Spoken Dialogue Understanding and Local Context

Peter A. Heeman

Technical Report 523
July 1994

UNIVERSITY OF
ROCHESTER
COMPUTER SCIENCE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	July 1994	technical report

4. TITLE AND SUBTITLE
Spoken Dialogue Understanding and Local Context

5. FUNDING NUMBERS
N00014-92-J-1512

6. AUTHOR(S)
Peter A. Heeman

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Computer Science Dept.
University of Rochester
734 Computer Studies Bldg.
Rochester, NY 14627-0226

8. PERFORMING ORGANIZATION REPORT NUMBER
TR 523

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research
Information Systems
Arlington, VA 22217

ARPA
3701 N Fairfax Drive
Arlington, VA 22203

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution of this document is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*
(see title page)

14. SUBJECT TERMS
spoken dialogue; speech repairs; utterances; discourse markers

15. NUMBER OF PAGES
51

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
unclassified	unclassified	unclassified	UL

Spoken Dialogue Understanding and Local Context

Peter A. Heeman

The University of Rochester
Computer Science Department
Rochester, New York 14627

Technical Report 523

July 1994

Abstract

Spoken dialogue poses many new problems to researchers in the field of computational linguistics. In particular, conversants must detect and correct speech repairs, segment a turn into individual utterances, and identify discourse markers. These problems are interrelated. For instance, there are some lexical items whose role in an utterance can be ambiguous: they can act as discourse markers, signal a speech repair, or even be part of the content of an utterance unit. So, these issues must be addressed together. The resolution of these problems will allow a basic understanding of how a speaker's turn can be broken down into individual contributions to the dialogue. We propose that this resolution must be and can be done using local context. They do not require a full understanding of the dialogue so far, nor, in most cases, a deep understanding of the current turn. Resolving these issues locally also means they can be resolved for the most part before later processing, and so will make a natural language understanding system more robust and able to deal with the unconstrained nature of spoken dialogue.

Funding gratefully received from the Natural Sciences and Engineering Research Council of Canada, from NSF under Grant IRI-90-13160, and from ONR/DARPA under Grant N00014-92-J-1512.

Contents

Chapter 1

Introduction

1.1 Spoken Dialogue

In the past few years, many researchers have started tackling the problems that arise in spoken dialogue. So far, the emphasis has been concentrated on single utterance queries, as exemplified by the ATIS project (MADCOW, 1992). The aim of the ATIS project is to make a spoken language system that can give air travel information. The following illustrates a typical utterance.

> I would like a flight that leaves after noon in San Francisco and arrives before 7 p.m. Dallas time

The queries, as the example demonstrates, look very text-like. The main exception is the presence of speech repairs, which Bear, Dowding, and Shriberg (1992) report happen in "10% of sentences longer than nine words." Good results in understanding have been achieved in this domain, as evidenced by the results of Dowding et al. (1993), who report syntactic and semantic understanding of 86% of all utterances. So, this task gives the allusion that spontaneous speech understanding corresponds to our intuitions about understanding text.

However, such spoken queries to not capture the problems inherit in natural dialogues. Unlike text and queries to a database system, the speech in natural dialogues is very unconstrained and often ungrammatical. As part of the TRAINS project (Allen and Schubert, 1991), which is a long term research project to build a conversationally proficient planning assistant, we are collecting a corpus of problem solving dialogues (Gross, Allen, and Traum, 1993; Heeman and Allen, 1994e). These dialogues involve two participants, one who is playing the role of a user and has a certain task to accomplish, and another, who is playing the role of the system by acting as a planning assistant. In these dialogues, we have found a number of phenomena that pose difficulties for natural language understanding systems.

In this thesis proposal, we will be focusing on detecting and correcting speech repairs, identifying utterance units, and identifying discourse markers. These problems are interrelated. For instance, there are some lexical items whose role in an utterance can be

ambiguous; for they can act as discourse markers, signal a speech repair, or even be part of the content of an utterance unit. The determination of which role it is playing will hence influence all three issues. So, these issues must be addressed together, and the resolution of these problems will allow a basic understanding of how a speaker's turn can be broken down into individual contributions to the dialogue.

1.1.1 Speech Repairs

The first problem is the abundance of speech repairs. A speech repair is a dysfluency where some of the words that the speaker utters need to be removed in order to correctly determine the speaker's meaning. In our corpus of problem solving dialogues, 25% of turns contain at least one repair, 67% of repairs occur with at least one other repair in the turn, and repairs in the same turn occur on average within 6 words of each other.

Speech repairs can be divided into three types: *abridged*, *modification* and *fresh starts*. An abridged repair is where the repair consists solely of a word fragment and/or editing terms, such as "um" and "uh". The following is an example.[1]

> we need to – um manage to get the bananas to Dansville more quickly (d93-14.3 utt50)

A modification repair is where the speech repair modifies what was just said.

> after the orange juice is at – the oranges are at the OJ factory (d93-19.3 utt59)

A fresh start is where the speaker abandons what she was saying and starts again.

> the current plan is we take – okay let's say we start with the bananas (d93-19.3 utt59)

These examples also illustrate how speech repairs can be divided into three intervals: the removed text, the editing terms, and the resumed text (Levelt, 1983; Nakatani and Hirschberg, 1993). The removed text, which might end in a word fragment, is the text that the speaker intends to replace. The end of the removed text is called the interruption point, which is marked in the above examples as "–". This is then followed by editing terms, which can be either filled pauses, such as "um", "uh", and "er", or cue phrases, such as "I mean", "I guess", and "well". The last interval is the resumed text, the text that is intended to replace the removed text. (All three intervals need not be present in a given speech repair.) In order to correct a speech repair, the removed text and the editing terms need to be deleted in order to determine what the speaker meant to say.[2]

---

[1]For each excerpt, we give the dialogue name and the utterance file that is is taken from. Dialogue names that start with "d91" are available in Gross, Allen, and Traum (1993), and ones that start with "d92" and "d93" are available in Heeman and Allen (1994e).

[2]The removed text and editing terms might still contain pragmatic information, as the following example displays, "Peter was ... well ... he was fired."

In order to understand what the speaker is saying, the hearer must detect the speech repair and determine the appropriate correction. All three types of repairs can be problematic. Consider modification repairs, which are often signaled by strong correspondences between the removed text and the resumed text. However, an algorithm that uses word correspondences could be led astray, since subsequent utterances often build on, or even repeat, what was previously said (Walker, 1993). Consider the following utterance (d93-8.3 utt79).

> that's all you need
> you only need one tanker

In this example, the end of the first utterance and the beginning of the second both contain the words "you" and "need". So, without doing any segmentation of the dialogue, one might mistakenly think that the first instance of "you need" is being replaced by "you only need". It is only when segmentation is taken into account that one discovers that the speaker was simply repeating herself for emphasis.

Even abridged repairs can be problematic. These repairs, which require only a word fragment or editing term to be removed, can sometimes be mistaken as modification repairs due to spurious word correspondences. For instance, consider the abridged repair given earlier, repeated here for the reader's convenience.

> we need to – um manage to get the bananas to Dansville more quickly

In this example, the text "manage to" might be postulated as replacing "need to", which would result in an incorrect correction. Even the editing terms, which include lexical items, such as "okay", "yeah", and "oh", and filled pauses, such as "um" and "uh", can be problematic, since they can also appear as part of the content of an utterance, or play some other discourse purpose other than to mark a repair.

More difficult yet are the fresh starts, in which a speaker abandons what she is saying and starts over. Does she abandon everything in the turn preceding the interruption point, including complete utterances? Consider the following excerpt (d93-10.3 utt20).

> go back to Elmira
> can they take tankers with them
> or how does – uh
> do tankers move by themselves

In the example above, the speaker probably only intends to cancel just the "how does". So how does the hearer determine how much of the preceding text needs to be deleted, especially where the fresh start is embedded in a turn?

1.1.2 Utterance Units

Brown and Yule (1983) discuss a number of ways in which speech differs from text. The syntax of spoken dialogue is typically much less structured than that of text: it contains

fragments, there is little subordination, and it lacks the meta-lingual markers between clauses. It also tends to come in installments and refinements, and makes use of topic-comment sentence structure. The following example illustrates what could be taken as an example of two fragments, or as an example of a topic-comment sentence (d92-1 utt5).

> so from Corning to Bath
> how far is that

Crystal (1980) presents some additional problems with viewing speech as sentences and clauses. So not only does speech lack the punctuation marks of written text, but it also lacks its syntactic well-formedness.

Although speech cannot always be mapped onto sentences, there is wide agreement that speech does come in sentence-like packages, which are called *utterances*. So, the question arises as to how a hearer can determine where one utterance ends and the next starts. From work done in linguistics, there is agreement that there are a number of factors that contribute to this, including intonation, pauses, and syntactic structure. But, no one has made a computation model that can deal with turns with more than one utterance unit. Even if segmenting a turn into utterance units is not needed for syntactic and semantic analysis, which is unlikely, it is necessary for correcting fresh starts and detecting discourse markers.

To illustrate how difficult segmenting a turn can be, consider the following single turn from a problem solving dialogue (d93-13.1 utt36).

> I don't know if I can give any help on that either let's see it takes one hour to load or unload any amount of cargo on a train so I guess that would be an hour no matter how many boxcars you have

The hearer must segment this into individual utterances. Consider the "or" halfway through. This could be a discourse marker, signaling that the speaker is giving a second alternative, that being a request to unload the oranges. Or, it could signal a speech repair; thus the speaker is saying that it takes an hour to unload a boxcar, but not necessarily an hour to load one. Or, it could be part of the sentential content, so either loading or unloading takes an hour.

Another example that illustrates the difficulty of segmenting a turn occurs on the phrase "no matter". It could end the preceding utterance, leaving the final utterance as the question "how many boxcars you have". (The ungrammaticality of this utterance cannot rule it out, since function words such as "do" are often shortened, and so hard to detect.) The other possibility is that "no matter" is part of the last clause; thus the speaker is saying that it takes an hour, no matter how many boxcars that one has. Below, we give a hand-segmented transcription of this turn, giving the interpretation that is suggested by the intonation.

> I don't know if I can give any help on that either
> let's see
> it takes one hour to load or unload any amount of cargo on a train
> so I guess that would be an hour
> no matter how many boxcars you have

So, a first step to understanding the content of a turn is to identify where each utterance unit starts and stops.

1.1.3 Discourse Markers

Phrases such as "so", "now", "firstly," "moreover", and "anyways", are referred to as discourse markers in discourse analysis. They are conjectured to give the hearer information about the discourse structure, and so aid the hearer in understanding how the new text relates to what was previously said (Litman and Hirschberg, 1990).

Although, some discourse markers, such as "firstly", and "moreover", are not commonly used in spoken dialogue, as noted above by Brown and Yule (1983), there are a lot of other discourse markers that are employed. These discourse markers are used to achieve a variety of effects: signal an acknowledgment or acceptance, hold a turn, signal a speech repair, stall for time, or signal an interruption or the return from one. These uses are concerned with the interactional aspects of discourse rather than adding to the content. These phrases, we feel, play a crucial role in understanding a turn, and so it is important that they be identified.

Consider the following utterance (d92-1 utt37).

> three hours
> then from Dansville to Corning

The first part of the utterance "three hours", is repeating what the other conversant just said, which was a response to the question "how far is it from Avon to Dansville". After repeating "three hours", the speaker then asks the next question, "from Dansville to Corning". Only by properly recognizing that "then" is acting as a discourse marker to relate the two questions, rather then as part of the content of a single utterance, will the other conversant properly understand the speaker's turn.

1.1.4 Incremental Understanding

Unlike written text, spoken dialogue understanding requires that the speaker's turn be understood incrementally, as the turn progresses, rather than waiting for the end of the speaker's turn. The following example illustrates how the hearer is in fact able to jump in and interrupt a speaker if a problem arises (d91-6.1 utt10).

User:	so we should move the engine at Avon
	Engine E
	to
System:	engine E one
User:	engine E one
	to Bath

In the excerpt above, the hearer interrupts the speaker after "to" is uttered in order to correct the identity of the engine to be used. So, the hearer must be incrementally understanding the speaker's turn in order to realize that a mistake was made. In fact, turn-taking,

which is negotiated by the two conversants, depends on the hearer having understood, or having not understood, what the speaker is saying (Sacks, Schegloff, and Jefferson, 1974). What this means is that any solution to identifying speech repairs, utterance units and discourse markers must be able to do so while processing the speech incrementally, rather than waiting till the end of the turn.

1.2 Thesis Proposal

Before we propose how these problematic issues in spoken dialogue can be resolved, we need to examine the alternative. Is it possible to structure the interaction between the user and the system so as to remove these problematic issues (Oviatt, 1994). Is it possible to make problem-solving dialogues more like the queries in the ATIS domain? In the ATIS domain, the user is allowed to think off-line and presses a button when he wants to speak. As well, the user is not expecting help in making the plan; he only expects answers to the queries that he makes. One could argue that a goal in understanding dialogue should be to find ways in which to constrain the user's actions in order to make understanding easier. But we feel that for problem-solving tasks such restrictions will inhibit the participants' ability to collaborate in the task at hand.

