Psychology and Language: An Introduction to Psycholinguistics.* New York: Harcourt Brace Jovanovich.

Cook, M., Smith, J., & Lalljee, M. (1987). Filled pauses and syntactic complexity. *Language and Speech, 17,* 11-16.

Couper-Kuhlen, E. (1992). Contextualizing discourse: The prosody of interactive repair. In Auer, P. and Di Luzio, A. (Eds.), *The Contextualization of Language*. Philadelphia: John Benjamins.

Cutler, A. (1983). Speakers' conceptions of the function of prosody. In Cutler, A. and Ladd, D.R. (Eds.), *Prosody: Models and Measurements.* New York: Springer-Verlag.

Dahl, D.A., Bates, M., Brown, M., Fisher, W., Hunicke-Smith, K., Pallett, D., Pao, C., Rudnicky, A., & Shriberg, E. (1994). Expanding the scope of the ATIS task: The ATIS-3 corpus. *Proceedings of the 1994 DARPA Speech and Natural Language Workshop,* pp. 43-48. Princeton, NJ.

Dechert, H.W. & Raupach, M. (Eds.). (1980). *Temporal Variables in Speech.* The Hague: Mouton.

Deese, J. (1980). Pauses, prosody, and the demands of production in language. In H. W. Dechert and M. Raupach (Eds.), *Temporal Variables in Speech: Studies in Honour of Frieda Goldman-Eisler.* The Hague: Mouton.

Delattre, P. (1965). *Comparing the Phonetic Features of English, French, German and Spanish*. Heidelberg: Julius Groos Verlag.

Dickerson, W.B. (1971). Hesitation phenomena in the spontaneous speech of non-native speakers of English. Doctoral dissertation, U. of Illinois at Urbana.

Du Bois, J.W. (1974). Syntax in mid-sentence. *Berkeley Studies in Syntax and Semantics, 1,* pp. III-1 to III-25.

Erman, B. (1987). *Pragmatic Expressions in English: A study of 'You Know', 'You See', and 'I mean' in Face to Face Conversation*. Stockholm: Almqvist and Wiksell International.

Esling, J.H. (1971). Preliminary report on a research project in acoustic phonetics with comparisons of the vowel systems of several languages and a discussion of the neutral position. Unpublished manuscript, U. of Michigan.

Fienberg, S.E. (1991). *The analysis of cross-classified categorical data*. Cambridge, MA: MIT Press.

Forney, G.D. (1973). The Viterbi algorithm. *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing, 61*, pp. 268-278.

Fowler, C.A. & Housum, J. (1987). Talkers' signalling of 'new' and 'old' words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language, 26*, pp. 489-504.

Fox Tree, J.E. (1993). *Comprehension after speech disfluencies.* Doctoral dissertation, Stanford University.

Fox Tree, J.E. & Clark, H.H. (1994). Pronouncing 'the' as /thiy/ to signal trouble in spontaneous conversation. Paper presented at the Meeting of the Psychonomics Society, St. Louis, MO.

Fromkin, V.A. (Ed.). (1980). *Errors in Linguistic Performance: Slips of the Tongue, Ear, Pen and Hand.* New York: Academic Press.

Geluykens, R. (1987). Tails (right-dislocations) as a repair mechanism in English conversation. In Nuyts, J. and de Schutter, G. (Eds.), *Getting One's Words into Line: On Word Order and Functional Grammar.* Dordrecht: Foris Publications.

Godfrey, J.J., Holliman, E.C. & McDaniel. (1992). SWITCHBOARD: Telephone speech corpus for research and development. *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing,* pp. 517-520. San Francisco: IEEE.

Goldman-Eisler, F. (1961). A comparative study of two hesitation phenomena. *Language and Speech, 4*, pp. 18-26.

Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in Spontaneous Speech.* New York: Academic Press.

Good, D.A. & Butterworth, B.L. (1980). Hesitancy as a conversational resource: Some methodological implications. In H.W. Dechert and M Raupach (Eds.), *Temporal Variables in Speech: Studies in Honour of Frieda Goldman-Eisler.* The Hague: Mouton.

Goodman, L.A. (1970). The multivariate analysis of qualitative data: Interactions among multiple classifications. J*ournal of the American Statistical Association, 65*, pp. 226-256.

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

Goodwin, C. (1986). Between and within: Alternative and sequential treatments of continuers and assessments. *Human Studies, 9*, pp. 205-217.

