The Computational Processing of Intonational Prominence: A Functional Prosody Perspective

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:25619466">http://nrs.harvard.edu/urn-3:HUL.InstRepos:25619466</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
The Computational Processing of Intonational Prominence: A Functional Prosody Perspective

Christine Hisayo Nakatani

TR-15-97
May 1997

Center for Research in Computing Technology
Harvard University
Cambridge, Massachusetts
THE COMPUTATIONAL PROCESSING OF INTONATIONAL PROMINENCE: A FUNCTIONAL PROSODY PERSPECTIVE

A Thesis presented

by

Christine Hisayo Nakatani

to

The Division of Engineering and Applied Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Computer Science

Harvard University

Cambridge, Massachusetts

May, 1997
DEDICATION

This thesis is dedicated with gratitude, love and admiration to
Setsuyo Yamanishi Asari
and
to the memory of
Yoshio John Asari
Intonational prominence, or accent, is a fundamental prosodic feature that is said to contribute to discourse meaning. This thesis outlines a new, computational theory of the discourse interpretation of prominence, from a FUNCTIONAL PROSODY perspective. Functional prosody makes the following two important assumptions: first, there is an aspect of prominence interpretation that centrally concerns discourse processes, namely the discourse focusing nature of prominence; and second, the role of prominence in language processing in general, and discourse processing in particular, is not essentially separate from the processing of other grammatical, nonprosodic information.

This thesis develops a computational theory of prominence interpretation by explaining how prominence serves as an inference cue in discourse processing. Prominence signals changes in the attentional status of entities in a discourse model, while nonprominence signals that the realized entities are already in discourse focus. Evidence for the new theory is provided by distributional analysis of a spontaneous narrative monologue. New discourse processing algorithms that integrate form of expression, grammatical function and intonational prominence information for reference resolution and attentional state modeling show how the principles of the theory may be applied in SPEECH UNDERSTANDING systems.

Finally, aspects of the new theory are explored in accent prediction experiments on a corpus of spontaneous and read direction-giving monologues. Machine learning is used to investigate the extent to which the analyzed higher-level linguistic features associated with prominence may combine with lower-level features that are known to influence accent assignment. Original constituent-based accent prediction experiments attempt to bootstrap off of established knowledge about citation-form accenting, and begin to develop an understanding of how the examined features of discourse context may be integrated into accent assignment systems for text-to-speech synthesis.
# Contents

1 Introduction .................................................. 1  
1.1 Under-utilization of Speech Cues in SLS's ......................... 1  
1.2 Integrating Speech and Text Processing ........................... 2  
1.3 Thesis Overview ............................................ 5  

2 Processing Intonational Prominence .............................. 7  
2.1 Descriptive Work ........................................... 8  
2.1.1 What is Prominence? ..................................... 8  
2.1.2 Associated Linguistic Phenomena .......................... 10  
2.2 Speech Modeling .............................................. 13  
2.2.1 Recognizing Prominence ................................... 13  
2.2.2 Synthesizing Prominence .................................. 15  
2.3 Prominence in Language Processing .............................. 15  
2.3.1 Cognitive Models .......................................... 16  
2.3.2 Text-to-speech Systems .................................... 18  
2.3.3 Message-to-speech Systems ................................ 19  
2.3.4 Summary .................................................. 21  
2.4 A Functional Prosody Perspective ............................... 21  

3 Narrative Study ............................................... 25  
3.1 The Given/New Study ......................................... 26  
3.1.1 Hypotheses .............................................. 26  
3.1.2 Analysis .................................................. 26  
3.1.3 Results .................................................. 27  
3.2 The Attentional Modeling Study ................................ 29  
3.2.1 Hypotheses .............................................. 29  
3.2.2 Discourse Structure Analysis of the Narrative ............. 29  
3.3 A Discourse-Based Interpretation of Accent Function .......... 33  
3.3.1 Manipulations of Local Focus: Pronouns ................ 35  
3.3.2 Manipulations of Global Focus: Explicit Forms .......... 37  
3.3.3 Evaluation .............................................. 40  
3.4 Algorithms for Processing Prominence in Discourse ............ 41  
3.4.1 Principles .............................................. 41  
3.4.2 Algorithms .............................................. 42  
3.5 Related Work ................................................. 45  
3.6 Discussion .................................................. 48
4 Task-oriented Monologue Study:

Part I 51

4.1 Corpus Design and Collection ........................................... 52
4.1.1 Motivation ............................................................. 52
4.1.2 Desiderata for a Spoken Language Corpus ......................... 54
4.1.3 Boston Directions Corpus ........................................... 54

4.2 Foundational Analyses .................................................. 57
4.2.1 Text Analysis ......................................................... 57
4.2.2 Prosodic Analysis ................................................... 59

4.3 Subcorpus ................................................................. 60

5 Baseline Accent Prediction Experiments 63

5.1 Previous Work ............................................................ 64
5.2 A Rule Learning System for Data Exploration ......................... 65
5.3 Classification of Accent Types .......................................... 66
5.4 Predictors ................................................................. 66
5.4.1 Part-of-Speech Tags ................................................... 66
5.4.2 Part-of-Speech Classes .............................................. 66
5.4.3 Phrase, Sentence and Task Position ................................ 67
5.4.4 Tonal Contours ........................................................ 68
5.4.5 NP-accent Assignment Using TTS .................................. 68
5.4.6 Word Lemma ........................................................... 68

5.5 Baseline Experiments ..................................................... 69
5.5.1 Part-of-Speech Tags and Classes .................................... 69
5.5.2 Adding Word Lemma ............................................... 70
5.5.3 Adding Linear Position and Tonal Contour Features .......... 71
5.5.4 Adding NP-accent Assignment ...................................... 72
5.6 Assessment ............................................................... 72

6 Task-oriented Monologue Study:

Part II 75

6.1 Syntactic Analysis .......................................................... 75
6.2 Lexico-syntactic and Lexico-semantic Analysis of Noun Phrase Referring Expressions ........................................ 76
6.2.1 Grammatical Function .............................................. 76
6.2.2 Form of Referring Expression ..................................... 77

6.3 Coreference Coding of Noun Phrase Referring Expressions ........ 77
6.4 Discourse Segmentation Analysis ...................................... 78
6.4.1 Methods ............................................................... 78
6.4.2 Segmentation Results on Expert Data ............................. 79

7 Accent Prediction Experiments on Higher-level Features 82

7.1 Classification of Constituent Accentuation ............................ 83
7.2 Predictors ................................................................. 84
7.2.1 Features from the Lex Model ....................................... 84
7.2.2 Phrasal, Sentence and Segment Position .......................... 85
7.2.3 Syntactic Constituency ............................................. 85
7.2.4 Lexico-syntactic and Lexico-semantic Features ............... 86
7.2.5 Global Focusing ...................................................... 86
This research was supported by a National Science Foundation (NSF) Graduate Research Fellowship, NSF Grant No. IRI-90-09018, NSF Grant No. IRI-93-08173, NSF Grant No. CDA-94-01024, the former AT&T Bell Laboratories and AT&T Labs – Research. The author gratefully acknowledges these institutions and especially the members of the Linguistics Research Department at Bell Labs and the Human-Computer Interface Research Department at AT&T Labs for providing a stimulating environment in which to learn and work. The views and conclusions contained in this document are those of the author.