Given that a goal of spoken dialogue research should not be to find ways to constrain the manner in which participants interact, we need to determine how the above problems can be tackled. Given that we must do incremental processing, these problems—detecting and correcting speech repairs, segmenting a turn into utterance units, and identifying discourse markers—cannot wait till the end of the turn before they are resolved. They must be resolved online, as the speech is spoken. Hence, we can not rely on a syntactic or semantic analysis of the entire turn.

Furthermore, since these problems concern interactional activity, rather than transactional ones, we feel that they are best handled outside of the range of syntactic and semantic processing. Consider, modification repairs. These repairs are often accompanied by a retracing of some of the text, which is postulated as a device that the speaker uses to enable the hearer to determine the correction (Levelt, 1983). Such retracing results in cross-serial word correspondences, which cannot be expressed in most grammar formalisms. Instead, they can easily be found by a pattern matcher outside the formalism (Bear, Dowding, and Shriberg, 1992; Heeman and Allen, 1994a). Furthermore, the detection of these repairs is not signaled by a clearly defined *edit signal* (cf. Labov, 1966). Rather, there are a number of clues, including word correspondences, presence of a word fragment, presence of editing terms, and prosodic clues that signal speech repairs, and these clues need to be combined (Nakatani and Hirschberg, 1993; Heeman and Allen, 1994b). Combining multiple clues is a task that also does not fit well into most grammar formalisms.

In this thesis proposal, we address how these issues in spoken dialogue can, for the most part, be resolved using local context, in order words, without analyzing the entire turn. In fact, we feel that these issues can be resolved as they are encountered, within a small window of words of their occurrence. The local clues that can be employed include lexical knowledge, intonational knowledge, as well as syntactic knowledge. The result will

be an analysis of the turn with speech repairs, utterance boundaries, and discourse phrases marked.

Although some syntactic and semantic knowledge might be needed to resolve some instances, we feel that the majority will not. This means that for the most part these issues will be resolved outside of syntactic and semantic processing and so those modules will have a chance to deal with the impoverished conditions found in spoken dialogue.

1.3 Related Work

1.3.1 Utterance Units

The segmentation of spoken speech into utterance units touches on several areas of research. Halliday (1967) proposed that intonation serves to break up the speech into informational units. Gee and Grosjean (1983) proposed that the intonational structure is the basic structure in language processing. In fact, we believe it is probably the best way to segment a turn into smaller parts, such as installments, refinements, and clauses, which exhibit more regular syntactic structure from a text standpoint. However, finding intonational boundaries reliably is a problem that is just starting to be addressed, and will prove difficult since the acoustic clues alone are probably too impoverished. "When we consider spontaneous speech (particularly conversation) any clear and obvious division into intonational-groups is not so apparent because of the broken nature of much spontaneous speech, including as it does hesitation, repetitions, false starts, incomplete sentences, and sentences involving a grammatical caesura in their middle" (Cruttenden, 1986, pg. 36).

The definition of utterance units can benefit from the work done on *grounding*. Spoken dialogue does not involve one-way understanding. The participants work together to achieve mutual understanding of what is said. Clark and his colleagues (Clark and Wilkes-Gibbs, 1986; Clark and Schaefer, 1989; Clark, 1992) have shown that an important part of dialogue is obtaining *mutual understanding* of the *contributions* to the dialogue, and they refer to this process as grounding. In fact an overhearer, who is not part of the grounding process, is less likely to understand the dialogue than the participants in the dialogue (Clark and Schober, 1989). These contributions might correspond to utterance units; so the intonational boundaries would then have an intentional interpretation.

Utterance units also need to be related to *speech acts*, which are the result of viewing language as action (Grice, 1957; Austin, 1962; Searle, 1969). Speakers act by way of their utterances to accomplish various effects. The actions include such things as promising, informing, and requesting. Starting with Cohen and Perrault (1979) and Allen and Perrault (1980), there has been considerable work done in formulating a computational model of speech actions and in stating the effects of such actions on the participants involved. Speech actions have also been extended to grounding and turn-taking (e.g., Traum and Hinkelman, 1992; Novick, 1991).

1.3.2 Speech Repairs

Previous work in speech repairs has examined different approaches to detecting and correcting speech repairs. One of the first was Hindle (1983), who added grammar rules to a deterministic parser to handle speech repairs. This work was based on research that indicated that the resumed text of a speech repair replaces text of the same category. However, Hindle assumed an edit signal would mark the interruption point, a signal that has yet to be found. Another approach, taken by Bear, Dowding, and Shriberg (1992), uses a pattern matcher to look for patterns of matching words. In related work, Dowding et al. (1993) employed a parser-first approach. If the parser and semantic analyzer are unable to make sense of an utterance, they look for speech repairs using the pattern matcher just mentioned. A number of researchers have also looked at intonational characteristics of speech repairs (O'Shaughnessy, 1992; Lickley, Shillcock, and Bard, 1991; Lickley and Bard, 1992; Bear, Dowding, and Shriberg, 1992; Nakatani and Hirschberg, 1993). Of these, Nakatani and Hirschberg have tried using these intonational features to detect speech repairs. They used hand-transcribed features, including duration, presence of fragments, presence of filled pauses, and lexical matching.

1.3.3 Discourse Markers

Many researchers have noted the importance of discourse markers (Cohen, 1984; Reichman-Adar, 1984; Sidner, 1985; Grosz and Sidner, 1986; Litman and Allen, 1987), both for determining discourse structure and for interpreting anaphoric reference. Little has actually been done in identifying them in spoken dialogue. One exception is work done by Litman and Hirschberg (1990), who looked at how intonational information can disambiguate lexical items that can either be a discourse marker or have a sentential reading.

1.4 Contribution

By making use of local context, we expect that a number of problems in dialogue understanding can be resolved, without the aid of syntactic and semantic analysis of the whole turn. The most important contribution will be to show that speech repairs can be detected and corrected using local clues. This will show that this phenomenon can be resolved before syntactic and semantic processing, and hence will simplify those processes.

Second, we will show that turns can be segmented into individual contributions, early on in speech understanding. This will undoubtedly use a combination of intonational information and local syntactic information. Again, this result will show that latter modules will not have to deal with this complexity in spoken dialogue understanding.

A third contribution will be the identification of discourse markers. Many lexical items that can be discourse markers can also act as part of the content of a larger contribution. For instance, phrases like "okay", "alright", can be the adjective in an utterance, can be part of a speech repair, or can stand on their own as an acknowledgment. We will show that these can be tagged as for which role that they are playing.

These contributions will be summed up in an algorithm that will take as input the word transcriptions and intonational phrase markings, and it will mark the parts of speech repairs, tag discourse markers, and segment turns into utterance units. This algorithm will run incrementally and will output the tags so higher level processing can take advantage of them.

On a more theoretical note, our work will give a better understanding of how conversants interact in a dialogue to deal with the local phenomena that can hinder understanding. As well, our work on utterance units will show that they are intonationally marked, and that conversants use them as the unit of exchange, and so are the cornerstone of theories on grounding.

Chapter 2

Related Work

The problems of detecting and correcting speech repairs, and identifying utterance units and discourse markers has received a lot of attention in the literature. Some of the relevant literature is in the field of computational linguistics, while the rest lies in philosophy, linguistics, psychology, and sociology. We start with a literature review of intonation, since intonation interacts with all the issues that we want to address. We then move to speech repairs, utterance units, and discourse markers. At the end of the chapter, we present work done in speech actions and conversational analysis. Speech actions captures how language usage relates to the mental state of an agent, so it gives a long range outlook as to what we can do as far as representing the content of a turn. Conversation analysis is concerned with the study of actual dialogues and describing the phenomena that arise in it. So this area will help us better understand the issues that arise in analyzing a speaker's turn.

2.1 Intonational Phrases

When people speak, they tend not to speak in a monotone. Rather, the pitch of their voice, as well as other characteristics, such as speech rate and loudness, varies as they speak. (Pitch is also referred to as the fundamental frequency, or $\mathbf{f_0}$ for short.) The study of intonation is concerned with describing this phenomenon and determining its communicative meaning. For instance, as most speakers of English implicitly know, a statement can be turned into a question by ending it with a rising pitch.

2.2 Representing Intonation

Pierrehumbert (1980) presented a model of intonation patterns. Her model describes English intonation as a series of highs (**H**) and lows (**L**). (The formulation that we use is a slight variant on this, and is described by Pierrehumbert and Hirschberg (1990).) The lowest level of analysis is at the word level, in which stressed words are marked with either a high or low *pitch accent.* The next level is the *intermediate phrase*, which consists of at least one stressed word, plus a high or low tone at the end of the phrase. This *phrasal tone* controls

the pitch contour between the last pitch accent and the end of the phrase. The highest level of analysis is the *intonational phrase*, which is made up of one or more intermediate phrases and ends with an additional high or low tone, the ***boundary tone***, which controls how the pitch contour ends. Together with the phrasal tone, an intonational phrase can end in four different ways, **H+H%**, **H+L%**, **L+H%**, and **L+L%**.

2.2.1 Intonation and Syntax

Many researchers have observed that intonational boundaries result in fragmented constituents (e.g., Gee and Grosjean, 1983). Researchers are addressing how parsing can interact with intonational boundaries. There seems to be a consensus that the intonational boundaries play an important part in disambiguating utterances, so can not be ignored. Price et al. (1991) found that hearers can resolve most syntactic ambiguous utterances based on prosodic information. Bear and Price (1990) explored how to make a parser use automatically extracted prosodic features to rule out extraneous parses. The prosodic information was represented as a numeric score between each pair of consecutive words, ranging from zero to five, depending on the amount of preboundary lengthening (normalized duration of the final consonants) and the pause duration between the words. Marcus and Hindle (1990) and Steedman (1990) also examined the role that intonational phrases play in parsing; but in their cases, how to represent the content of a phrase, which is often incomplete from a syntactic standpoint.

Not only is prosodic information available to help in resolving syntactic ambiguity, but as Beach demonstrated (1991), hearers also have the capacity to use it at an early point in the sentence. Beach used the beginning parts of sentences that were ambiguous between minimal attachment and non-minimal attachment. The hearers had to determine if the noun phrase following a verb was a direct object or the subject of a sentence complement. Beach recorded sentences with appropriate prosodic patterns in terms of the duration of the verb and the pitch on the noun phrase and had the hearers judge the initial parts of sentences, such as "Jay believed the gossip ...". Beach found that hearers could do the task, even when the noun phrase was removed. In a second experiment, Beach found that pitch and duration act like cue trading relations.

2.2.2 Intonation and Pragmatics

Intonation has long been thought to be related to pragmatics. Important work in this area was done by Halliday (1967), who proposed that intonation gives information about given versus new information. In fact, he claims, speakers break their thoughts into information units, which are realized phonologically by intonation.

Pierrehumbert and Hirschberg (1986; 1990) looked at the role that intonation plays in discourse interpretation. They claim that the choice of tune "[conveys] a particular relationship between an utterance, currently perceived beliefs of a hearer or hearers, ... and anticipated contributions of subsequent utterances ...[and] that these relationships are compositional —composed from the *pitch accents*, *phrase accents*, and *boundary tones* that make up tunes" (Pierrehumbert and Hircshberg, 1990, pg. 271). In their theory, pitch

accents contain information about the status of discourse referents, phrasal tones about the relatedness of intermediate phrases, and boundary tones about whether the phrase is "forward-looking" or not.

Intonation can also give information about discourse structure. Grosz and Hirschberg (Grosz and Hirschberg, 1992; Hirschberg and Grosz, 1992) investigated whether subjects could segment a dialogue based on Grosz and Sidner's theory of how discourse structure relates to the intentional and attentional state of the participants (1986). As their corpus, they used AP news stories read by a professional newscaster. They had a group of subjects segment the stories based on text alone, and another group use the spoken stories to segment. Subjects had to identify the local structural features of direct quotes and parentheticals, and global structural features such as the beginning of a new discourse segment, and the ending of a segment. They found considerable agreement between subjects in segmenting the dialogues. They also found that intonational features, namely pitch range of an utterance, pauses between utterances, and change in pitch range from one utterance to a next, correlate with the segmentation that the subjects produced. They found, however, that "the relationship between structure and intonational features is sometimes a complex one—a given discourse structural feature may be signaled by several intonational variables, which may or may not be independent" (Grosz and Hirschberg, 1992, pg. 432).

Nakajima and Allen (1992) also found that the intonation contour can be used to give information about the local discourse structure. In their work, they used a corpus of spoken dialogues, not text read by a trained speaker. They found that the onset frequency, final pitch, and pitch ratio (of first peak of current utterance to first peak of previous utterance) can be used to distinguish whether the utterance unit is a speech act continuation, a topic shift, topic continuation, or an elaboration.