Guaïtella, I. (1993). Functional, acoustical and perceptual analysis of vocal hesitations in spontaneous speech. *Proceedings of the ESCA Workshop on Prosody. Working Papers 41*, pp. 128-131. Dept. of Linguistics and Phonetics, Lund, Sweden.

't Hart, J., Collier, R., & Cohen, A. (1990). *A Perceptual Study of Intonation*. Cambridge, England: Cambridge University Press.

Hawkins, P. R. (1971). The syntactic location of hesitation pauses. *Language and Speech, 14*, pp. 277-288.

Heeman, P. & Allen, J. (1994). Detecting and correcting speech repairs. *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pp. 295-302. Association for Computational Linguistics.

Hieke, A.E. (1981). A content-processing view of hesitation phenomena. *Language and Speech, 24* (2), pp. 147-160.

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

Hockett, C.F. (1967). Where the tongue slips, there slip I. In *To Honor Roman Jacobson: Vol. 2*. The Hague: Mouton.

Howell, P. & Young, K. (1991). The use of prosody in highlighting alterations in repairs from unrestricted speech. *The Quarterly Journal of Experimental Psychology, 43A* (3), pp. 733-758.

James, D.M. (1972). Some aspects of the syntax and semantics of interjections. *Papers from the Eighth Regional Meeting of the Chicago Linguistic Society*, pp. 162-172. Chicago Linguistic Society.

James, D.M. (1973). Another look at, say, some grammatical constraints on, oh, interjections and hesitations. *Papers from the Ninth Regional Meeting of the Chicago Linguistic Society*, pp. 242-251. Chicago Linguistic Society.

Kasl, S.V., & Mahl, G.V. (1965). The relationship of disturbances and hesitations in spontaneous speech to anxiety. *Journal of Personality and Social Psychology, 1*, pp. 425-433.

Kowal, S., O'Connell, D.C. & Sabin, E.J. (1975). Development of temporal patterning and vocal hesitations in spontaneous narratives. *Journal of Psycholinguistic Research, 4* (3), pp. 195-207.

Kowtko, J.C. & Price, P.J. (1989). Data collection and analysis in the air travel planning domain. *Proceedings of the DARPA Speech and Natural Language Workshop*, pp. 119-125. Cape Cod.

Lalljee, M.G. & Cook, M. (1969). An experimental investigation of the function of filled pauses in speech. *Language and Speech, 12*, pp. 24-29.

Langer, H. (1990). Syntactic normalization of spontaneous speech. *Proceedings of COLING 90,* pp. 180-183.

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

Levelt, W.J.M. (1989). *Speaking: From Intention to Articulation*. Cambridge, Mass.: MIT Press.

Levelt, W.J.M. & Cutler, A. (1983). Prosodic marking in speech repair. *Journal of Semantics, 2* (2), pp. 205-217.

Levin, H. & Silverman, I. (1965). Hesitation phenomena in children's speech. *Language and Speech, 8*, pp. 67-85.

Lickley, R.J. (1994). *Detecting Disfluency in Spontaneous Speech.* Doctoral dissertation, University of Edinburgh.

Lickley, R.J. & Bard, E.G. (1992). Processing disfluent speech: Recognising disfluency before lexical access. *Proceedings of the International Conference on Spoken Language Processing,* pp. 935-938. Banff, Alberta, Canada.

Lickley, R.J., Bard, E.G., & Shillcock R.C. (1991). Understanding disfluent speech: Is there an editing signal? *Proceedings of the International Congress of Phonetic Sciences,* pp. 98-101. Aix-en-Provence, France.

Lickley, R.J., Shillcock R.C. & Bard, E.G. (1991). Processing disfluent speech: How and when are disfluencies found? *Proceedings of Eurospeech 91,* pp. 1499-1502. Genova, Italy.

Lounsbury, F.G. (1954). Transitional probability, linguistic structure and systems of habit-family hierarchies. In Osgood, C.E. and Sebeok, T. (Eds.), *Psycholinguistics: A Survey of Theory and Research Problems*. Baltimore: Waverly Press, Inc.

Lutz, K.C. & Mallard, A.R. (1986). Disfluencies and rate of speech in young adult nonstutterers. *Journal of Fluency Disorders, 11*, pp. 307-316.

Maclay, H. & Osgood, C.E. (1959). Hesitation phenomena in spontaneous English speech. *Word, 15*, pp. 19-44.