The research reported in Chapter 3 of this thesis has been partially published and presented elsewhere (Nakatani, 1993; Nakatani, 1994b; Nakatani, 1994a; Nakatani, 1995; Nakatani, 1997a; Nakatani, 1997b). The author thanks the organizers of these efforts for the opportunity to present this work, and participants and anonymous reviewers for their comments.

The design and collection of the Boston Directions Corpus, reported in Chapter 4, Section 4.1, is joint work with Barbara Grosz and Julia Hirschberg and has been partially published elsewhere (Nakatani, Grosz, and Hirschberg, 1995; Hirschberg, Nakatani, and Grosz, 1995; Hirschberg, 1995). The related research reported in Chapter 6, Section 6.4, is joint work with Julia Hirschberg and has been published elsewhere (Hirschberg and Nakatani, 1996). The author thanks Barbara Grosz and Julia Hirschberg for permitting the presentation of joint work in this thesis.

Full bibliographic information is found in the References.
This thesis would not have been created without the inspiration and encouragement of many individuals. First, I thank my thesis committee: Barbara Grosz (advisor), Julia Hirschberg and Stuart Shieber. I am one amongst many who have been motivated by Barbara’s seminar on computational discourse modeling to take up hard but interesting problems in NLP. Barbara spent many office hours discussing discourse modeling with me and I have learned a great deal from it. I thank her also for reviewing my written work over the years and for providing support for this thesis. Julia has served for exactly nine years as my mentor in prosody research, and her efforts to shape me up into a speech person have been nothing short of herculean. The years of apprenticeship under Julia at Bell Labs evoke many happy memories of tinkering, toiling, sharing and discovering. The work in this thesis bears her influence throughout. I am especially grateful to her for the many consultations on empirical matters, large and small, and for critiquing my chapters in real-time. Stuart quite efficiently helped me to revise the original thesis draft in crucial places and to sharpen the scientific arguments, and I thank him.

Many other individuals directly aided this research. I thank Virginia Merlini, who collected the narrative speech, and those researchers who shared tools, methods and advice on how to use them, especially during corpus collection and analysis. They are Chinatsu Aone, Eric Brill, William Cohen, Michael Collins, Giovanni Flammia, Lynette Hirschman, Diane Litman, Lloyd Nakatani, Becky Passonneau, Richard Sproat, the ToBI group and Victor Zue. Gregory Ward reviewed my linguistic coding schemes and Lyn Walker reviewed my chapter drafts; both are much improved as a result. This thesis has also benefited from the able research assistance of David Ahn, Charles Reiss, Alejandro Sepulveda, Philip Maymin and Stefanie Jannedy. Alan Capil executed editorial work with extraordinary skill and made the office cheerful and orderly. Peg Schafer and Corwyn Miyagishima provided technical support beyond the ordinary. The exceptional care of Dr. Emil Pascarelli, Dr. Wenfei Xie, Kathy Phillips, P.T. and Michael Weiss, P.T. helped me heal from work-related RSI. I am also grateful to Lloyd Nakatani for creating a special foot keyboard that enabled me to complete this thesis.

For making the journey worthwhile, I am indebted to numerous friends, colleagues and family. Karen Lochbaum, Andy Kehler and Stan Chen could not have been kinder colleagues. Karen taught me the ropes of getting through Harvard from beginning to end and her friendship is much appreciated. Andy made me laugh and let me cry, and Stan made me laugh and give better talks. I thank my other peers at Harvard and members of the Discourse Working Group, the Boston-area Prosody Forum and the local research community — especially Christine Bartels, Nancy Chang, Giovanni Flammia, Mari Ostendorf, Lisa Selkirk, Stefanie Shattuck-Hufnagel and Candy Sidner — for their collegiality; professors whose teaching has influenced me — especially Aravind Joshi, Susumu Kuno, Mark Liberman and Ellen Prince; those who have informally taught, advised and encouraged me through the years — Cinzia Avesani, Mary Beckman, Nick Campbell, Peter Gordon, Don Hindle, Jan van Kuppevelt, Ana Pereira, Fernando Pereira, Martha Pollack, Jan van Santen, Chilin Shih, Richard Sproat, Jacques Terken and Gregory Ward; friends and colleagues who sustain me — especially Meiting and Eric Brill, Duff Campbell, Sunny Chae, Susan Curtis, Ken Fan, Pascale Fung, Ellen Germain, Hong Hu, Vivie Lee, Kenjiro Matsuda, Jae Son, Lyn Walker, Steve Whittaker and Dekai Wu; and my relatives — Setsuyo Asari, Shunji Asari, Yoichi, Carol, David and Lisa Asari, Mitsume Yamanishi and Don, Betty, Allison and Kimberly Hiroto — for their unfailing support. Laura and Warren Nakatani always try to bring out something better in me, and Lloyd and Kayo Nakatani taught me the importance of strength and integrity in all endeavors. Finally, for his patience and caring, I thank Robert Chu.
Chapter 1

Introduction

Spoken language communication between human and machine is one of the oldest, most popularized promises of ARTIFICIAL INTELLIGENCE (AI) research. This problem of SPEECH UNDERSTANDING involves not only an understanding of the organization of thoughts into words and sentences, but also knowledge of the rich sound structure of language and its relationship to meaning.

1.1 Under-utilization of Speech Cues in SLS’s

Researchers have just begun to scratch the surface of the speech understanding problem by creating experimental SPOKEN LANGUAGE SYSTEMS (SLS). With few exceptions, these systems are built by linking together separate, stand-alone speech and NATURAL LANGUAGE PROCESSING (NLP) components, as illustrated by the prototypical SLS architecture given in Figure 1.1. Simply put, the processing model for the architecture in Figure 1.1 transforms speech into text, and then text into speech. First, the user produces a spoken utterance, which serves as input to an AUTOMATIC SPEECH RECOGNIZER (ASR). The ASR module outputs its best hypothesis or $n$-best hypotheses as to the words making up the speech stream to an utterance interpretation or NATURAL LANGUAGE UNDERSTANDING (NLU) component. The NLU component then applies morphological, lexical, syntactic, semantic and perhaps pragmatic analyses to the word string(s) and constructs a representation of the meaning(s) of the word string(s) in logical form or some other semantic representation language. A discourse management module then processes this meaning representation in

Figure 1.1: Prototypical architecture for spoken language system (SLS), with natural language processing (NLP) components inserted between automatic speech recognition (ASR) and text-to-speech synthesis (TTS).
context and outputs a message structure encoding the system’s reply to a sentence generator. The sentence generator creates a text string, which is passed along to a text-to-speech (TTS) module for processing and then is synthesized out loud to the user.

One obstacle in building SLS’s is that uneven efforts have been devoted to the research and development of individual components. For example, ASR and utterance interpretation are relatively developed technologies, while discourse management and sentence generation in context have received less attention and are less tested. A more serious problem with this prototypical SLS architecture is that it simply sandwiches NLP components between ASR and TTS technologies that send and receive text strings respectively. This yields not only a language interpretation module that is largely deaf and dumb, but also speech processing components that interface at a distance through text, not speech. Such circumstances arise from the fact that ASR and TTS models traditionally have borne little relation to one another, with recognizers trying to filter out acoustic and prosodic variation as noise to get at the underlying words and synthesizers trying to squeeze the last bit of information out of the text to bring the words back to life in spoken form. This disconnect between ASR and TTS is so extreme that many leading edge systems simply opt to output pre-recorded speech, or synthesize canned output using an off-the-shelf synthesizer. Conversely, TTS and MESSAGE-TO-SPEECH (MTS) synthesis systems focus on generating speech from text or message structures, leaving aside the question of how similar aspects of meaning for the coming message can be derived from speech.