Lastly, research has indicated that intonation plays a significant role in turn-taking. Ford and Thompson (1991) claim that intonational phrase endings almost always involve grammatical completion and semantic completion, and hence these units will be a major component of the turn-taking protocol. So, participants in a conversation can use intonational clues to determine where a new turn could start.

2.2.3 Detecting Intonational Boundaries

Detecting intonational boundaries is very difficult. Instead, most research work has examined how durational clues, such as preboundary lengthening and pausal durations, correlate with boundary types. Ostendorf et al. (1990) and Wightman et al. (1992) have found that if they normalize the duration of consonants following the last vowel in a word to take account of its normal duration, then these durations correlate with the type of intonational tone (no tone, phrasal tone, or boundary tone) that follows the word. However, this work has mostly concentrated on read speech, which is not plagued by speech repairs.

Another line of attack is to automatically classify intonational boundaries. Wang and Hirschberg (1992) have looked at predicting intonational phrases from textual information, such as category of the current word, category of the constituent being built, distance from last boundary, and the presence of a predicted accent. They used an automatic classifier to discover which types of information were most informative. This work is relevant for it

shows that there are textual clues that predict intonational boundaries and that they can be learned, and so can be used to augment acoustic clues.

2.3 Speech Repairs

Most of the current work in detecting and correcting speech repairs starts with the seminal work of Levelt (1983). Levelt was primarily interested in speech repairs as evidence for how people produce language and how they monitor it to ensure that it meets the goals it was intended for. From studying task-oriented monologues, Levelt put forth a number of claims. The first is that when a speaker notices a speech error, she will only interrupt the current word if it is in error. Second, repairs obey the following well-formedness rule (except those involving syntactically or phonologically ill-formed constructions). The concatenation of the text before the interruption point (with some completion to make it well formed) followed by the conjunction "and" followed by the text after the interruption point must be syntactic well-formed. For instance, "did you go right – go left" is a well-formed repair since "did you go right and go left" is syntactically well-formed; whereas "did you go right – you go left" is not since "did you go right and you go left" is not.

Third, Levelt hypothesized that listeners can use the following rules for determining the extent of the removed text (the continuation problem). If the last word before the interruption is of the same category as the word before, then delete the last word before the interruption. Otherwise, find the closest word prior to the interruption that is the same as the first word after the interruption. That word is the start of the text to be removed. Levelt found that this strategy would work for 50% of all repairs (including fresh starts), get 2% wrong, and have no comment for the remaining 48%.[1] In addition, Levelt showed that different editing terms make different predictions about whether a repair is a fresh start or not. For instance, that "uh" strongly signals an abridged or modification repair, whereas a word like "sorry" signals a fresh start.

2.3.1 Intonational Clues

One area that Levelt only briefly touched on was how prosody can be used in detecting and correcting speech repairs. In work done with Cutler (Levelt and Cutler, 1983), they showed that lexical stress often marks lexical repairs, repairs in which the speaker uttered the wrong word. They also suggested that intonation can play a bigger part in detecting a repair and determining its correction.

A number of other researchers have also explored the role that intonation plays in detecting speech repair. Lickley, Shillcock and Bard (1991) found that the hearer could perform the task using low-pass filtered speech, which removes segmental information, leaving what amounts to the intonation contour. From this, Lickley and Bard (1992) proposed that since people are often unaware of speech repairs, the speech processing mechanism must have very early access to them. They found that subjects were able to recognize speech repairs after (and not before) the onset of the first word following the interruption point and before they

[1] Levelt incorrectly calculated the percent that this algorithm gets correct to be 52%.

were able to recognize the word. This shows that listeners are able to use prosodic clues present across the interruption point without recourse to lexical or syntactic knowledge.

Other researchers have been more specific in terms of which intonational clues are useful. O'Shaughnessy (1992) suggests that duration and pitch can be used. Bear, Dowding, and Shriberg (1992) discuss acoustic clues for filtering potential repair patterns, for identifying potential cue words of a repair, and for identifying fragments. Nakatani and Hirschberg (1993) suggest that speech repairs can be detected by small but reliable differences in pitch and amplitude and by the length of pause at a potential interruption point. However, none of these results have been incorporated into a system that automatically detects speech repairs.

2.3.2 Computational Approaches

Previous computational work has explored a variety of approaches to detecting and correcting speech repairs. A way to compare the effectiveness of these approaches is to look at their recall and precision rates. For detecting repairs, the recall rate is the number of correctly detected repairs compared to the number of repairs, and the precision rate is the number of detected repairs compared to the number of detections (including false positives). But the true measures of success are the correction rates. Correction recall is the number of repairs that were properly corrected compared to the number of repairs. Correction precision is the number of repairs that were properly corrected compared to the total number of corrections.

One of the first computational approaches was by Hindle (1983), who addressed the problem of correcting self-repairs by adding rules to a deterministic parser that would remove the necessary text. Hindle assumed the presence of an edit signal that would mark the interruption point, and was able to achieve a recall rate of 97% in finding the correct repair. For modification repairs, Hindle used three rules for expuncting text. The first rule "is essentially a non-syntactic rule" that matches repetitions (of any length); the second matches repeated constituents, both complete; and the third, matches repeated constituents, in which the first is not complete, but the second is.

However, Hindle's results are difficult to translate into actual performance. First, his parsing strategy depends upon the "successful disambiguation of the syntactic categories." Although syntactic categories can be determined quite well by their local context (as is needed by a deterministic parser), Hindle admits that "[self-repair], by its nature, disrupts the local context." Second, Hindle's algorithm depends on the presence of an edit signal; so far, the abrupt cut-off that some have suggested signals the repair (cf. Labov, 1966) has been difficult to find, and it is unlikely to be represented as a binary feature (cf. Nakatani and Hirschberg, 1993).

The SRI group (Bear et al., 1992) employed simple pattern matching techniques for detecting and correcting modification repairs.[2] For detection, they were able to achieve a recall rate of 76%, and a precision of 62%, and they were able to find the correct repair 57%

[2]They referred to modification repairs as *nontrivial* repairs, and to abridged repairs as *trivial* repairs; however, these terms are misleading. Consider the utterance "send it back to Elmira uh to make OJ". Determining that the corrected text should be "send it back to Elmira to make OJ" rather than "send it back to make OJ" is not trivial.

of the time, leading to an overall correction recall of 43% and correction precision of 50%. They also tried combining syntactic and semantic knowledge in a "parser-firs" approach—first try to parse the input and if that fails, invoke repair strategies based on word patterns in the input. In a test set containing 26 repairs (Dowding et al., 1993), they obtained a detection recall rate of 42% and a precision of 84.6%; for correction, they obtained a recall rate of 30% and a recall rate of 62%.

Nakatani and Hirschberg (1993) investigated using acoustic information to detect the interruption point of speech repairs. In their corpus, 74% of all repairs are marked by a word fragment. Using hand-transcribed prosodic annotations, they trained a classifier on a 172 utterance training set to identify the interruption point (each utterance contained at least one repair). On a test set of 186 utterances containing 223 repairs, they obtained a recall rate of 83.4% and a precision of 93.9% in detecting speech repairs. The clues that they found relevant were duration of pause between words, presence of fragments, and lexical matching within a window of three words. However, they do not address the problem of determining the correction or distinguishing modification repairs from abridged repairs.

Young and Matessa (Young and Matessa, 1991) have also done work in this area. In their approach, speech repairs are corrected after a opportunistic case-frame parser analyzes the utterance. In fact the aim of their work is to detect and correct any misunderstandings that the parser makes, be them speech repairs, ungrammatical or ill-formed utterances, or even utterances that cannot be represented in the case-frame representation. Their system first looks for parts of the input utterance that were not used by the parser, and then uses semantic and pragmatic knowledge (of the limited domain) to correct the interpretation. For instance they use knowledge about slot-filler constraints and relations among slot fillers in order to make sense of the input that hasn't been resolved. For speech repairs, they make use of functional equivalence heuristics (Matessa and Young, 1994) and report correction rates between 95 and 99% of all fresh starts and modification repairs.

We feel that this approach addresses speech repairs too late in the process. There is good evidence that semantic knowledge is not needed to process speech repairs, and so when a less restricted setting is used, in which participants can carry on a collaborative dialogue, analysis at the semantic level might be too late to deal with the complexities. Their techniques will probably only work in limited domains as the syntactic and semantic analysis is very domain dependent.

2.4 Utterance Units

From our review of the literature, there does not seem to be any computational work in segmenting a turn into utterance units. Current computational approaches, such as employed in the ARPA data collection (MADCOW, 1992), allow the user to think off-line, with turn-taking "negotiated" by the user pressing a button when he wants to speak. More relevant research has been done in the fields of linguistics and psychology, in spoken language generation, and in developing spoken dialogue transcription schemes.

An interesting question concerning utterances is whether a listener is able to predict the end of an utterance. Although such a result should not be surprising, given theories

of intonation, it is not clear if intonational clues suffice. Using read speech, Grosjean (1983) took sentences in which the base part of the sentences could be continued by a prepositional phrase. He found that listeners at the potentially last word were able to determine if the utterance was over, and if not, how much remained. Grosjean also found that pitch, duration, and amplitude clues on the potentially last word correlated with the results.

In other work, Gee and Grosjean (1983) discuss *performance structures*, which are based on such experimental paradigms as measuring pauses between words, recall, and relatedness judgments. The structures that are derived are found to be relatively invariant across tasks, and are supposed to capture the psychological reality of the linguistic structure of an utterance. However, these structures do not always correspond well to syntactic structure, which tend to be right branching, rather than being more balanced, as what tends to happen with performance structures.

Gee and Grosjean put forward a model to account for the performance structures that arise. Their model builds up intermediate phrases and intonational phrases based on syntactic constituents. Every time a head of a phrase is encountered, unless it is a function word (such as a preposition), it is posited as ending the current intermediate phrase. They also have rules for building these intermediate phrases into intonational phrases. With this model, which works on local information in a fairly left to right manner, they are able to better account for performance data, as well as unify prosody with linguistic structure. They suggest the following.

> [The] prosodic structures reflect properties of the logical form of the sentence. This suggests the possibility that such structures are not just prosodic structures, but really a basic linguistic structure, perhaps the only one, or at least the critical one in processing (pg. 448).

Although Gee and Grosjean were not trying to explicitly account for spoken dialogue, this theory of performance structure, in terms of intonational phrases, helps solidify the relationship between utterance units and linguistic theory.

Using Gee and Grosjean's model of how linguistic content can be broken into intonational phrases, Bachenko and Fitzpatrick (1990) tried doing speech generation from text. They found that Gee and Grosjean's model falls short of providing a comprehensive theory, since their rules were too limited and the syntax too under-specified to achieve good coverage. However, it did provide them with the theoretical framework from which to proceed.

The next work of relevance is that of Nakajima and Allen (1992; 1993), which was done as part of the TRAINS project. As we mentioned above, they were interested in using prosody to determine discourse structure. Since they used spoken dialogues, they were forced to deal with the issue of defining utterance units (UU). They proposed the following principles.

Grammatical Principle: Place the UU boundary where a period could be put. In case of conjunction, the UU boundary is set just before the conjunction.

Pragmatic Principle: The UU should correspond to a basic speech-act. In other words, UU should represent the speaker's basic intention. Note that this does not rule out the case where one speech act continues over several UUs.

Conversational Principle: A UU boundary should be placed whenever the speaker changes. This includes the case of short acknowledgments such as *hnn-hnn* or *yes.*

Prosodic Principle: The UU boundary is placed whenever a medium length or longer pause occurs. The pause threshold is set to 750 msec which is a bit longer than the pauses called *search pauses* or *repair pauses.*

This formulation of utterance units only indirectly makes use of intonation, and is very subjective, especially for the pragmatic principle, which runs the risk of being circular.

In later work done in the TRAINS project, the segmentation into utterance units was changed to correspond "roughly" to intonational phrases (Gross, Allen, and Traum, 1993). The end of an utterance unit would be signaled by one of the following: a boundary tone, a pause in speech longer than a single beat, or a resetting of the pitch level, starting a new intonational phrase. The last option is presumably for fresh starts or when the speaker interrupts herself; thus the interruption point can signal an utterance boundary. The second option will also catch some of the modification speech repairs as well as pauses in the middle of an intonational phrase. Of course, this will not cause a problem in the TRAINS model, due to the presence of a `continuation` and a `repair` grounding move (Traum and Hinkelman, 1992).

2.5 Discourse Markers

Many researchers have noted the importance of discourse markers (Cohen, 1984; Reichman-Adar, 1984; Sidner, 1985; Grosz and Sidner, 1986; Litman and Allen, 1987). These markers serve to inform the reader about the structure of the discourse—how the current part relates to the rest. For instance, words such as "now", "anyways" signal a return from a digression. Words such as "firstly" and "secondly" signal that the speaker is giving a list of options. The structure of the text is also important because in most theories of discourse, it helps the listener resolve anaphoric references.