MADCOW. (1992). Multi-site data collection for a spoken language corpus. *Proceedings of the fifth DARPA Speech and Natural Language Workshop*, pp. 7-14. Morgan Kaufmann.

Mahl, G.F. (1956). Disturbances and silences in the patient's speech in psychotherapy. *Journal of Abnormal and Social Psychology, 53*, pp. 1-15.

Martin, J.G. & Strange, W. (1968). The perception of hesitation in spontaneous speech. *Perception and Psychophysics, 3* (6), pp. 427-438.

Martin, J.G. (1967). Hesitation in speakers' production and listeners' reproduction of utterances. *Journal of Verbal Learning and Verbal Behavior, 6*, pp. 903-909.

Moore, R.K., & Browning, S.R. (1992). Results of an exercise to collect 'genuine' spoken enquiries using Wizard of Oz techniques. *Proceedings of the Institute of Acoustics 1992 Autumn Conference on Speech and Hearing*, pp. 613-620. Windermere.

Nakatani, C.H. & Hirschberg, J. (1994). A corpus-based study of repair cues in spontaneous speech. *Journal of the Acoustical Society of America, 95* (3), pp. 1603-1616.

Nooteboom, S.G. (1980). Speaking and unspeaking: Detection and correction of phonological and lexical errors in spontaneous speech. In Fromkin, V.A. (Ed.), *Errors in Linguistic Performance: Slips of the Tongue, Ear, Pen and Hand.* New York: Academic Press.

O'Connell, D.C. & Kowal, S. (1983). Pausology. In Sedelow, W.A. and Sedelow, S.Y. (Eds.), *Computers in Language Research 2.* The Hague: Mouton.

O'Shaughnessy, D. (1992). Recognition of hesitations in spontaneous speech. *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing,* pp. 521-524. San Francisco, CA. IEEE.

O'Shaughnessy, D. (1993). Analysis and automatic recognition of false starts in spontaneous speech. *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, pp. 724-727. Minneapolis, MN. IEEE.

O'Shaughnessy, D. (1994). Correcting complex false starts in spontaneous speech. *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing,* pp. 349-352. Adelaide, S. Australia. IEEE.

Okada, M. & Otsuka, H. (1993). Incremental elaboration in generating spontaneous speech. *Proceedings of the International Symposium on Spoken Dialogue: New Directions in Human and Man-Machine Communication*, pp. 49-52. Waseda, Japan.

Osgood, C.E. & Sebeok, T. (Eds.). (1954). *Psycholinguistics: A Survey of Theory and Research Problems.* Baltimore Waverly Press.

Ostendorf, M., Price, P., & Shattuck-Hufnagel, S. (1994). *Evaluating the use of prosodic information in speech recognition and understanding.* Research Report, National Science Foundation and Advanced Research Projects Administration.

Oviatt, S.L. (To Appear). Predicting spoken disfluencies during human-computer interaction. *Computer Speech and Language*.

Peters, A.M. & Menn, L. (1993). False starts and filler syllables: Ways to learn grammatical morphemes. *Language, 69* (4), pp. 742-777.

Pierrehumbert, J. (1980). The phonology and phonetics of English intonation. Doctoral dissertation, MIT.

Postma, A., Kolk, H. & Povel, D-J. (1990). On the relation among speech errors, disfluencies and self-repairs. *Language and Speech, 33* (1), pp. 19-29.

Postma, A. & Kolk, H. (1992). The effects of noise masking and required accuracy on speech errors, disfluencies, and self-repairs. *Journal of Speech and Hearing Research, 35*, pp. 537-544.

Redeker, G. (1991). Linguistic markers of discourse structure. *Linguistics, 29*, pp. 1139-1172.

Reynolds, A. & Paivio, A. (1968). Cognitive and emotional determinants of speech. *Canadian Journal of Psychology, 22*, pp. 164-175.

Rochester, S. R. (1973). The significance of pauses in spontaneous speech. *Journal of Psycholinguistic Research, 2* (1), pp. 51-81.

Sadanobu, T. & Takubo, Y. (1993). The discourse management function of fillers: A case of 'eeto' and 'ano(o)'. *Proceedings of the International Symposium on Spoken Dialogue: New Directions in Human and Man-Machine Communication*, pp. 271-274. Waseda, Japan.

Schachter, S., Christenfeld, N., Ravina, B. & Bilous, F. (1991). Speech disfluency and the structure of knowledge. *Journal of Personality and Social Psychology, 60* (3), pp. 362-367.