In sum, the interfaces between SLS components are extremely impoverished, consisting of a word string passed from ASR to the NLP components, and from the NLP components back to TTS. The result is that speech cues to meaning are ignored or under-utilized during both spoken language understanding and speech generation.

1.2 Integrating Speech and Text Processing

While this insertion of an NLP subsystem between ASR and TTS technologies fits conveniently with the modern evolution of language processing, the impoverished interfaces between components in SLS architectures do not reflect our scientific knowledge about the relationship of speech cues to linguistic meaning. Evidence is mounting that if all of the acoustic-prosodic structure of spoken language is lost to the language understanding components, certain aspects of meaning will not be accurately disambiguated or fully interpreted. And if the relationship of derived linguistic meaning to acoustic-prosodic speech features is not carefully modeled, speech synthesizers will never excel at communicating contextual aspects of meaning that lend naturalness to artificial speech.

To appreciate the wide ranging interactions of speech and text cues, consider some of the many speech understanding phenomena that have already been investigated:

- **Speech can disambiguate text.** Lexical stress in many languages distinguishes homographs, such as the noun con’tent and the adjective con’tent¹. Intonational prominence is said to disambiguate anaphoric bindings (Hirschberg and Ward, 1991a) and the scope of focus-sensitive operators (Rooth, 1985; Rooth, 1991). Intonational phrasing and durational information can disambiguate syntactic constituent structure (Lehiste, 1979; Wales and Toner, 1979), inter alia, and can be used to score parses for ASR string hypotheses (Ostendorf, Wightman, and Veilleux, 1993). Prominence and phrasing together can be used to disambiguate sentential and discourse structural usages of cue phrases or clue words, such as now and so (Hirschberg and Litman, 1993).

- **Text can disambiguate speech.** Language models trained on textual co-occurrence information captured by n-grams can improve the accuracy of word string hypotheses in ASR using lexical class and syntactic information (Jelinek, 1976; Jelinek and Mercer, 1980).
• **Speech can signal discourse-level aspects of meaning.** Pitch range, energy and timing information are correlated with topic structure (Silverman, 1987; Avesani and Vayra, 1988; Hirschberg and Grosz, 1992; Ayers, 1992; Swerts, Gelyukens, and Terken, 1992; Hirschberg and Nakatani, 1996). This intonational variation can be modeled in speech synthesis systems to efficiently communicate topic structure (Silverman, 1987; Davis and Hirschberg, 1988). Prosodic cues to discourse structure can also be combined with text-based information to identify discourse segment boundaries in spoken language (Passonneau and Litman, 1997; Swerts and Ostendorf, 1995). Finally, intonational contours can help identify uses of propositionally redundant information and thereby shed light on discourse structure (Kießling et al., 1993; Walker, 1993).

• **Speech can convey illocutionary or paralinguistic information.** Intonational contours can disambiguate questions that occur in declarative form from statements (Hirschberg, 1989; Daly and Zue, 1990); convey propositional attitudes such as (un)certainty and (in)credulity (Liberman and Sag, 1974; Hirschberg and Ward, 1991b) or signal the speaker’s social status with respect to his or her listener(s) (McLemore, 1991). Acoustic-prosodic cues such as glottalization, abrupt pauses and other hesitation phenomena can alert the listener to the occurrence of a performance error or disfluency (Hindle, 1983; Bear, Dowding, and Shriberg, 1992; Nakatani and Hirschberg, 1994; Shriberg, 1994). Finally, variation in fundamental frequency, speaking rate and amplitude can convey speaker emotion (Scherer, Ladd, and Silverman, 1984; Ladd et al., 1985; Cahn, 1988; Higuchi, Hirai, and Sagisaka, 1997).

In short, the interactions between speech and text are pervasive — extending into areas of lexical, syntactic, semantic, pragmatic and discourse processing and user modeling — and fundamental to determining diverse aspects of linguistic meaning.

Considering the abundance of phenomena to model and the fact that sometimes speech disambiguates text and at other times text disambiguates speech, it seems desirable to avoid having to develop special-purpose mechanisms for handling each of these phenomena in isolation. Rather, the argument for computational elegance and efficiency favors a uniform framework for representing and interpreting speech features in SLS architectures. When the traditional SLS architecture is conceptually revised in this manner, a new, SPEECH UNDERSTANDING architecture arises as presented in Figure 1.2. The term, *speech understanding*,

Figure 1.2: Architecture for speech understanding system, integrating text and speech processing.

is chosen to remind us that acoustic and prosodic features of speech need to be understood in systems that use spoken language as the input or output modality.

The key to such a system is to start out with speech and text on equal footing, demanding that automatic *speech* recognizers not simply supply text transcriptions of spoken words, but actually recognize acoustic and prosodic features of speech that are crucial to language interpretation. The recognizer in effect outputs an orthographic transcription and an acoustic-prosodic or SPEECH TRANSCRIPTION, time-aligned with reference to the speech signal. Next, the utterance interpreter is free to make use of the speech transcription to
interpret relevant speech features in relation to features of the associated text. Once the contributions and effects of certain speech features are recorded in the appropriate semantic structures, it would be possible to discharge those features of the speech transcription that are fully interpreted at the stage of utterance interpretation. Those features of the speech transcription that are needed by the discourse manager, such as pitch range and pause information, could be preserved as annotations of the semantic formalism. The semantic formalism would indirectly encode a reduced speech transcription, just as reduced syntactic information may be usefully recorded for aspects of discourse processing (Sproat and Ward, 1987; Kehler, 1995). However, since it is not yet known precisely which speech features bear on utterance interpretation and which bear on discourse interpretation, it may be advisable at this stage of speech understanding research to make the full speech transcription available at all stages of interpretation. As better models of speech interpretation are developed, refinements in representations can be made.

When discourse-level interpretation is completed, the discourse manager, as before, outputs a message structure encoding the system’s reply. But this time it is received by a MTS module instead of a TTS synthesizer. This MTS component realizes a complete integration of text generation and speech generation, formerly carried out in sequence by a sentence generator and then a TTS system. By reversing the various interpretation rules for speech and text features employed during utterance and discourse interpretation, the MTS component can make choices for speech and text realization in parallel, in a system of mutual constraints.

Undoubtedly, the realization of this speech understanding architecture will require many new research and development efforts. But it is also the case that there is a solid starting point from which researchers can work toward this vision of speech understanding. The concept of parallel transcriptions of different facets of a language stream underpins the notion of a ToBI transcription in the ToBI prosodic transcription standard (Silverman et al., 1992; Beckman and Ayers, 1994; Beckman and Hirschberg, 1994; Pitrelli, Beckman, and Hirschberg, 1994). As stated by Beckman and Ayers, “A ToBI transcription consists minimally of a recording of the speech, an associated electronic or paper record of the fundamental frequency contour, and (the transcription proper) symbolic labels for events arranged in four parallel tiers” (1994, Section 1.1). Two of these tiers, the tone and break-index tiers, represent the core prosodic transcription, capturing melodic prominence, phrasing and contours and prosodic juncture information respectively. The transcription of words and recording of paralinguistic events (such as disfluencies or coughs) make up the orthographic and miscellaneous tiers respectively. This multi-tier model, which is discussed further in Chapter 4, can be readily extended by adding on additional tiers to express additional properties.