Spoken dialogue also employs a number of other discourse markers that are not as closely tied to the discourse structure. Words such as "mm-hm" and "okay" function as acknowledgments. Words such as "well", "like", "you know", "um", and "uh" can act as 'fillers'. Due to their lack of sentential content, and their relevance to the discourse process (including preventing someone from stealing the turn), they are also regarded as discourse markers.

Little has actually been done in identifying them in spoken dialogue. One exception is work done by Litman and Hirschberg (1990), who looked at how intonational information can distinguish between a discourse marker or sentential reading for a set of ambiguous lexical items. This work was based on hand-transcribed intonational features and only examined discourse markers that were only one word long. The corpus consisted of segments

from a radio call in show "The Harry Gross Show: Speaking of Your Money" (Pollack, Hirschberg, and Webber, 1982). They found that discourse uses of such words are either an intermediate phrase by themselves, or they are first in an intermediate phrase and either de-accented or have a **L*** accent. Sentential uses were either in the middle of a phrase or, if first, bore a **H*** accent. Litman and Hirschberg found that the hand-coded acoustics clues disambiguated the usage of non-conjunct discourse markers (other than "and"s and "or"s) 93.4% of the time, while a part-of-speech tagger correctly identified 75% of them.

2.6 Speech Actions

One of the big developments towards relating language usage to the mental state of agents has been to view language as actions (Grice, 1957; Austin, 1962; Searle, 1969). Speakers act by way of their utterances to accomplish various effects. The actions in this case are *speech actions*, and include such things as promising, informing, and requesting. Taking this approach has allowed utterances to be given an intentional account. The work on speech act analysis is relevant to this thesis proposal because the utterance units that we want to identify undoubtedly have an intentional account.

Not only can actions be analyzed by themselves, but these actions can also be used in a plan-based approach. This idea was first suggested by Bruce (1975), and formalized by Cohen and Perrault (1979), who developed a system that takes an agent's goal and finds a speech action that will accomplish it, and by Allen and Perrault (1980), who, given an observed action, try to determine the agent's goal. This work was based on the STRIPS planning formalism (Fikes and Nilsson, 1971), in which actions are represented by a *precondition list*, an *add list*, and a *delete list*. The preconditions state the propositions that must be true for an action to be executed, the add list gives the propositions that will be true after the action has been executed, and the delete list gives the propositions that will not hold. The STRIPS formalism was expanded by Sacerdoti (1975) so that actions include a *decomposition list*, which allows actions to be built from more primitive actions. Although STRIPS suffers from serious deficits (Allen, 1990), its simplicity has attracted many followers. Below we give an example of one of Allen and Perrault's speech actions.

`INFORM(?speaker,?hearer,?prop)`
prec: ?speaker KNOW ?prop
effect: ?hearer KNOW prop
body: hearer BELIEVE speaker WANT hearer KNOW prop

The work of Cohen, Allen, and Perrault has been extended to account for stretches of speech longer than a single utterance by finding a complete plan, and for dealing with aspects beyond informs and requests.

Litman and Allen's work (1987) focused on showing how clarifications can be handled in a plan-based formalism. Just as speakers can execute domain actions, they can also execute discourse actions, such as `identify-parameter`. Plan recognition starts with a surface speech action, which is then chained to a discourse action, through decomposition.

The discourse action in turn might refer to another plan, such as referring to an action in a domain plan. This will cause the appropriate action to be added to an existing domain plan (through plan elaboration) or will cause a new plan to be created if the appropriate plan doesn't exist. By doing this plan reasoning, Litman and Allen are able to make sense of such things as clarification dialogues, and changes in topic. They also employed linguistic knowledge about discourse markers to help guide the plan recognition problem.

The thrust of Traum's work (Traum, 1991; Traum and Hinkelman, 1992) is on providing a computation model of grounding. Grounding is the process where conversational participants add to the common ground of a conversation (Clark and Schaefer, 1989; Clark and Brennan, 1990). Traum models the grounding process by proposing that utterances move through a number of states, pushed by grounding acts, which include initiate, continue, repair, request repair, acknowledge, and request acknowledge. Once an utterance has been acknowledged, it will reside in mutual belief as a proposal of the person who initiated it. By modeling the grounding process, Traum can better deal with how participants in conversation collaborate, and can account for the abundance of acknowledges that tend to occur.

Heeman and Hirst (Heeman and Hirst, 1992; Hirst et al., 1994) have also done work based on the work of Clark and his colleagues (Clark and Wilkes-Gibbs, 1986; Clark and Schaefer, 1989). In this case, it was in providing a computational model of how participants collaborate in accomplishing the linguistic task of referring. Their speech acts include specific acts for referring and actions for collaborating in building a plan for the referring expression, such as `s-accept` for accepting a plan, `s-postpone` for postponing a judgment until more is added, `s-reject` for rejecting a part of a plan, and `s-actions` for adding additional actions to a plan. With these, Heeman and Hirst are able to account for the fragments that tend to occur in such dialogues in terms of speech actions related to the collaborative task. Although based on similar psycholinguistic models, this work differs from Traum's in that Heeman and Hirst assume that being in a collaborative state sanctions the acceptance of actions that further the collaborative task.

2.7 Conversation Analysis

Conversation Analysis is a subfield of sociology that is concerned with the study of actual dialogues in order to discover recurring patterns.[3] Work done in this field is starting to enter into computation approaches to natural language (e.g., Traum and Allen, 1994; McRoy, 1993), for it allows expectations and social norms to be used in understanding a speaker's utterance, thus short-circuiting means-end reasoning about the other participant's top-level goals and beliefs, and in deciding what to do next. The contributions made in the field of conversation analysis are also of interest in this proposal, for they shed light on the notion of utterance units.

[3] A good introduction to this field is contained in Levinson (1983) and in Fox (1987).

2.7.1 Turn-Taking

One of the most obvious features of dialogue is that for the most part only one person is talking at a time. In fact, conversation can be divided into turns. But how do conversants co-ordinate how long each will speak for and who will speak next. Sacks, Schegloff and Jefferson (1974) propose that there is a turn-taking protocol, and they give rules governing it. The basic unit is the *turn-constructional unit*, and these units are the lowest element at which a change in speaker can occur. These units can be sentential, clausal, phrasal, and lexical constructions (pg. 702). After each turn-constructional unit, there is a *transition relevance place*, where either the current speaker can select the next speaker, as in asking a question to a particular person, or the next speaker can self-select.

2.7.2 Adjacency Pairs

Adjacency pairs are pairs of consecutive utterances produced by different speakers. A speaker's utterance is often influenced by what the other conversant just said. In fact, Schegloff and Sacks (1973) claim that there is a preferred response for each utterance type and a less preferred second. For instance, the preferred response to a question is the answer. In fact the first part makes the second part conditionally relevant. If the recipient does not respond with the preferred response, but instead with the less-preferred second, then this response should be marked.

2.7.3 Exchanges

Along the lines of adjacency pairs are exchanges. Here, rather than concentrate on pairs of utterances across a change in speaker, we are interested in finding a unit of analysis that we can break a dialogue into. From theorizing about the function of intonation in discourse, Coulthard and Brazil (1981) propose that dialogue is made up of *transactions*, and that these units are the "largest structural unit of discourse" (pg. 88). Transactions are made up of *exchange* structures, with each part being marked intonationally. For instance an eliciting exchange would have a structure along the lines of the following: initiation, response, followup.

2.7.4 Repairs

The fourth focus of conversation analysis deals with repairs. Repairs, in this case, are any attempts to fix previous utterances because they were insufficient for what they were intended. Repairs can be initiated (noticed) by the speaker or by the other participant, they can be made either, and they can happen in the same turn as the utterance or later on. Schegloff, Jefferson, and Sacks (1977) propose that self-initiated, self-repairs in the same turn are most preferred, and least preferred are other-initiated, other-repairs.

Chapter 3

Completed Work

In this section, we present the work that has been completed as part of this thesis proposal. This work provides a lot of the foundation which the thesis is based on. The first section outlines the work done on collecting a corpus of spoken dialogues, and tools and standards for transcribing the events that we are interested in. The second section discusses our work on detecting and correcting two types of speech repair—*modification* repairs and *abridged* repairs.

3.1 Dialogue Corpus

As part of the TRAINS project (Allen and Schubert, 1991), which is a long term research project to build a conversationally proficient planning assistant, we are collecting a corpus of problem solving dialogues. The dialogues involve two participants, one who is playing the role of a user and has a certain task to accomplish, and another, who is playing the role of the system by acting as a planning assistant. The first set of dialogues were collected in 1991 by Gross, Allen, and Traum (1993), and they discuss the manner in which this was done and provide transcriptions. In 1992 and 1993, more dialogues were collected, this time using the Waves software (Entropic Research Laboratories), which allowed time-aligned word transcriptions to be produced. We describe the manner in which these dialogues were collected in a technical note (Heeman and Allen, 1994e). These dialogues, along with the word transcriptions, will be available on CD-ROM. In all, the entire corpus consists of 112 dialogues totaling almost eight hours in length and containing about 62,000 words and 6300 speaker turns.

3.1.1 Dialogue Tools

Although there are tools and standards for annotating single speaker spoken utterances, such as the WAVES software (Talkin, 1989), and the ToBI annotation scheme (Silverman et al., 1992; Beckman and Hirschberg, 1994; Beckman and Ayers, 1994), such resources do not exist for dialogues. So, we have built a toolkit and guidelines for bridging the gap between dialogues and single speaker utterances (Heeman and Allen, 1994d). The toolkit

addresses the problem of setting up the initial dialogue audio file, obtaining a breakup of the dialogue into single speaker utterances files, printing the contents of a dialogue, and in updating the breakup. The guidelines give rules for segmenting the dialogue into utterance files, so that local phenomena, such as speech repairs, are not separated.

3.1.2 Annotated Speech Repairs

The second aspect of the corpus work has focused on developing an annotation scheme for speech repairs (Heeman and Allen, 1994c). This work is based on the work done by Bear et al. (1993), but extends it to better deal with overlapping repairs. The annotation scheme allows the user to label the word correspondences that exist between the removed text and the resumed text. We annotate this using the labels **m** for word matching and **r** for word replacements (words of the same syntactic category). Each pair is given a unique index. Other words in the removed text and resumed text are annotated with an **x**. Also, editing terms (filled pauses and clue words) are labeled with **et**, and the moment of interruption with **int**, which will occur before any editing terms associated with the repair, and after the fragment, if present. The annotation scheme also allows the interruption point to be indexed and the word labels to refer to its interruption point, thus allowing overlapping repairs to be annotated. Below is a sample annotation, with removed text "go to oran-", editing term "um", and resumed text "go to" (d93-14.2 utt60).

```
go| to|  oran-| um| go| to| Corning
m1| m2| x| int| et| m1| m2|
```

A speech repair can also be characterized by its *repair pattern*, which is a string that consists of the repair labels (word fragments are labeled as **-**, the interruption point by a period, and editing terms by **e**). The repair pattern for the example is **mm-.emm**.

3.2 Detecting and Correcting Speech Repairs

In our corpus of problem solving dialogues, 25% of turns contain at least one repair, 67% of repairs occur with at least one other repair in the turn, and repairs in the same turn occur on average within 6 words of each other. As a result, no spoken language system will perform well without an effective way to detect and correct speech repairs. In the following, we present research that has been reported elsewhere (Heeman and Allen, 1994b; Heeman and Allen, 1994a), in which we address the problem of detecting and correcting modification and abridged speech repairs using local clues. This work is empirically based and we use the TRAINS dialogue corpus for training and test data.

3.2.1 Repair Indicators

In order to correct speech repairs, we first need to detect them. If we were using prosodic information, we could focus on the actual interruption point (cf. Nakatani and Hirschberg,

	Total	with Fragment	with Editing Term
Modification Repair	450	14.7%	19.3%
Word Repetition	179	16.2%	16.2%
Larger Repetition	58	17.2%	19.0%
Word Replacement	72	4.2%	13.9%
Other	141	17.0%	26.2%
Abridged Repair	267	46.4%	54.3%
Total	717	26.5%	32.4%

Table 3.1: Occurrence of Types of Repairs

1993); however, our initial work was restricted to lexical clues, and so we need to be more lenient.

Table 3.1 gives a breakdown of the modification speech repairs and the abridged repairs in the training set, based on hand-annotations of the repairs.[1] Modification repairs are broken down into four groups, single word repetitions, multiple word repetitions, one word replacing another, and others. In the second and third columns, we give the percentage of the repairs that include fragments and editing terms.

This table shows that strictly looking for the presence of fragments and editing terms will miss at least 41% of speech repairs. So, we need to look at word correspondences in order to get better coverage for detecting repairs. In order to keep the false positive rate down, we restrict ourselves to the following types of word correspondences: (1) word matching with at most three intervening words, denoted by **m-m**; (2) two adjacent words matching two others with at most 6 words intervening, denoted by **mm–mm**; and (3) adjacent replacement, denoted by **rr**. Table 3.2 gives the number of repairs in the training corpus that can be detected by each clue, based on the hand-annotations. For each clue, we give the number of repairs that it will detect in the first column. In the next three columns, we give a breakdown of these numbers in terms of how many clues apply. As the table shows, most repairs are signal by only one of the 3 clues.