Schegloff, E.A. (1979). The relevance of repair to syntax-for-conversation. In Givon, T. (Ed.), *Syntax and Semantics* 12: *Discourse and Syntax.* New York: Academic Press.

Schegloff, E.A. (1987). Recycled turn beginnings: A precise repair mechanism in conversation's turn-taking organization. In G. Button and J. R. E. Lee (Eds.), *Talk and Social Organisation.* Clevedon: Multilingual Matters Ltd.

Schegloff, E.A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language, 53*, pp. 361-382.

Schiffrin, D. (1987). *Discourse Markers.* New York: Cambridge University Press.

Selting, M. (1988). The role of intonation in the organization of repair and problem handling sequences in conversation. *Journal of Pragmatics, 12*, pp. 293-322.

Shattuck-Hufnagel, S. (1986). The representation of phonological information during speech production planning: Evidence from vowel errors in spontaneous speech. *Phonology Yearbook, 3*, pp. 117-149.

Shriberg, E.E. (1991). Intonation of filled pauses in spontaneous speech. Presentation, Conference on Grammatical Foundations of Prosody and Discourse, UC Santa Cruz, July 1991.

Shriberg, E.E. (1994). Language modeling for SWITCHBOARD. Presentation, SRI International, March 1994.

Shriberg, E.E. (in preparation (a)). Filled pauses and transition probability in the ATIS corpus.

Shriberg, E.E. (in preparation (b)). A model of repetition frequency in the SWITCHBOARD corpus.

Shriberg, 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.* Harriman, NY.

Shriberg, E.E. & Lickley, R.J. (1992a). Intonation of clause-internal filled pauses. *Proceedings of the International Conference on Spoken Language Processing,* pp. 991-994. Banff, Alberta, Canada.

Shriberg, E.E. & Lickley, R.J. (1992b). The relationship of filled-pause F0 to prosodic context. *Proceedings of the IRCS Workshop on Prosody in Natural Speech, Technical Report IRCS-92-37,* pp. 201-209. University of Pennsylvania, Institute for Research in Cognitive Science, Philadelphia, PA.

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

Shriberg, E., Wade, E. & Price, P. (1992). Human-machine problem solving using spoken language systems (SLS): Factors affecting performance and user satisfaction. *Proceedings of the DARPA Speech and Natural Language Workshop,* pp. 49-54.

Siegel, G.M., Lenske, J, & Broen, P. (1969). Suppression of normal speech disfluencies through response cost. *Journal of Applied Behavior Analysis, 2*, p. 265.

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

Suhm, B., Levin, L., Coccaro, N., Carbonell, J., Horiguchi, K., Isotani, R., Lavie, A., Mayfield, L., Rose, C.P., Van Ess-Dykema, C., & Waibel, A. (1994). Speech-language integration in a multi-lingual speech translation system. *Proceedings of the AAAI-94 Workshop on Integration of Natural Language and Speech Processing.* Seattle, WA.

Taylor, I. (1969). Content and structure in sentence production. *Journal of Verbal Learning and Verbal Behavior, 8*, pp. 246-250.

d'Urso, V. & Zammuner, V. (1990). The perception of pause in question-answer pairs. *Bulletin of the Psychonomic Society, 28* (1), pp. 41-43.

Van Wijk, C. & Kempen, G. (1987). A dual system for producing self repairs in spontaneous speech: Evidence from experimentally elicited corrections. *Cognitive Psychology, 19* (4), pp. 403-440.

Viterbi, A.J. (1967). Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Transactions on Information Theory, IT-13*, pp. 260-269.

Voss, B. (1979). Hesitation phenomena as sources of perceptual errors for non-native speakers. *Language and Speech, 22* (2).

Wade, E., Shriberg, E.E., & Price, P.J. (1992). User behaviors affecting speech recognition. *Proceedings of the International Conference on Spoken Language Processing*, pp. 995-998. Banff, Alberta, Canada.

Wheatley, B., Doddington, G., Hemphill, C., Godfrey, J., Holliman, E., McDaniel, J., & Fisher, D. (1992). Robust automatic time alignment of orthographic transcriptions with unconstrained speech. *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, pp. 521-524. San Francisco, CA, IEEE.

Wingate, M.E. (1984). Fluency, disfluency, dysfluency, and stuttering. *Journal of Fluency Disorders, 17*, pp. 163-168.