Implicit in ToBI is the idea that all tiers are related to each other via time-alignment, and the time-course of the represented events is supplied by the recording of the speech itself. Thus, the spiral notebook metaphor from phonology can be applied to ToBI, in which the speech signal serves as a sort of binding in the spiral notebook and each page expresses properties of the speech along relatively independent dimensions. The formalism for each page or tier is arbitrarily complex and is in principle unconstrained by the representational formalism of any other dimension. At this stage of research, there remain many open questions concerning the relationship of text to speech, such as that of syntactic constituency to prosodic phrasing. Because of this, it makes sense not to impose formal constraints between different representational systems prematurely, but rather to let these constraints emerge from research on descriptively adequate representations of important dimensions of speech and text. The spiral notebook metaphor embodied in ToBI is thus a general way of thinking about speech transcriptions and text transcriptions. Particular choices of tiers and of formalisms in each tier can be revisited without undoing the general framework.

A speech recognizer that could automatically generate a reasonable speech transcription is achievable in the near future. Prosody recognition systems using the ToBI standard as well as other prosodic transcription schemes are improving with the availability of new training corpora and new prosodic recognition models (see Chapter 2), and recent work has introduced unsupervised learning methods for prosodic recognition.
A speech transcription containing the equivalent of ToBI’s core prosodic transcription would provide in and of itself a solid basis for incorporating speech features from ASR into the NLP components for language understanding.

1.3 Thesis Overview

This thesis explores how one fundamental aspect of a prosodic transcription, namely INTONATIONAL PROMINENCE, can be made use of not only in language interpretation, but also in speech generation to better convey contextual aspects of meaning associated with the use of prominence. While much research has been done on this topic, few attempts have been made to relate and reconcile a wealth of theoretical proposals and empirical findings. In this thesis, a uniform account of the role of prominence in speech understanding is developed in various stages.

In Chapter 2, a review of previous research identifies open problems and conflicts among existing accounts. The research issues to be addressed in this thesis are then formulated within the framework of FUNCTIONAL PROSODY. The functional prosody perspective on prominence processing makes two important assumptions. First, there is an aspect of prominence interpretation that centrally concerns discourse processes, namely the discourse focusing nature of prominence. Second, the role of prominence in language processing in general, and discourse processing in particular, is not essentially separate from the processing of other grammatical, nonprosodic information.

Chapter 3 presents a first study on intonational prominence in a spontaneous narrative. Guided by distributional analysis, new principles governing the discourse functions of accent are developed, which overturn, reconcile and extend previous research findings on the interpretation of prominence. In particular, these new principles specify the interactions of intonational prominence with grammatical function, form of referring expression and the attentional status of discourse entities in a computational discourse model. These principles are incorporated into high-level discourse processing algorithms that make use of prominence information for reference resolution and attentional modeling.

A second study, presented in the next four chapters, Chapter 4 through Chapter 7, investigates intonational prominence in a new corpus of direction-giving monologues. The design, collection and analysis of this corpus are described in Chapter 4 and Chapter 6. The labeled direction-giving corpus is used in two sets of accent prediction experiments using machine learning techniques. The first set of machine learning experiments, reported in Chapter 5, establishes performance benchmarks for word-based accent prediction using a variety of lower-level linguistic features studied in the literature. For the second set of machine learning experiments, reported in Chapter 7, the accent prediction task is redefined as the prediction of patterns of deviation from citation form accentuation of noun phrase constituents. These original experiments, using a variety of higher-level linguistic features, show which factors contribute most heavily to accenting decisions in discourse. Finally, the findings of this thesis are summarized in Chapter 9.

This thesis contributes to our understanding of the computational processing of intonational prominence in several areas. It is distinguished from previous accounts of the discourse functions of prominence by the generality of the postulated principles governing prominence interpretation and the specificity of their application to detailed classes of prominent and nonprominent expressions; by the breadth of linguistic factors from text and speech that are considered in relation to prominence; by the methodological rigor applied in speech and text analysis; by its presentation of empirical evidence for and the empirical evaluation of proposals concerning the processing of prominence; by its treatment of prominence in high-level discourse-processing algorithms that can be used for discourse interpretation in speech understanding systems; and by its attempt to integrate new high-level features influencing prominence into automatically trained accent prediction systems that can be used in speech synthesis systems. The theoretical and empirical results of this thesis are based on American English, although cross-linguistic evidence for certain findings has been
uncovered in recent studies of intonational prominence and discourse structure in Italian (Avesani, 1996) and in Japanese (Venditti, 1996). Finally, the research in this thesis makes few commitments to particular linguistic or computational linguistic theories, but rather synthesizes important findings from different disciplines (e.g. computational linguistics, linguistics, psycholinguistics) and perspectives (e.g. TTS, MTS, functional syntax). The principles, algorithms and systems presented in this thesis may be usefully incorporated into components of speech understanding systems that utilize diverse representational formalisms and linguistic theories.
Chapter 2

Processing Intonational Prominence

Research on the role of speech features in language understanding can be roughly divided into three areas: descriptive work, speech modeling and language processing. Descriptive work is aimed at identifying a specific speech feature and the linguistic phenomenon or phenomena with which it is associated. Speech modeling research seeks to develop an understanding of the acoustic and prosodic properties of the speech feature sufficient for its automatic recognition and generation. Finally, language processing research concentrates on how the speech feature is handled in cognitive and computational models that capture the relationship of the feature to the related linguistic phenomenon or phenomena. Research in these three areas often occurs simultaneously, each area being influenced by findings in the other two. In addition, speech understanding research requires efforts in multiple disciplines, including acoustics, computational linguistics, linguistics, psycholinguistics and speech engineering.

One basic speech feature that has been studied at the descriptive, speech modeling and language processing levels, by researchers representing all of these disciplines, is INTONATIONAL PROMINENCE. Descriptive work has yielded a working definition of the feature, in terms of both its physical properties and its major areas of contribution to linguistic meaning. With respect to speech modeling, robust speech technologies are being developed for prominence recognition and sophisticated theoretical models for generating prominence in intonational contours have been applied in speech synthesis systems. Sections 2.1 and 2.2 review various results from the descriptive and speech modeling areas of research on prominence. The third area of research, language processing, puts to use the speech technologies and linguistic analyses by modeling the role of prominence in speech interpretation and generation. Section 2.3 reviews language processing work on prominence, including cognitive processing models based on psycholinguistic experimentation as well as computational processing models that have been employed especially in speech generation systems.

While much has been learned about prominence itself, how prominence relates to other linguistic structures is a matter requiring considerable further research. The approaches taken in tackling this issue have differed significantly in their view of the linguistic nature of prominence — predictable from syntax, as a grammatical entity existing within a rule-governed prosodic structure that mediates the relationship between it and its linguistic meaning, or as a free and direct signifier of communicative intentions. This thesis assumes that prominence acts as a functional morpheme in a prosodic grammar. Section 2.4 discusses the implications of such a perspective — the FUNCTIONAL PROSODY perspective — on the linguistic nature of prominence, for framing the problems of prominence processing addressed in this thesis. The research goals of this thesis are motivated by the vision of speech understanding outlined in Chapter 1, but are sufficiently narrow and well-defined within the framework of functional prosody to allow for the empirical testing of theoretical results and the formulation of various algorithms that integrate prominence into language processing.
2.1 Descriptive Work

2.1.1 What is Prominence?

**INTONATIONAL PROMINENCE** can be defined as the acoustic-prosodic marking of words as perceptually salient, relative to other words in a sentence or utterance. The physical and perceptual characteristics of prominence have been described by acousticians and psycholinguists, while its linguistic properties have been described by phoneticians and phonologists.