Although the **m–m** clue and **mm–mm** clue do not precisely locate the interruption point of a repair, we can, none the less, detect them; in fact, we can detect 97.7% (708/725) of all the repairs. But, we still will have a problem with false positives, and detecting the extent of the repair.

3.2.2 Determining the Correction

Based on the work done at SRI (Bear, Dowding, and Shriberg, 1992), we next looked at the speech repair patterns in our annotated training corpus. If we can automatically determine the pattern, then the deletion of the removed text along with the editing terms gives the

[1]Eight repairs were excluded from this analysis. These repairs could not be automatically separated from other repairs that overlapped with them.

	Total		1 clue	2 clues	3 clues
Fragment	190		127	58	5
Editing Terms	232		164	63	5
m-m	331	412	296	111	5
mm–mm	94				
rr	59				
others	9		n.a.	n.a.	n.a.
Total	717		587	116	5

Table 3.2: Repair Indicators

correction. Since the length of the pattern can be quite long, especially when editing terms and word fragments are added in, the number of possible templates becomes very large. In our training corpus of 450 modification repairs, we found 72 different patterns (not including variations due to editing terms and fragments). All patterns with at least 2 occurrences are listed in Table 3.3.

m.m	179	mmx.mm	4
r.r	72	mrm.mrm	3
mm.mm	41	mmmr.mmmr	3
mr.mr	17	mm.mxm	3
mx.m	15	r.xr	2
mmm.mmm	14	mxxx.m	2
rm.rm	12	mx.mx	2
m.xm	6	mmrm.mmrm	2
mmr.mmr	5	mmmx.mmm	2
m.xxm	5	mmmm.mmmm	2
x.xx	4	m.mx	2
x.	4		

Table 3.3: Repair Patterns and Occurrences

3.2.2.1 Rules for Adding to the Pattern

Rather than doing template matching, we build the repair pattern on the fly. When a possible repair is detected, the detection itself puts constraints on the repair pattern. For instance, if we detect a word fragment, the location of the fragment limits the extent of the editing terms. It also limits the extent of the resumed text and removed text, and so restricts word correspondences that can be part of the repair.

In this section, we present the rules we use for building repair patterns. These rules not

only limit the search space, but more importantly, are intended to keep the number of false positives as low as possible, by capturing a notion of 'well-formedness' for speech repairs.

The four rules listed below follow from the model of repairs that we presented in the introduction. They capture how a repair is made up of three intervals—the removed text, which can end in a word fragment, possible editing terms, and the resumed text—and how the interruption point follows the removed text and precedes the editing terms.

1. Editing terms must be adjacent.
2. Editing terms must immediately follow the interruption point.
3. A fragment, if present, must immediately precede the interruption point.
4. Word correspondences must straddle the interruption point and can not be marked on a word labeled as an editing term or fragment.

The above rules alone do not restrict the possible word correspondences enough. Based on an analysis of the hand-coded repairs in the training corpus, we propose the following additional rules.

Rule (5) captures the regularity that word correspondences of a modification repair are rarely, if ever, embedded in each other. Consider the following exception.

how would that – how long that would take

In this example, the word correspondence involving "that" is embedded inside of the correspondence on "would". The speaker actually made a uncorrected speech error (and so not a speech repair) in the resumed text, for he should have said "how long would that take." Without this ungrammaticality, the two correspondences would not have been embedded, and so would not be in conflict with the following rule.

5. Word correspondences must be cross-serial; a word correspondence cannot be embedded inside of another correspondence.

The next rule is used to limit the application of word correspondences when no correspondences are yet in the repair pattern. In this case, the repair would have been detected by the presence of a fragment or editing terms. This rule is intended to prevent spurious word correspondences from being added to the repair. For instance in the following example, the correspondence between the two instances of "I" is spurious, since the second "I" in fact replaces "we".

I think we need to uh I need

So, when no correspondences are yet included in the repair, the number of intervening words needs to be limited. From our test corpus, we have found that 3 intervening words, excluding fragments and editing terms is sufficient.

6. If there are no other word correspondences, there can only be 3 intervening words, excluding fragments and editing terms, between the first part and the second part of the correspondence.

The next two rules restrict the distance between two word correspondences. Figure 3.1 shows the distance between two word correspondences, indexed by i and j. The intervals x and y are the sequences of words that occur between the marked words in the removed text and in the resumed text, respectively. The word correspondences of interest are those that are adjacent, in order words, the ones that have no labeled words in the x and y intervals.

$$\mathbf{m_i} \underbrace{\cdots}_{x} \mathbf{m_j} \cdots \mathbf{int} \cdots \mathbf{m_i} \underbrace{\cdots}_{y} \mathbf{m_j}$$

Figure 3.1: Distance between correspondences

For two adjacent word correspondences, Rule (7) ensures that there is at most 4 intervening words in the removed text, and Rule (8) ensures that there are at most 4 intervening words in the resumed text.

7. In the removed text, two adjacent matches can have at most 4 intervening words ($|x| \leq 4$).
8. In the resumed text, two adjacent matches can have at most 4 intervening words ($|y| \leq 4$).

The next rule, Rule (9), is used to capture the regularity that words are rarely dropped from the removed text, instead they tend to be replaced.

9. For two adjacent matches, the number of intervening words in the removed text can be at most one more than the number of intervening words in the resumed text ($|x| \leq |y| + 1$).

The last rule, Rule (10), is used to restrict word replacements. From an analysis of our corpus, we found that word replacement correspondences are rarely isolated from other word correspondences.

10. A word replacement must either only have fragments and editing terms between the two words that it marks, or there must be a word correspondence in which there are no intervening words in either the removed text or the resumed text ($x = y = 0$).

3.2.2.2 An Example

To illustrate the above set of well-formedness constraints on repair patterns, consider the example given above, "I think we need to – uh I need." The detection clues will mark the word "uh" as being a possible editing term, giving the partial pattern given below.

```
I think we need to uh| I need
                     et|
```

Now let's consider the two instances of "I". Adding this correspondence to the repair pattern will violate Rule (6), since there are four intervening words, excluding the editing terms. The correspondence between the two instances of 'need' is acceptable though, since it straddles the editing term and there are only two intervening words between the corresponding words, excluding editing terms.

Even with the correspondence between the two instances of 'need', the matching between the 'I's still cannot be added. There are 2 intervening words between "I" and "need" in the removed text, but none in the resumed side, so this correspondence violates Rule (9). The word replacement of "we" by the second instance of "I", does not violate any of the rules, including Rule (10), so it is added, resulting in the following labeling.

```
I think we| need| to uh| I| need|
          r|    m|    et| r|    m|
```

3.2.2.3 Algorithm

The algorithm for labeling potential repair patterns encodes the assumption that speech repairs can be processed one at a time. The algorithm runs in lockstep with a part-of-speech tagger (Church, 1988), which is used for deciding possible word replacements. Words are fed in one at a time. The detection clues are checked first. If one of them succeeds, and there is not a repair being processed, then a new repair pattern is started. Otherwise, if the clue is consistent with the current repair pattern, then the pattern is updated; otherwise, the current one is sent off to be judged, and a new repair pattern is started.

When a new repair is started, a search is made to see if any of the text can contribute word correspondences to the repair. Likewise, if there is currently a repair being built, a search is made to see if there is a suitable word correspondence for the current word. Anytime a correspondence is found, a search is made for any additional correspondences that it might sanction.

Since there might be a conflict between two possible correspondences that can be added to a labeling, the one that involves the most recent pair of words is preferred. For instance, in the example above, the correspondence between the second instance of "I" and "we" is preferred over the correspondence between the second instance of "I" and the first.

The last issue to account for is the judging of a potential repair. If the labeling consists of just cue phrases, then it is judged as not being a repair.[2] Otherwise, if the interruption point of the potential repair is uniquely determined, then it is taken as a repair. This will be the case if there is at least one editing term, a word fragment, or there are no unaccounted for words between the last removed text part of the last correspondence and the resumed text part of the first correspondence.

[2]This prevents phrases such as "I guess" from being marked as editing terms when they have a sentential meanings, as in "I guess we should load the oranges."

3.2.2.4 Results of Pattern Building

The input to the algorithm is the word transcriptions, augmented with turn-taking markers. Since we are not trying to account for fresh starts, break points are put in to denote the cancel, and its editing terms are deleted (this is done to prevent the algorithm from trying to annotate the fresh start as a repair). The speech is not marked with any intonational information, nor is any form of punctuation inserted. The results are given in Table 3.4.

	Training Set	Test Set
Detection Recall	94.9%	91.5%
Detection Precision	55.8%	45.3%
Correction Recall	89.2%	85.9%
Correction Precision	52.4%	42.5%

Table 3.4: Results of Pattern Matching

The pattern builder gives many false positives in detecting speech repairs due to word correspondences in fluent speech being mis-interpreted as evidence of a modification repair. Also, in correcting the repairs, word correspondences across an abridged repair cause the abridged repair to be interpreted as a modification repair, thus lowering the correction recall rate.[3] For example, the following abridged repair has two spurious word correspondences, between "need to" and "manage to".

we need to – um manage to get the bananas to Dansville more quickly

These spurious word correspondences will cause the pattern builder to hypothesize that this is a modification repair, and so propose the wrong correction.

3.2.3 Statistical Filter

We make use of a part-of-speech tagger to not only determine part-of-speech categories (used for deciding possible word replacements), but also to judge modification repairs that are proposed by the pattern builder. For modification repairs, the category transition probabilities from the last word of the removed text to the first word of the resumed text have a different distribution than category transitions for fluent speech. So, by giving these distributions to the part-of-speech tagger (obtained from our test corpus), the tagger can decide if a transition signals a modification repair or not.

Part-of-speech tagging is the process of assigning to a word the category that is most probable given the sentential context (Church, 1988). The sentential context is typically approximated by only a set number of previous categories, usually one or two. Good part-of-speech results can be obtained using only the preceding category (Weischedel et al., 1993),

[3] About half of the difference between the detection recall rate and the correction recall rate is due to abridged repairs being misclassified as modification repairs.

which is what we will be using. In this case, the number of states of the Markov model is N, where N is the number of tags. By using the Viterbi algorithm, the part-of-speech tags that lead to the maximum probability path can be found in linear time with respect to the number of words to be tagged.

Figure 3.2 gives a simplified view of a Markov model for part-of-speech tagging, where C_i is a possible category for the ith word, w_i, and C_{i+1} is a possible category for word w_{i+1}. The category transition probability is simply the probability of category C_{i+1} following category C_i, which is written as $P(C_{i+1}|C_i)$, and the probability of word w_{i+1} given category C_{i+1} is $P(w_{i+1}|C_{i+1})$. The category assignment that maximizes the product of these probabilities is taken to be the best category assignment.

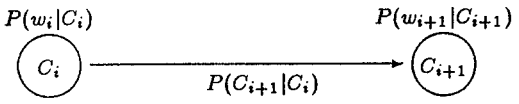

Figure 3.2: Markov Model of Part-of-Speech Tagging

3.2.3.1 A Simple Model of Speech Repairs

Modification repairs are often accompanied by a syntactic anomaly across the interruption point. Consider the following example, "so it takes two hours to go to – from Elmira to Corning" (d93-17.4 utt57), which contains a "to" followed by a "from". Both should be classified as prepositions, but the event of a preposition followed by another preposition is very rare in well-formed speech, so there is a good chance that one of the prepositions might get erroneously tagged as some other part of speech. Since the category transitions across interruption points tend to be rare events in fluent speech, we simply give the tagger the category transition probabilities around interruption points of modification repairs. By keeping track of when this information is used, we not only have a way of detecting modification repairs, but part-of-speech tagging is also improved.

To incorporate knowledge about modification repairs, we let R_i be a variable that indicates whether the transition from word w_i to w_{i+1} contains the interruption point of a modification repair. Rather than tag each word, w_i, with just a category, C_i, we tag it with $R_{i-1}C_i$: the complex tag consisting of the category and the presence of a modification repair.[4] This effectively multiplies the size of the tagset by two. From Figure 3.2, we see that we now need the following probabilities, $P(R_iC_{i+1}|R_{i-1}C_i)$ and $P(w_i|R_{i-1}C_i)$.

To keep the model simple, and ease problems with sparse data, we make several independence assumptions.

(1) Given the category of a word, a repair before it is independent of the word. (R_{i-1} and w_i are independent, given C_i.) So $P(w_i|R_{i-1}C_i) = P(w_i|C_i)$.

[4]Changing each tag to C_iR_i would result in the same model. We chose $R_{i-1}C_i$ since it simplifies the development of the model that we propose in Figure 3.3.