The perceptual correlates of intonational prominence in English include pitch, timing and intensity (Fry, 1955). That is, words that are marked with changes in fundamental frequency, longer duration and greater amplitude or energy — which are the acoustic correlates of pitch, timing and intensity respectively — are generally perceived as prominent. Perceptual experiments, starting with Fry (1955), established that pitch may serve as the primary cue to prominence, but subsequent experiments on whispered and monotonized speech have also shown that pitch is not a necessary cue (Cutler and Darwin, 1981). Acoustic studies by Klatt (1979), van Santen and Olive (1990) and Eefting (1991) and showed that the presence of prominence increases word duration.

There is also consensus at a general level on the phonetic and phonological nature of prominence. Various schools or traditions share in common the notion that prominence is best represented as a feature of the intonational structure of a sentence, and that the alignment of prominence with particular syllables of particular words is mediated by mapping rules between intonation structure and a metrical grid that represents the rhythmic structure of spoken language (Liberman, 1975; Liberman and Prince, 1977; Selkirk, 1984). The metrical grid does not entirely determine the location of prominence, but it constrains its alignment with the text along several dimensions. In particular, the metrical grid provides a hierarchical representation that encodes the relationships among different levels of lexical stress. These stress levels correspond to the primary and secondary stress levels found in dictionary entries, such as *a₁-ba-cus, a-ro²-ma-ther¹-a-py* and *in²-ter-disc¹-i-pli-na²-ry*. In general, for each word there is one privileged syllable receiving the highest stress, the **PRIMARY STRESS**. Acoustic-phonetic cues to prominence fall on primary stress-marked syllables. Further, metrical well-formedness rules operating over phrasal domains may alter the alignment of prominence by changing the pattern of primary stress syllables, giving rise to phenomena such as **STRESS SHIFT** (Liberman and Prince, 1977). When the word, *sixteen*, for example, is placed right before a word with primary stress on the first syllable, such as *pandas*, the primary stress in *sixteen* is shifted from the second syllable to the first syllable to preserve the preferred alternation of weak and strong syllables, resulting in the stress assignment, *six¹-teen pan¹-das*.

The treatments of intonation arising out of autosegmental and metrical phonology (Liberman, 1975; Liberman and Prince, 1977; Pierrehumbert, 1980; Selkirk, 1984), although differing in detail, share a crucial assumption, namely that a qualitative distinction exists between intonational prominence and lexical stress. Lexical stress concerns the relative prominence of syllables within a word and is a feature of words in isolation. In contrast, as expressed in the stated definition, intonational prominence concerns the perceptual salience of words in context. It may be treated in phonological theories as a level of stress above and beyond lexical stress (Liberman, 1975; Liberman and Prince, 1977; Pierrehumbert, 1980; Beckman, 1986), but it cannot be reduced to the notion of lexical stress. On this view, lexical stress determines a potential locus for the acoustic-prosodic highlighting of a word relative to other words in context (Bolinger, 1958). Theories of intonation that adhere to this distinction between lexical stress and intonational prominence have been grouped together under the rubric, **INTONATIONAL PHONOLOGY** (Ladd, 1992).

In theories of intonational phonology, prominence is often equated with some notion of **ACCENT** (Bolinger, 1958), such as **PITCH ACCENT** (Pierrehumbert, 1980) or **SENTENCE ACCENT** (Gussenhoven, 1984). The latter term is often used interchangeably with **NUCLEAR STRESS** (Chomsky and Halle, 1968). In Chomsky and Halle’s theory of stress (1968), the nuclear stress was said to be the highest stress within a sentence.
To avoid blurring the important distinction made between the prominence of syllables within words and the prominence of words within sentences, the term, *stress*, will be reserved in this thesis for the former notion and the terms, *prominence* and *accent*, for the latter. Further, throughout this thesis, usage of the term, *intonational prominence*, or simply, *prominence*, underscores the theory-neutral nature of this phenomenon as the acoustic-prosodic marking of salient words in context. The terms, *pitch accent, accent* and *accentuation*, refer to the more technical notion of prominence communicated primarily (but not exclusively) by pitch excursions, and are used in discussion of empirical analyses of prominence as pitch accent (Pierrehumbert, 1980).

As pitch movement is often a salient perceptual cue to prominence, intonational prominence or accent is often described in terms of its realization in the pitch contour or *fundamental frequency* (F0) of speech. There are at least three major areas of disagreement in treatments of prominence in various intonational systems. First, the description of the pitch levels of prominence is a matter of debate that does not centrally concern this thesis. The second area of disagreement is whether prominence provides pitch targets in the intonational contour, or specifies pitch movements with particular slopes. The third area concerns the domain of prominence, or of the accentual unit. Different positions on these issues lead to important differences in how patterns of pitch movements that are associated with prominence are described in intonational systems. In certain systems, most notably those of the ‘British’ School (Crystal, 1969) and the ‘Dutch’ School (Cohen and ’t Hart, 1967; Van Katwijk, 1974), prominence is associated with pitch motions, which are said to lend prominence to the words on which or near which they occur. Examples in English are the *high-rise* typical of yes/no questions in English, and the declarative *fall*. Both of these describe pitch movements occurring at the end of sentences. These two contours, the high-rise and fall, are illustrated by the first and last contributions of speaker A in the examples in Figure 2.1 taken from Ladd (1980). The illustrated pitch motions, referred to as *tails*, often contain several pitch movements in one accentual unit. The pitch movements are often distinguished by the pitch levels associated with them.

In intonational systems that have been grouped into the ‘American’ School (Ladd, 1980), which includes foundational work on Swedish as well (Bruce, 1977), melodic contours at the end of phrases are not analyzed as holistic accentual units, but rather are decomposed into sequences of targets, namely pitch accent(s) and phrasal accent(s) (Bruce, 1977; Pierrehumbert, 1980; Beckman and Pierrehumbert, 1986). In this thesis, an ‘American’ School system of intonation is used for intonational analysis, namely Pierrehumbert’s theory of American English intonation (1980). Pierrehumbert defined a regular language (i.e. finite-state grammar) for describing the prosody of American English using a hierarchy of tonal elements — pitch accents, phrase accents and boundary tones. The regular grammar for Pierrehumbert’s system of American

```
A: fini
   Did he

B: Well, he’s over at the grad school office right now, and...

A: Ye fin
   ah, but did he

Figure 2.1: Examples of high-rise and fall (British School).
```
English intonation is given below:

\[
\text{Intonational Phrase} \rightarrow (\text{Intermediate Phrase})^+ \quad \text{Boundary Tone} \quad (2.1)
\]

\[
\text{Intermediate Phrase} \rightarrow (\text{Pitch Accent})^+ \quad \text{Phrase Accent} \quad (2.2)
\]