(2) Given the category of a word, a repair before that word is independent of a repair following it and the category of the next word. (R_{i-1} is independent of R_iC_{i+1}, given C_i.) So $P(R_iC_{i+1}|R_{i-1}C_i) = P(R_iC_{i+1}|C_i)$.

Another manipulation that we can do is use the definition of conditional probabilities to rewrite $P(R_iC_{i+1}|C_i)$ as $P(R_i|C_i) * P(C_{i+1}|C_iR_i)$. This manipulation allows us to view the problem as tagging null tokens between words as either the interruption point of a modification repair, $R_i = \tau_i$, or as fluent speech, $R_i = \phi_i$. The resulting Markov model is shown in Figure 3.3. Note that the context for category C_{i+1} is both C_i and R_i. So, R_i depends (indirectly) on the joint context of C_i and C_{i+1}, thus allowing syntactic anomalies to be detected.[5]

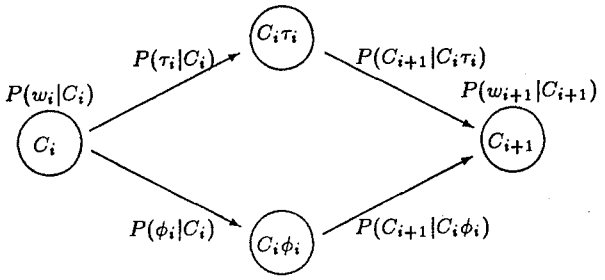

Figure 3.3: Markov Model of Repairs

Table 3.6 gives results for this simple model running on our training corpus. In order to remove effects due to editing terms and word fragments, we temporarily eliminate them from the corpus. Also, for fresh starts and change-of-turn, the algorithm is reset, as if they marked the end of a sentence. To eliminate problems due to overlapping repairs, we include only data points in which the next word is not intended to be removed (based on our hand annotations). This gives us a total of 19587 data points with 384 modification repairs, of which the statistical model found 169 of them, and a further 204 false positives. This gives us a recall rate of 44.2% and a precision of 45.3%. In the test corpus, there are 98 modification repairs, of which the model found 30, and a further 23 false positives; this gives a recall rate of 30.6% and a precision rate of 56.6%.

From Table 3.1, we can see that the recall rate of fragments as a predictor of a modification repair is 14.7% and their precision is 34.7%.[6] So, the method of statistically tagging modification repairs has more predictive power, and so should be used as a clue for detecting them. Furthermore, this method is doing something more powerful than just detecting word repetitions or category repetitions. Of the 169 repairs that it found, 109 were word repetitions and an additional 28 were category repetitions. So, 32 of the repairs that were found were from less obvious syntactic anomalies.

[5]Probabilities for fluent transitions are from the Brown corpus and probabilities for repair transitions are from the training data.

[6]The precision rate was calculated by taking the number of fragments in a modification repair (450*14.7%) over the total number of fragments ($450 * 14.7\% + 267 * 46.4\%$).

3.2.3.2 Adding Additional Clues

In the preceding section we built a model for detecting modification repairs by simply using category transitions. However, there are other sources of information that can be exploited, such as the presence of fragments, editing terms, and word matchings. The problem is that these clues do not always signal a modification repair. For instance, a fragment is twice as likely to be part of an abridged repair than it is to be part of a modification repair. One way to exploit these clues is to try to *learn* how to combine them, using a technique such as CART (Brieman, Friedman, and Olshen, 1984). However, a more intuitive approach is to adjust the transition probabilities for a modification repair to better reflect the more specific information that is known. Thus, we combine the information such that the individual pieces do not have to give a 'yes' or a 'no' answer, but rather, all can contribute to the decision.

Fragments: Assuming that fragments can be detected automatically (cf. Nakatani and Hirschberg, 1993), the question arises as to what the tagger should do with them. If the tagger treats them as lexical items, the words on either side of the fragment will be separated. This will cause two problems. First, if the fragment is part of an abridged repair, category assignment to these words will be hindered. Second, and more important to our work, is that the fragment will prevent the statistical model from judging the syntactic well-formedness of the word before the fragment and the word after, preventing it from distinguishing a modification repair from an abridged repair. So, the tagger needs to skip over fragments. However, the fragment can be viewed as the "word" that gets tagged as a modification repair or not. (The 'not' in this case means that the fragment is part of an abridged repair.) When no fragment is present between words, we view the interval as a null word. So, we augment the model pictured in Figure 3.3 with the probability of the presence of a fragment, F_i, given the presence of a repair, R_i, as is pictured in Figure 3.4. Since there are two alternatives for F_i—a fragment, f_i, or not, $\overline{f}_i$—and two alternatives for R_i—a repair or not, we need four statistics.

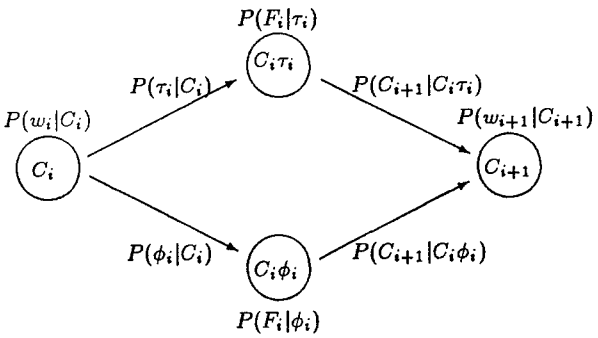

Figure 3.4: Incorporating Fragments

From our training corpus, we have found that if a fragment is present, a modification repair is favored—$P(f_i|\tau_i)/P(f_i|\phi_i)$—by a factor of 28.9. If a fragment is not present, fluent speech is favored—$P(\overline{f}_i|\phi_i)/P(\overline{f}_i|\tau_i)$, by a factor of 1.17.

Editing Terms: Editing terms, like fragments, give evidence as to the presence of a modification repair. So, we incorporate them into the statistical model by viewing them as part of the "word" that gets tagged with R_i, thus changing the probability on the repair state from $P(F_i|R_i)$ to $P(F_iE_i|R_i)$, where E_i indicates the presence of editing terms. To simplify the probabilities, and reduce problems due to sparse data, we make the following independence assumption.

(3) Given that there is a modification repair, the presence of a fragment or editing terms is independent. (F_i and E_i are independent, given R_i.) So $P(F_iE_i|R_i) = P(F_i|R_i) * P(E_i|R_i)$.

An additional complexity is that different editing terms do not have the same predictive power. So far we have investigated "um" and "uh". The presence of an "um" favors a repair by a factor of 2.7, while for "uh" it is favored by a factor of 9.4. If no editing term is present, fluent speech is favored by a factor of 1.2.

Word Matchings: In a modification repair, there is often a correspondence between the text that must be removed and the text that follows the interruption point. The simplest type of correspondence is word matchings. In fact, in our test corpus, 80% of modification repairs have at least one matching. This information can be incorporated into the statistical model in the same way that editing terms and fragments are handled. So, we change the probability of the repair state to be $P(F_iE_iM_i|R_i)$, where M_i indicates a word matching. Again, we assume that the clues are independent of each other, allowing us to treat this clue separately from the others.

Just as with editing terms, not all matches make the same predictions about the occurrence of a modification repair. Bear, Dowding, and Shriberg (1992) looked at the number of matching words versus the number of intervening words. However, this ignores the category of the word matches. For instance, a matching verb (with some intervening words) is more likely to indicate a repair than say a matching preposition or determiner. So, we classify word matchings by category and number of intervening words. Furthermore, if there are multiple matches in a repair, we only use one, the one that most predicts a repair. For instance in the following repair, the matching instances of "take" would be used over the matching instances of "will", since main verbs were found to more strongly signal a modification repair than do modals (d93-18.3 utt43).

how long will that take – will it take for engine one at Dansville

Since the statistical model only uses one matching per repair, the same is done in collecting the statistics. So, our collection involves two steps. In the first, we collect statistics on all word matches, and in the second, for each repair, we count only the matching that most strongly signals the repair. Table 3.5 gives a partial list of how much each matching favors a repair broken down by category and number of intervening words. Entries that are marked with "–" do not contain any datapoints and entries that are blank are below the baseline rate of 0.209, the rate at which a modification repair is favored (or actually disfavored) when there is no matching at all.

Cat	Number of Intervening Words					
	0	1	2	3	4	5
DT	935.5	38.5	2.7	2.2	0.7	0.8
IN	–	171.7	59.6	22.9	10.4	6.3
IS	490.0	55.8	5.9	3.2		
MD	–	6706.5	199.8	37.1	12.4	2.4
NN	–	68.0	32.2	10.4	0.3	0.2
NNP	144.3	9.2	6.2	6.7	3.3	2.8
PREP	16433.6	2.8				
PRP	8242.3	15.2	2.9	1.2	0.5	
RB	25.2	19.4	6.9	6.4	3.9	3.6
TO	5170.7	1.6	0.5	0.4		
VB	5170.6	216.3	71.5	31.2	18.1	7.0

Table 3.5: Factor by which a repair is favored

The problem with using word matching is that it depends on identifying the removed text and its correspondences to the text that follows the interruption point. However, a good estimate can be obtained by considering all word matchings with at most eight intervening words. This is in fact what we did for obtaining the results reported in Table 3.6, which is discussed next.

3.2.3.3 Results of Statistical Model

Table 3.6 summarizes the results of incorporating additional clues into the Markov model. The first column gives the results without any clues, the second with fragments, the third with editing terms, the fourth with word matches, and the fifth, with all of these clues incorporated. Of the 384 modification repairs in the training corpus, the full model predicts 305 of them versus 169 by the simple model. As for the false positives, the full model incorrectly predicted 207 versus the simple model at 204. So, we see that by incorporating additional clues, the statistical model can better identify modification repairs.

	Simple Model	Frag-ments	Edit Terms	Word Match	Full
Training:					
Recall	44.0%	50.0%	45.1%	76.5%	79.4%
Precision	45.3%	47.8%	46.5%	54.9%	59.6%
Testing:					
Recall	30.6%	43.9%	32.7%	74.5%	76.5%
Precision	56.6%	62.3%	59.3%	58.4%	62.0%

Table 3.6: Results of Markov Models

3.2.4 Overall Results

As we mentioned in Section 3.2.2.4, the pattern builder on its own gives many false positives. These are due to word correspondences in fluent speech being mis-interpreted as evidence of a modification repair, and due to word correspondences across an abridged repair causing the abridged repair to be interpreted as a modification repair. This results in a correction recall rate of 86% on the test corpus, but a precision rate of 43%. However, these rates can be improved by coupling the pattern builder with the statistical model, which will eliminate most of the false positives.

Potential repairs found by the pattern builder are divided into two groups. The first includes abridged repairs, and modification repairs involving only word repetitions. These are classified as repairs outright. The rest of the modification repairs are judged by the statistical model. Any potential repair that it rejects, but which contains a word fragment or filled pause is accepted as an abridged repair. Table 3.7 gives the results of the combined approach on the training and test sets.

	Training Corpus	Test Corpus
Detection		
Recall	91%	83%
Precision	96%	89%
Correction		
Recall	88%	80%
Precision	93%	86%

Table 3.7: Overall Results

Comparing our results to others that have been reported in the literature must be done with caution. Such a comparison is limited due to differences in both the type of repairs that are being studied and in the datasets used for drawing results. Bear, Dowding, and Shriberg (1992) use the ATIS corpus, which is a collection of queries made to an automated airline reservation system. As stated earlier, they removed all utterances that contained abridged repairs. Using a technique based on simple pattern matching, they obtained a detection recall rate of 76% and a precision of 62%, and for correction, a recall rate of 43% and a precision of 50%. It is not clear whether their results would be better or worse if abridged repairs were included. Dowding et al. (1993) used a similar setup for their data. They employed a "parser-first" strategy; if the utterance could not be syntactically and semantically interpreted as is, they looked for the simple patterns of Bear, Dowding, and Shriberg. As part of a complete system, they obtained a detection recall rate of 42% and a precision of 85%; and for correction, a recall rate of 30% and a precision of 62%. Lastly, Nakatani and Hirschberg (1993) also used the ATIS corpus, but focused only on detection. They used a classifier that they trained on hand-coded acoustic and textual clues. Their test corpus consisted entirely of utterances that contained at least one repair. This makes it hard to evaluate their results, reporting a detection recall rate of 83% and precision of 94%. Testing on an entire corpus would clearly decrease their precision. As for our own

data, we used a corpus of natural dialogues that were segmented only by speaker turns, not by individual utterances, and we focused on modification repairs and abridged repairs, with fresh starts being marked in the input so as not to cause interference in detecting the other two types.

The performance of our algorithm for correction is significantly better than other previously reported work, with a recall rate of 80.2% and a precision rate of 86.4% on a fair test. While Nakatani and Hirschberg report comparable detection rates, and Hindle reports better correction rates, neither of these researchers attack the complete problem of both detection and correction. Both of them also depend on externally supplied annotations not automatically derived from the input. As for the SRI work, their parser-first strategy and simple repair patterns cause their rates to be much lower than ours. A lot of speech repairs do not look ill-formed, such as "and a boxcar of – and a tanker of OJ", and "and bring – and then bring that orange juice," and are mainly signaled by either lexical or acoustic clues.