The highest level unit in the grammar, the major or INTONATIONAL phrase, is comprised of one or more minor or INTERMEDIATE phrases, plus a BOUNDARY TONE, as defined in Rule 2.1. Each intermediate phrase is made up of a sequence of one or more PITCH ACCENTS, followed by a PHRASE ACCENT, as defined in Rule 2.2. A pitch accent is defined as a low (L* — read low-star) or high (H*) pitch excursion, or a combination of both low and high excursions (L*+H — read as low-star plus high, L+H*, H*+L, H+L*), in the melodic contour of the sentence.\(^1\) The tonal targets are time-aligned with the vocalic nucleus of the highest stressable syllable of a word and are said to render the associated word perceptually prominent. The specific tone aligned with the vocalic nucleus in an accented word is marked by the asterisk. In Pierrehumbert’s theory, the sentence accent or nuclear stress is referred to as the NUCLEAR ACCENT, and is defined as the most prominent pitch accent in the intermediate phrase, which often occurs in phrase-final position. There is no explicit marking of the nuclear accent in the abstract tonal transcription. The phrase accent is a high (H-) or low (L-) tonal target controlling the interpolation of the melodic contour from the end of the last pitch accent to (roughly) the end of the intermediate phrase. The boundary tone, either high (H%) or low (L%), serves as a tonal target for melodic interpolation from the immediately preceding phrase accent to the end of the utterance. In Pierrehumbert’s theory, only two tones, low and high, are used to express the relative height of pitch targets, and changes in pitch level are described by potentially subtle pitch range relationships between successive pitch targets.

### 2.1.2 Associated Linguistic Phenomena

In early theoretical work on prominence, it was proposed that the communicative function of prominence is to highlight information that is new or important (Bolinger, 1972). On this view, the universal function of accent is to serve as a form of acoustic-prosodic deixis, telling the listener, Watch that! (Brown and Yule, 1983). This somewhat vague generalization has been more closely analyzed and refined by studies focusing on distinct subclasses of the highlighting function of prominence. These hypothesized functions of prominence can be roughly divided into the following categories: emphasis (Bolinger, 1972; Schmerling, 1976; Ladd, 1980; Fuchs, 1984), contrast (Bolinger, 1961; Lakoff, 1971; Jackendoff, 1972), semantic focus (Ladd, 1979b; Gussenhoven, 1983; Rooth, 1985; Selkirk, 1993), and the given/new distinction (Halliday, 1967; Bruce, 1977; Brown, 1983; Fuchs, 1984; Terken, 1984; Hirschberg and Pierrehumbert, 1986; Gundel, Hedberg, and Zacharski, 1989).

It is no coincidence that all of these categories affect information structure or information packaging, and as would be suspected, many relationships exist among these uses of prominence. It is helpful, nonetheless, to keep in mind paradigmatic cases of these phenomena. In illustrating these functions of prominence, a basic model of mutual belief is assumed to underlie all definitions of information status. The model rests on two assumptions that are quite minimal and are standardly applied in computational discourse modeling. First, it is assumed that information that has been shared in the preceding portion of a discourse makes up an important part of the speaker’s and hearer’s mutual beliefs about the discourse. This dynamically changing set of mutual beliefs may be referred to as the DISCOURSE MODEL. Second, when a discourse participant makes a contribution to the discourse, it is generally the case that s/he expresses NEW

---

\(^1\)Pierrehumbert’s theory allows that timing and intensity cues alone can communicate prominence, but it has not been easy to define hard and fast rules for identifying such cases, in this or other theories, based on acoustic measures of duration, amplitude and energy.
information as well as given information (Halliday, 1967; Prince, 1981). The speaker intends the hearer to integrate the new information into the discourse model by virtue of its relation to one or more established mutual beliefs communicated by the given information. These two parts of information structure, namely the new information and the related given information, are referred to by different names in various schools of pragmatics and discourse analysis. These include foreground/background (Firbas, 1971), discourse-new/discourse-old (Prince, 1988), comment/topic (Kuppevelt, 1993), rHEME/THemE (Kuno, 1975) and predicated/unpredicated information (Hirschberg and Pierrehumbert, 1986). It is implicit in this model that information status is defined with three points of reference: the speaker’s beliefs, the hearer’s beliefs and the speaker’s and hearer’s mutual beliefs about the discourse model. For example, information can be new to the discourse and to the hearer, but is of course already known to the speaker who communicates it.

Given this basic model, the linguistic phenomena influenced by intonational prominence — emphasis, contrast, semantic focus and given/new — may be more uniformly described. Emphasis may be loosely described as the highlighting of information that is already represented in the discourse model. Intonational emphasis is achieved by extreme changes in the physical correlates of prominence — pitch, duration and intensity. In the following naturally-occurring example from the spontaneous narrative analyzed in Chapter 3, the third and final reference to Freud is a clear case of intonational emphasis.\(^2\)

\[
\text{They all put Freud on a pedestal} \\
\text{HE is an icon okay} \\
\text{HE can do no wrong}
\]

It has been hypothesized that although changes in the physical correlates of prominence are gradient, the perception of non-emphatic versus emphatic prominence is categorical (Ladd, 1993). Above certain thresholds for the physical correlates, a word is perceived not merely as more prominent but as emphatically intoned.

When marking contrast, intonational prominence is often referred to as contrastive stress (Bolinger, 1961). It may involve either the contradiction or comparison of entities or propositions in the discourse model. Bolinger (1961) cites the following example from Coleman (1914), in which contrast between the individual words, blue and black, involves contradiction, while the contrast between you and I involves comparison.

\[
\text{YOU may call it DARK blue, I should say it was BLACK.}
\]

Contradiction is perhaps the classic usage of contrastive stress. But there are many instances of contrast that do not involve contradiction. One oft-analyzed set of examples is due to Lakoff (1971):\(^3\)

(a) John called Mary a Republican, and then SHE insulted HIM.
(b) John called Mary a Republican, and then she INSULTED him.
(c) # John praised Mary, and then SHE insulted HIM.
(d) John praised Mary, and then she INSULTED him.
(e) John insulted Mary, and then SHE insulted HIM.
(f) # John insulted Mary, and then she INSULTED him.

On Lakoff’s analysis, in sentence (a), it is presupposed that calling someone a Republican is an insult. So, the semantic roles assigned to John and Mary are contrasted between the two clauses. In (b), this presupposition must not hold in order for the sentence to be felicitous. Rather, in (b) a contrast is drawn between

\(^2\)In linguistic data presented throughout this thesis, the intonational prominence of a word will be indicated by capitalization of the entire word.

\(^3\)Following convention, pragmatic (as opposed to syntactic) ill-formedness, technically referred to as infelicity, is marked by the symbol ‘#’.
the two predicates. This analysis is supported by the unambiguous sentences, (c)-(f). These sentences are all cases of contrast as comparison, as they contain no contradictions of any sort. While the notion of contradiction is rather well-defined, a precise definition of contrast in the comparative sense has eluded linguists (although subclasses of this phenomenon such as scalar implicature (Hirschberg, 1985), comparison induced by parallelism, e.g. (Hobbs, 1985; Dalrymple, Shieber, and Pereira, 1991; Kehler, 1995), and focus constructions (Rooth, 1985) have been analyzed). It is these cases involving comparison, as opposed to the contradiction of entities or propositions in the discourse model, that are on the borderline between contrast and focus.

Studies too numerous to list exhaustively have investigated the focus-marking role of intonational prominence. Many of these, especially since Bolinger (1972), have adopted a semantic notion of focus, under which the new information in a discourse contribution — the foreground, comment or rhyme — is said to be intonationally highlighted to help the listener focus on this important information when interpreting the sentence. The classic test for semantic focus is the question-answer test: the potentially focused constituent is replaced with a wh-expression (e.g. who, what, why), to make a question out of the utterance; if the utterance is an appropriate answer to the question, then the replaced constituent is a semantic focus. Thus, to borrow Jackendoff’s (1972) question-answer example, the answer in the exchange below is felicitous,

A: Who ate the beans?
B: FRED ate the beans.

while the following answer is not.

A: Who ate the beans?
B: # Fred ate the BEANS.

The exact manner in which intonational prominence conveys semantic focus information is a matter of considerable debate. Theories developed in relation to syntactic or prosodic theories of phrase structure have shaped this problem as one of FOCUS PROJECTION of F-MARKERS associated with lexical items that are intonationally prominent (Jackendoff, 1972; Schmerling, 1976; Gussenhoven, 1984; Selkirk, 1984; Rochemont, 1986). Some theories concerning intonation and focus assume that nuclear stress is the marker of intonational focus for a sentence, and is therefore the only prominence feature relevant to focus interpretation (Schmerling, 1976; Gussenhoven, 1984; Rooth, 1985). Others state that all prominences, nuclear or not, potentially convey semantic focusing information (Bolinger, 1972; Hirschberg and Pierrehumbert, 1986; Selkirk, 1993). Recent psycholinguistic studies suggest that nuclear accents are distinguished perceptually from non-nuclear accents, even when acoustic variation in nuclear accent type is present (Ayers, 1996). More research in this direction is needed to establish the precise role of this perceptual distinction in linguistic interpretation.

In addition to semantic focus, intonational prominence has been associated with a discourse notion of focus (Brown, 1983; Fuchs, 1984; Terken, 1984; Hirschberg and Pierrehumbert, 1986; Terken and Hirschberg, 1994). These studies have investigated different variations on the hypothesis that new information is accentuated and given information is deaccented. In a study of task-oriented speech, Brown (1983) confirmed a general tendency for given information to be unaccented, and new information to be accented, where lexical items referring to referents previously mentioned in the discourse were considered given. In her corpus, very few instances of discourse-old information (less than 5%) were accented. Brown also examined form of referring expression and syntactic information and concluded that when syntax and form of expression convey conflicting information statuses, form of expression provides the reliable clue.

Other researchers have observed that given information can be made prominent in more regular ways than suggested by Brown’s results. Fuchs (1984) hypothesized that given items may be left unaccented to signify their givenness, or may be accented “to establish them as a point of relevance/‘newness’ with
respect to the question of immediate concern at that point” (p. 144). Later work addressed from different linguistic perspectives the circumstances under which given information is accented. For example, Horne (1991) investigated metrical-phonological constraints, Selkirk (1993) examined syntactic factors and Hirschberg (1993) explored discourse modeling accounts. Needham (1990) demonstrated that deaccentuation can also be licensed for INFERRABLE referents (Prince, 1981) that referred to part of a previously mentioned discourse entity, contra Brown (1983). All of these studies have relied on the notion of given/new as discourse-given/discourse-new (Prince, 1988). That is, the first-mention of an entity in the discourse was judged to convey new information, while subsequent mentions were judged to convey given information.

Several of these proposals, nonetheless, allow that accenting behavior may be influenced in a dynamic way by the unfolding discourse model. Hirschberg (1993) noted that for professionally read news stories, discourse-old information could be accented when it was not focused upon in the recent context. A more local dynamism in the discourse model was shown to affect the deaccenting of given information in short discourses in which subjects described the movements of geometric objects (Terken and Hirschberg, 1994). By controlling the lexical and syntactic forms of utterances in their elicitation method, Terken and Hirschberg demonstrated that structural properties, such as the persistence of grammatical function and surface position, contributed to the deaccentuation of given information. Local and global influences of discourse structure on accentuation were more directly investigated by Terken (1984). In elicited, instruction-giving monologues, Terken found that the first introduction of a discourse topic was generally accented and subsequent references to the topic were often expressed as unaccented pronouns. In contrast, non-initial references to non-topical entities were often expressed as accented full forms. Terken’s results suggest that with the introduction of each new discourse topic, the accentability of information is re-evaluated. The refinements of the given/new hypothesis contributed by these studies can be summarized as follows: information is predicted to be intonationally prominent when first introduced into a discourse, while continually focused upon information such as a discourse topic is predicted to lose its prominence (Terken, 1984; Terken and Hirschberg, 1994). Further, discourse-old information that is not continually focused upon may be made prominent when reintroduced into the discourse (Terken, 1984; Hirschberg, 1993).

2.2 Speech Modeling

The speech modeling of prominence has reached a level of sophistication that is not yet exploited in either automatic speech interpretation or automatic speech generation. That is to say, while experimental techniques for prominence recognition have been developed, their usage in experimental NLU systems has yet to be tested. Similarly, technologies for generating intonational contours have outstripped the ability of synthesis systems to automatically make use of the full-range of realizable intonational variation in linguistically appropriate ways. Recent results in both the recognition and synthesis of prominence are reviewed below.

2.2.1 Recognizing Prominence

As would be expected, the speech recognition of intonational prominence has focussed on the modeling of the physical correlates of prominence: fundamental frequency movement, duration, amplitude and energy. Work in this area can be divided into special-purpose techniques that identify the location of prominent syllables or words, and general prosodic modeling approaches that identify prominences as elements of an intonational contour.

Examples in the first category include emphasis detectors developed for applications such as speech skimming (Chen and Withgott, 1992; Arons, 1994). Chen and Withgott’s experimental system used an hidden Markov model to detect prominences based on energy and fundamental frequency parameters in
conversational speech. The performance of their algorithm was tested in terms of its effectiveness at selecting summarizing phrases that human subjects selected. The kappa score of 0.52 showed weak reliability of the system, but it also was shown that human subjects achieved no greater inter-labeler reliability. Arons’ implemented speech skimmer (1994) used a simpler model of changes in fundamental frequency to identify locations in the speech segment for controlling speech playback. Task-dependent evaluation showed the system performed better than the random or arbitrarily determined selection of locations for controlling playback.

While these systems may serve the purposes of certain applications like speech skimming, their performance as true detectors of intonational prominence is unevaluated. Prominence detection techniques that are embedded in more general prosodic recognition systems have achieved detection rates above 80% when tested on hand-labelled prosodic databases. For example, Campbell (1993) used normalized segmental duration and energy to detect prominences in a corpus of three sets of 100 sentences each, containing contrastive or focus-shifting accent. The prominence detector was 92% correct on the first set of sentences which were read in context, 78% correct for the second set of sentences read in randomized order (i.e. out-of-context) and 72% for the third set of spontaneous sentences elicited in interactive conversation. This prominence detector is being used to partially guide the retrieval of speech segments for large-unit concatenative speech synthesis (Campbell, 1996). Wightman and Campbell (1994) have also integrated this model into a full prosodic recognition system, using hidden Markov models, that detects prominences as well as phrasal and junctural boundaries (i.e. break-indices). This system achieved a performance rate of 86% for prominence detection on the Boston University Radio News corpus of professionally read news speech (Ostendorf, Price, and Shattuck-Hufnagel, 1995).