3.2.5 Overlapping Repairs

One novel aspect of our algorithm is that it handles overlapping repairs. Two repairs overlap if part of the text is used in both repairs. Such repairs occur fairly frequently in our corpus, and for the most part, our method of processing repairs, even overlapping ones, in a sequential fashion appears successful. Out of the 725 modification and abridged repairs in the training corpus, 23% of them are overlapping repairs, and our algorithm is able to detect and correct 86.6% of them, which is just slightly less than the 88% correction recall rate for all modification and abridged repairs in the training corpus.

Consider the following example (d93-14.2 utt26), which contains four speech repairs, with the last one overlapping the first three.

> and pick up um the en- I guess the entire um p- pick up the load of oranges at Corning

The algorithm is fed one word at a time. When it encounters the first "um", the detection rule for editing terms gets activated, and so a repair pattern is started, with "um" being labeled as an editing term. The algorithm then processes the word "the", for which it can find no suitable correspondences. Next is the fragment "en-". This causes the detection rule for fragments to fire. Since this fragment comes after the editing term in the repair being built, adding it to the repair would violate Rule (2) and Rule (3). So, the algorithm must finish with the current repair, the one involving "um". Since this consists of just a filled pause, it is judged as being an actual repair.

Now that the algorithm is finished with the repair involving "um", it can move on to the next one, the one signaled by the fragment "en-". The next words that are encountered are "I guess", which get labeled as editing terms. The next token is the word "the", for which the algorithm finds a word correspondence with the previous instance of "the". At this point, it realizes that the repair is complete (since there is a word correspondence and all words between the first marked word and the last are accounted for) and so sends it off to

be judged by the statistical model. The model tags it as a repair. Deleting the removed text and the editing terms indicated by the labeling results in the following, with the algorithm currently processing "the".

and pick up the entire um p- pick up the load of oranges at Corning

Continuing on, the next potential repair is triggered by the presence of "um", which is labeled as an editing term. The next token encountered, a fragment, also indicates a potential repair, but adding it to the labeling will violate Rule (2) and Rule (3). So, the pattern builder is forced to finish up with the potential repair involving "um". Since this consists of just a filled pause, it is accepted. This leaves us with the following text, with the algorithm currently processing "p-", which it has marked as a fragment.

and pick up the entire p- pick up the load of oranges at Corning

The next word it encounters is "pick". This word is too far from the preceding "pick" to allow this correspondence to be added. However, the detection clue **mm–mm** does fire, due to the matching of the pair of adjacent words "pick up". This clue is consistent with "p-" being marked as the word fragment of the repair, and so these correspondences are added. The next token encountered is "the", and the correspondence for it is found. Then "load" is processed, but no correspondence is found for it, nor for the remaining words. So, the repair pattern that is built contains an unlabeled token, namely "entire". But due to the presence of the word fragment, the interruption point can be determined. The repair pattern is sent off to be judged, which tags it as a repair. This leaves the following text not labeled as the removed text nor as the editing terms of a repair.

and pick up the load of oranges at Corning

Due to the sequential processing of the algorithm, as the above example demonstrates, and its ability to commit to a repair without seeing the entire utterance, overlapping repairs do not pose a major problem.

Some overlapping repairs can cause problems however. Problems can occur when word correspondences are attributed to the wrong repair. Consider the following example (d93-15.2 utt46).

you have w- one you have two boxcar

This utterance contains two speech repairs, the first is the replacement of "w-" by "one", and the second the replacement of "you have one" by "you have two". Since no analysis of fragments is done, the correspondence between "w-" and "one" is not detected. So, our greedy algorithm decides that the repair after "w-" also contains the word matches for "you" and "have", and that the occurrence of "one" after the "w-" is an inserted word. Due to the presence of the partial and the word matching, the statistical model accepts this proposal, which leads to the erroneous correction of "one you have two boxcars," which blocks the subsequent repair from being found.

3.2.6 Conclusion

This section described a method of locally detecting and correcting abridged and modification speech repairs. Our work shows that a large percentage of speech repairs can be resolved prior to parsing. Our algorithm assumes that the speech recognizer produces a sequence of words and identifies the presence of word fragments. With the exception of identifying fresh starts, all other processing is automatic and does not require additional hand-tailored transcription.

There is an interesting question as to how good the performance can get before a parser is required in the process. Clearly, some examples require a parser. For instance, we can not account for the replacement of a noun phrase with a pronoun, as in "the engine can take as many um – it can take up to three loaded boxcars" without using syntactic knowledge. On the other hand, we can expect to improve on our performance significantly before requiring a parser. The scores on the training set, as indicated in Table 3.7, suggest that we do not have enough training data yet. In addition, we do not yet use any prosodic cues.

Chapter 4

Future Direction

We propose that the problems of detecting and correcting speech repairs, identifying utterance units, and identifying discourse markers can be solved using local context. This would allow these problems to be resolved before syntactic and semantic processing. In the following, we outline the work that needs to be completed.

4.1 Segmenting a Turn

One of the problems that we need to tackle is how utterance units can be identified in spoken dialogue, where the utterance units are the contributions to a dialogue. We feel that these contributions correspond to intonational phrases. Our intuitions are in agreement with Halliday (1967), that speakers segment their speech into chunks, and that the phonological realization of this is the intonational phrase. So, as a first step we need to give stronger evidence of this hypothesis; and second, we need to address how they can be detected.

4.1.1 Evidence for the Intonational Phrase

Much work has been done on grounding. Clark and colleagues (Clark and Wilkes-Gibbs, 1986; Clark and Schaefer, 1989; Clark and Brennan, 1990) have done an impressive job in giving psychological support for the hypothesis that conversants collaborate in understanding. What hasn't been done is to identify the unit of grounding, which we claim is the intonational phrase. Due to implicit acceptance, it is not possible to give direct support that the intonational phrase is the unit of grounding. We can, however, ask subjects to label possible acknowledgment points in a dialogue, and compare this to where transcribers, trained in prosody, marked intonational phrases.

The subjects would be given instructions to mark places in a dialogue where they feel a "uh-huh" could be uttered. In American English, the "uh-huh" is viewed as an acknowledgment of understanding, and so its placement would mark points in the dialogue that the subjects felt grounding could take place.

In a preliminary set of experiments, we had two subjects, untrained in prosody, mark a dialogue with possible acknowledgments. We also had two transcribers, trained in prosody,

mark intonational phrase boundaries. The results are preliminary, but there seems to be some definite correlation between these two measures. More work is needed to better define the instructions, and for comparing the results from the two annotation methods.

4.1.2 Detecting Intonational Phrases

It is unlikely that intonational phrases can be detected using only prosodic information. "When we consider spontaneous speech (particularly conversation) any clear and obvious division into intonational-groups is not so apparent because of the broken nature of much spontaneous speech, including as it does hesitation, repetitions, false starts, incomplete sentences, and sentences involving a grammatical caesura in their middle" (Cruttenden, 1986, pg. 36). Syntactic considerations will also play a part. Hence, it might be possible to combine the various clues using a hidden Markov model, in much the same way that we detect modification repairs (Heeman and Allen, 1994b).

At our disposal for detecting intonational phrases, we have the following resources. First, some of the TRAINS dialogues have been hand-annotated with prosodic features, including intonational boundaries. This corpus can be used for training and test data. Second, we have the HTK toolkit for signal analysis. These tools allow the user to construct and train continuous hidden Markov models. Third, there is the possibility of collaborative work with other researchers on extracting prosodic features. With these resources, and our previous experience with using hidden Markov models for detecting syntactic anomalies, we should be able to detect utterance endings by combining these clues.

4.2 Speech Repairs

As discussed in the previous chapter, we have already investigated two types of speech repairs, those being modification repairs and abridged repairs. However, there is still more to do with these two types, and, as well, fresh starts need to be accounted for. What is missing from our previous work is a model of how the segmentation of a turn into utterance units and the identification of discourse markers interacts with speech repairs.

4.2.1 Modification and Abridged Repairs

Our completed work in this area (Heeman and Allen, 1994b; Heeman and Allen, 1994a) has examined how modification and abridged repairs can be detected using only local context. The context that we used consisted of lexical clues, word correspondences, and syntactic well-formedness. Good results have been achieved using just these factors. However, one problem for this model is how to account for turns like the following (d93-8.3 utt79).

> that's all you need
> you only need one tanker

Here, the local clues that we use force an interpretation in which "you only need" is hypothesized as replacing "you need". To avoid such problems we need to incorporate prosodic

clues into our mode. Nakatani and Hirschberg (1993) have explored this area using hand-transcribed clues. The challenge will be to incorporate clues that can be automatically detected.

We also need to model how modification repairs interact with intonational phrases. We feel that modification repairs can only be in a certain relationship to intonational phrases. Either the modification repair is completely contained in an intonational phrase, or the moment of interruption is at an intonational boundary. Of course, this is not much help in the above example, for the hypothesized interruption point would be at the intonational phrase boundary.

4.2.2 Fresh Starts

For fresh starts, the hearer is faced with the problem of determining the extent of the removed text. In simple exchanges, the speaker undoubtedly cancels back to the beginning of her turn. However, in spoken dialogue, this is often not the case, and the hearer has the task of determining the extent, with relatively few lexical clues. One hypothesis is that for fresh starts, the speaker intends to cancel back to the beginning of the intonational phrase. If this is true, then this would give evidence that intonational phrases play a special role in understanding. Furthermore, one could hypothesize that in order to play this role, intonational phrases must be the grounding unit, and so fresh starts simply cancel the current unit.

If speech repairs are local to intonational phrases, and so the grounding unit, then this would simplify models of incremental processing, for speech repairs would not enter into the picture, since they would be locally resolved. In fact, our work on detecting and correcting modification and abridged repairs (Heeman and Allen, 1994b; Heeman and Allen, 1994a) shows that most of these can be resolved using only using local context and prior to high-level syntactic and semantic analysis.

4.3 Discourse Markers

In our work on detecting and correcting modification and abridged repairs, we use a simplistic model of determining if a given phrase is an editing term. Due to the frequency of phrases like "okay" and "yeah" to mark acknowledgments, we do not allow them as editing terms, which leads to some speech repairs going undetected.

So, we need to better model discourse markers. A prime indicator will be intonational boundaries. Litman and Hirschberg (1990) show that discourse markers are intonationally distinct from sentential uses. In another part of their study, they showed that they could also be detected using a part-of-speech tagger, although not as reliably as with intonational clues. We feel that by incorporating a part-of-speech tagger, with intonational clues and with a model of utterance units, we should be able to reliably detect them in spoken dialogue.

Acknowledgments

I wish to thank my advisor James Allen for his help in choosing this thesis topic, and for his invaluable help in our work on detecting and correcting speech repairs.

I would also like to thank Len Schubert, Mike Tanenhaus, Hannah Blau, George Ferguson, Chung Hee Hwang, Marc Light, Massimo Poesio, Elizabeth Shriberg, Ramesh Sarukkai, David Traum, and Ed Yapratoom for enlightening conversations and helpful feedback.

I also wish to thank Bin Li, Greg Mitchell, and Mia Stern for their help in both transcribing the TRAINS dialogues and giving us useful comments on our annotation scheme for speech repairs.

Bibliography

Allen, James, James Hendler, and Austin Tate, editors. 1990. *Readings in Planning.* Morgan Kaufmann Publishers.

Allen, James F. 1990. Formal models of planning. In James Allen, James Hendler, and Austin Tate, editors, *Readings in Planning.* Morgan Kaufmann Publishers, pages 50–54.

Allen, James F. and C. Raymond Perrault. 1980. Analyzing intention in utterances. *Artificial Intelligence*, 15:143–178. Reprinted in (Grosz, Sparck Jones, and Webber, 1986).

Allen, James F. and Lenhart K. Schubert. 1991. The TRAINS project. Technical Report 382, Department of Computer Science, University of Rochester, May.

Austin, J. L. 1962. *How to do things with words.* New York: Oxford University Press.

Bachenko, J. and E. Fitzpatrick. 1990. A computational grammar of discourse-neutral prosodic phrasing in english. *Computational Linguistics*, 16(3):155–170.

Beach, Cheryl M. 1991. The interpretation of prosodic patterns at points of syntactic structure ambiguity: Evidence for cue trading relations. *Journal of Memory and Language*, 30(6):644–663.

Bear, John, John Dowding, and Elizabeth Shriberg. 1992. Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, pages 56–63.

Bear, John, John Dowding, Elizabeth Shriberg, and Patti Price. 1993. A system for labeling self-repairs in speech. Technical Note 522, SRI International, February.

Bear, John and Patti Price. 1990. Prosody, syntax, and parsing. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics*, pages 17–22, Pittsburgh, June.

Beckman, Mary E. and Gayle M. Ayers. 1994. Guidelines for ToBI labelling, version 2.0. Manuscript and accompanying speech materials, Ohio State University, (obtain by writing to tobi@ling.ohio-state.edu).

Beckman, Mary E. and Julia Hirschberg. 1994. The ToBI annotation conventions. Manuscript, Ohio State University, (obtain by writing to tobi@ling.ohio-state.edu).

Brieman, Leo, Jerome H. Friedman, and Richard A. Olshen. 1984. *Classification and Regression Trees.* Monterrey, CA: Wadsworth & Brooks.

Brown, Gillian and George Yule. 1983. *Discourse Analysis.* Cambridge: Cambridge University Press.

Bruce, Bertram C. 1975. Generation as a social action. In *Theoretical Issues in Natural Language Processing (TINLAP-1)*, pages 64–67. Reprinted in (Grosz, Sparck Jones, and Webber, 1986).

Church, K. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Preceedings of the 2nd Conference on Applied Natural Language Processing*, pages 136–143, Febuary.

Clark, Herbert H., editor. 1992. *Arenas of Language Use.* University of Chicago Press and CSLI.

Clark, Herbert H. and S. E. Brennan. 1990. Grounding in communication. In L.B. Resnick, J. Levine, and S.D. Behreno, editors, *Perspectives on Socially Shared Cognition.* APA.

Clark, Herbert H. and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13:259–294. Reprinted in (Clark, 1992), pages 144–175.

Clark, Herbert H. and Michael F. Schober. 1989. Understanding by addressees and overhearers. *Cognitive Psychology.*

Clark, Herbert H. and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22:1–39. Reprinted in (Clark, 1992), pages 107–143.

Cohen, Philip R. and C. Raymond Perrault. 1979. Elements of a plan-based theory of speech acts. *Cognitive Science*, 3(3):177–212. Reprinted in (Grosz, Sparck Jones, and Webber, 1986).

Cohen, Robin. 1984. A computational theory of the function of clue words in argument understanding. In *Proceedings of the 10th International Conference on Computational Linguistics (COLING '84)*, pages 251–255.

Coulthard, Malcolm and David Brazil. 1981. Exchange structure. In M. Coulthard and M. Montgomery, editors, *Studies in Discourse Analysis.* Routledge & Kegan Paul, pages 82–106.

Cruttenden, Alan. 1986. *Intonation.* Cambridge: Cambridge University Press.

Crystal, D. 1980. Neglected grammatical factors in conversational English. In S. Greenbaum, G. Leech, and J. Svartvik, editors, *Studies in English Linguistics.* Longman.

Dowding, John, Jean Mark Gawron, Doug Appelt, John Bear, Lynn Cherny, Robert Moore, and Douglas Moran. 1993. Gemini: A natural language system for spoken-language understanding. In *Proceedings of the 31th Annual Meeting of the Association for Computational Linguistics*, pages 54–61.

Fikes, Richard E. and Nils J. Nilsson. 1971. Strips: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence*, 2:189–208. Reprinted in (Allen, Hendler, and Tate, 1990).

Ford, Cecelia and Sandra Thompson. 1991. On projectability in conversation: Grammar, intonation, and semantics. Presented at the *Second International Cognitive Linguistics Association Conference*, August.

Fox, Barbara A. 1987. *Discourse Structure and Anaphora.* Cambridge: Cambridge University Press.

Gee, James Paul and Francois Grosjean. 1983. Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognitive Psychology*, 15(3):411–458.

Grice, H. P. 1957. Meaning. *Philosophical Review*, 66:377–388.

Grosjean, Francois. 1983. How long is the sentence? Predicting and prosody in the on-line processing of language. *Linguistics*, 21(3):501–529.

Gross, Derek, James Allen, and David Traum. 1993. The Trains 91 dialogues. Trains Technical Note 92-1, Department of Computer Science, University of Rochester, June.

Grosz, Barbara and Julia Hirschberg. 1992. Some intonational characteristics of discourse structure. In *Proceedings of the 2nd International Conference on Spoken Language Processing (ICSLP-92)*, pages 429–432, October.

Grosz, Barbara J. and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Grosz, Barbara J., Karen Sparck Jones, and Bonnie Lynn Webber, editors. 1986. *Readings in Natural Language Processing.* Morgan Kaufmann Publishers.

Halliday, M. A. 1967. Notes on transitivity and theme in English: Part 2. *Journal of Linguistics*, 3:199–244.

Heeman, Peter and James Allen. 1994a. Detecting and correcting speech repairs. In *Proceedings of the 32th Annual Meeting of the Association for Computational Linguistics*, pages 295–302, Las Cruces, New Mexico, June.

Heeman, Peter and James Allen. 1994b. Tagging speech repairs. In *ARPA Workshop on Human Language Technology*, Princeton, March.

Heeman, Peter A. and James Allen. 1994c. Annotating speech repairs. Unpublished manuscript.

Heeman, Peter A. and James Allen. 1994d. Dialogue transcription tools. Trains Technical Note 94-1, Department of Computer Science, University of Rochester, August.

Heeman, Peter A. and James Allen. 1994e. The Trains 93 dialogues. Unpublished manuscript.

Heeman, Peter A. and Graeme Hirst. 1992. Collaborating on referring expressions. Technical Report 435, Department of Computer Science, University of Rochester, August.

Hindle, Donald. 1983. Deterministic parsing of syntactic non-fluencies. In *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*, pages 123–128.

Hirschberg, Julia and Barbara Grosz. 1992. Intonational features of local and global discourse structure. In *Proceedings of the DARPA Workshop on Speech and Natural Language Processing*, pages 441–446, Arden House, February.

Hirschberg, Julia and Janet Pierrehumbert. 1986. The intonational structuring of discourse. In *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics*, pages 136–144.

Hirst, Graeme, Susan McRoy, Peter Heeman, Philip Edmonds, and Diane Horton. 1994. Repairing conversational misunderstandings and non-understandings. *Speech Communications*. To Appear.

Labov, William. 1966. On the grammaticality of everyday speech. Paper presented at the Linguistic Society of America Annual Meeting.

Levelt, Willem J. M. 1983. Monitoring and self-repair in speech. *Cognition*, 14:41–104.

Levelt, Willem J. M. and Anne Cutler. 1983. Prosodic marking in speech repair. *Journal of Semantics*, 2(2):205–217.

Levinson, Stephen C. 1983. *Pragmatics*. Cambridge University Press.

Lickley, R. J. and E. G. Bard. 1992. Processing disfluent speech: Recognizing disfluency before lexical access. In *Proceedings of the 2nd International Conference on Spoken Language Processing (ICSLP-92)*, pages 935–938, October.

Lickley, R. J., R. C. Shillcock, and E. G. Bard. 1991. Processing disfluent speech: How and when are disfluencies found? In *Proceedings of the 2nd European Conference on Speech Communication and Technology (Eurospeech 91)*, pages 1499–1502, Genova, Italy, September.

Litman, Diane and Julia Hirschberg. 1990. Disambiguating cue phrases in text and speech. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING '90)*, pages 251–256.

Litman, Diane J. and James F. Allen. 1987. A plan recognition model for subdialogues in conversations. *Cognitive Science*, 11(2):163–200, April–June.

MADCOW. 1992. Multi-site data collection for a spoken language corpus. In *Proceedings of the DARPA Workshop on Speech and Natural Language Processing*, pages 7–14, February.

Marcus, Mitchell and Donald Hindle. 1990. Description theory and intonation boundaries. In Gerry T. M. Altmann, editor, *Cognitive Models of Speech Processing*. MIT Press, pages 483–512.

Matessa, Michael and Sheryl R. Young. 1994. Algorithms for processing spontaneous speech: Restarts, mid-utterance corrections and ill-formed input. Unpublished manuscript.

McRoy, Susan W. 1993. Abductive interpretation and reinterpretation of natural language utterances. Doctoral dissertation, Technical Report CSRI 288, Department of Computer Science, University of Toronto.

Nakajima, Shin'ya and James Allen. 1992. Prosody as a cue for discourse structure. In *Proceedings of the 2nd International Conference on Spoken Language Processing (ICSLP-92)*, pages 425–428, October.

Nakajima, Shin'ya and James F. Allen. 1993. A study on prosody and discourse structure in cooperative dialogues. Trains Technical Note 93-2, Department of Computer Science, University of Rochester, September.

Nakatani, Christine and Julia Hirschberg. 1993. A speech-first model for repair detection and correction. In *Proceedings of the 31th Annual Meeting of the Association for Computational Linguistics*, pages 46–53.

Novick, David. 1991. Controlling interaction with meta-acts. Technical Report CS/E 91-001, Oregon Graduate Institute of Science and Technology.

O'Shaughnessy, Douglas. 1992. Analysis of false starts in spontaneous speech. In *Proceedings of the 2nd International Conference on Spoken Language Processing (ICSLP-92)*, pages 931–934, October.

Ostendorf, M., P. Price, J. Bear, and C. W. Wightman. 1990. The use of relative duration in syntactic disambiguation. In *Proceedings of the DAPRA Speech and Natural Language Workshop*, Hidden Valley, Pennsylvania, June.

Oviatt, Sharon L. 1994. Predicting and managing spoken disfluencies during human-computer interaction. In *ARPA Workshop on Human Language Technology*, Princeton, March.

Pierrehumbert, J. B. 1980. The phonology and phonetics of english intonation. Doctoral dissertation, Massachusetts Institute of Technology.

Pierrehumbert, Janet and Julia Hircshberg. 1990. The meaning of intonational contours in the interpretation of discorse. In Philip R. Cohen, Jerry Morgan, and Martha E. Pollack, editors, *Intentions in Communication*, SDF Benchmark Series. MIT Press, pages 271–311.

Pollack, M. E., J. Hirschberg, and B. Webber. 1982. User participation in the reasoning processes of expert systems. Technical Note MS-CIS-82-9, University of Pennsylvannia, July.

Price, P. J., M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong. 1991. The use of prosody in syntactic disambiguation. *Journal of the Accoustic Society of America*, 90(6):2956–2970, December.

Reichman-Adar, Rachel. 1984. Extended person-machine interface. *Artificial Intelligence*, 22:157–218.

Sacerdoti, Earl D. 1975. Planning in a hierarchy of abstraction spaces. *Artificial Intelligence*, 5:115–135. Reprinted in (Allen, Hendler, and Tate, 1990).

Sacks, Harvey, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735, December.

Schegloff, E. A. and H. Sacks. 1973. Opening up closings. *Semiotica*, 7:289–327.

Schegloff, Emanuel A., Gail Jefferson, and Harvey Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language*, 53:361–382.

Searle, J. R. 1969. *Speech acts: An essay in the philosophy of language.* Cambridge: Cambridge University Press.

Sidner, Candace L. 1985. Plan parsing for intended response recognition in discourse. *Computational Intelligence*, 1(1):1–10.

Silverman, K., M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. 1992. ToBI: A standard for labelling English prosody. In *Proceedings of the 2nd International Conference on Spoken Language Processing (ICSLP-92)*, pages 867–870.

Steedman, Mark J. 1990. Syntax and intonational structure in a combinatory grammar. In Gerry T. M. Altmann, editor, *Cognitive Models of Speech Processing.* MIT Press, pages 457–482.

Talkin, David. 1989. Looking at speech. *Speech Technology*, 4:74–77.

Traum, David R. 1991. Towards a computational theory of grounding in natural language conversation. Technical Report 401, Department of Computer Science, University of Rochester.

Traum, David R. and James F. Allen. 1994. Discourse obligations in dialogue processing. In *Proceedings of the 32^{th} Annual Meeting of the Association for Computational Linguistics*, pages 1–8, Las Cruces, New Mexico, June.

Traum, David R. and Elizabeth A. Hinkelman. 1992. Conversation acts in task-oriented spoken dialogue. *Computational Intelligence*, 8(3).

Walker, Marilyn A. 1993. Informational redundancy and resource bounds in dialogue. Doctoral dissertion, Institute for Research in Cognitive Science report IRCS-93-45, University of Pennsylvania, December.

Wang, Michelle Q. and Julia Hirschberg. 1992. Automatic classification of intonational phrase boundaries. *Computer Speech and Language*, 6:175–196.

Weischedel, Ralph, Marie Meteer, Richard Schwartz, Lance Ramshaw, and Jeff Palmucci. 1993. Coping with ambiguity and unknown words through probabilistic models. *Computational Linguistics*, 19(2):359–382.

Wightman, Colin W., Stefanie Shattuck-Hufnagel, Mari Ostendorf, and Patti J. Price. 1992. Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Accoustic Society of America*, 91(3):1707–1717, March.

Young, Sheryl R. and Michael Matessa. 1991. Using pragmatic and semantic knowledge to correct parsing of spoken language utterances. In *Proceedings of the 2nd European Conference on Speech Communication and Technology (Eurospeech 91)*, Genova, Italy, September.