Ostendorf and Ross (1997) developed a dynamical system model for full prosodic recognition. This model incorporates parameters that directly reflect aspects of the phonological theory of intonation defined by Pierrehumbert and colleagues (Pierrehumbert, 1980; Beckman and Pierrehumbert, 1986), such as pitch range and normalized fundamental frequency contour for syllable units, as well as parameters demonstrated to be useful in previous work, such as syllable duration. This system also adopted principles of the Fujisaki model of intonation generation (1969), but adapted them by making them probabilistic. The dynamical system approach yielded 89% accuracy for accent detection (not judging tonal labelings) on the Boston University Radio News corpus (Ostendorf, Price, and Shattuck-Hufnagel, 1995).

Finally, studies of lexical stress detection, which is related to prominence detection, have provided techniques that make extensive use of dictionary information to locate potential sites for prominence (Hieronymus, McKelvie, and McInnes, 1992; Bishop, 1992).

Prominence detection has recently been embedded within experimental spoken language systems, which is the scenario envisioned in the speech understanding architecture proposed in Chapter 1. Ostendorf (1996) reports that given word transcriptions, which can be realistically assumed only in architectures in which ASR word hypotheses are computed prior to prominence detection and are highly accurate, prominence can be detected at 81% accuracy for the ATIS database of spontaneous dialogue speech (MADCOW, 1992). Jackson and colleagues (1991) carried out a pilot study on semantic interpretation, testing a template matching system that was used as a back-off model when parsing failed. While 68% (36/53) of the template matching errors involved missing keywords, 90% of those missing keywords were in the processed utterance and were intonationally prominent. This pilot study therefore illustrates how prominence information can be used in a real language understanding task. In spoken language understanding systems, prosody can serve as an independent knowledge source that can supply interpretation information unobtainable from or overlooked by word-based ASR and by full syntactic and semantic processing.
2.2.2 Synthesizing Prominence

Of the above systems, only the dynamical system model for prosodic recognition is reversible for use in prosodic generation. This is perhaps not surprising, since that model strove to directly model parameters of at least two established models for intonation synthesis, namely Pierrehumbert’s prosodic grammar model and Fujisaki and Kawai’s (1988) accent filtering and superposition model. Several other uses of these two prosodic models in speech synthesis are discussed below.

A synthesis model based on Pierrehumbert’s theory of intonation was developed quite directly from the interpolation rules, mentioned above, that describe the phonetic realization of abstract phonological descriptions (Pierrehumbert, 1981; Anderson, Pierrehumbert, and Liberman, 1984). In this model, the tonal features at the three levels of prosodic structure — pitch accent, phrase accent and boundary tone — provide fundamental frequency targets, relative to each other and to a topline, which models the maximum fundamental frequency values that can be realized in a given pitch range.

A different manner of producing intonational contours from accent targets is given by the Fujisaki model (1969), in which contours are synthesized by superimposing two component signals generated by physiologically motivated functions. The first signal is the phrase component, modeled as a critically damped second-order system responding to an impulse function. The second signal is the accent component, modeled as a critically damped second-order system responding to rectangular functions that capture the relative heights and locations of the prominences in a sentence. This model can be refined and enhanced to capture various linguistic and prosodic insights, as in Möbius’ system for German intonation synthesis (1997). As remarked on, intonation systems differ in their choice of units of analysis. The Pierrehumbert and Fujisaki models share the notion that accents define specific tonal targets, and that the generation of intonational contours involves connecting these pitch targets, whether by direct interpolation or by superposition of an underlying phrasal contour. A very different system of intonation generation has been developed based on the ‘British’ and ‘Dutch’ Schools of intonational phonology. In Dutch synthesis, for example, patterns of pitch movements, such as accent-lending rises and accent-lending falls, are time-aligned with words that are assigned prominence by the system (’t Hart and Cohen, 1973; ’t Hart and Collier, 1975). Generation of the contour then involves connecting these pre-computed segments of the contour together. In the Tilt system developed for British English (Taylor, 1994), no distinctions are made between pitch accents, phrase accents and boundary tones. Contours are simply described as a smoothed sequence of rise, fall and connective elements, that carry their own duration and amplitude specifications as well as height and slope parameters. The Tilt model has proven to be robust, but since it requires an extra step of mapping from a Tilt description to an abstract representation of phonological prominence in the British system to recover the prominence locations, it is difficult to evaluate this system with regard its abstract modeling of prominence.

Within any speech synthesis system, issues of timing, such as segmental variation, phrase lengthening effects and inter-phrasal pausing, arise at all levels of phonological description. There is no model of segmental or syllable timing inherent in any of the synthesis models discussed above. However, since duration is a basic perceptual cue to prominence, the proper assignment of duration at the segmental and syllable levels is important. Many synthesis systems, as well as the duration-based prominence recognition systems described above, rely on the conventional segmental duration modeling used in speech recognition for time-aligning intonational contours and speech segments. Speech timing models based on time warping have been proposed by van Santen (1994) as a generalization of segmental duration modeling.

2.3 Prominence in Language Processing

Work on the cognitive and computational processing of prominence has been concentrated on two areas, its role in marking semantic and discourse focus and its relationship to linguistic properties of text such as
lexical category and syntactic information. Section 2.3.1 reviews psycholinguistic studies of prominence processing, while Sections 2.3.2 and 2.3.3 review computational models that have been used for assigning accent in text-to-speech and message-to-speech synthesis systems respectively.

2.3.1 Cognitive Models

Many psycholinguistic studies have explored the role of prominence in language processing. The results have been interpreted along two lines, which can be characterized as abstract and concrete (Cutler and Ladd, 1983). On the abstract view, prominence is treated as a functional morpheme, c.f. (Pierrehumbert, 1980; Selkirk, 1993), that can be related to features of the morphosyntactic representation of an utterance. On the concrete view, which focuses on the acoustic-phonetic characteristics of a prosodic phenomenon, c.f. (Fowler and Housum, 1987), the communicative force of prominence is limited to that of acoustic-prosodic deixis. These two views arise not so much as conflicting accounts, but as alternative interpretations of a large body of experimental results.

A number of experiments have focused on whether it is accent location, or acoustic-prosodic properties of accent itself, that render accented words especially accessible to lexical processing modules. Phoneme detection experiments have shown that words that bear sentence accent are processed more quickly than words that do not (Cutler and Foss, 1977; Shields, McHugh, and Martin, 1974). Sentence accents also speed the detection of mispronounced words (Cole, 1978). Other experiments explored whether such effects arise from the interaction of accent and rhythmic structure, or from only the acoustic-prosodic cues carried by accenting. Meltzer and colleagues (1976), Martin (1979), and Buxton (1983) found that altering the temporal structure of normal speech at the syllable level delayed target detection. Cutler (1976) showed that prosodic aspects of the portion of an utterance preceding a sentence accent provided listeners with the same detection advantage whether the word in sentence accent location was the original, accented word or a spliced in version of the word spoken with neutral prosody. A later experiment (Cutler, 1987) established that while different prosodic cues may allow the listener to predict or attend to words in sentence accent location, for stimuli that were processed to have mismatches among cues, sentence accent did not facilitate speech processing.

A different line of investigation has pursued the relationship of accenting to information structure through language comprehension experiments. Experiments on single-accent sentences demonstrated that intonational focusing was a stronger marker of new information than word order variation that is produced by passivization (Most and Saltz, 1979) or fronting in Dutch (Nootseboom et al., 1981; Kruyt, 1985). Other studies established straightforward effects of focus-marking in question-answer constructions (Eady and Cooper, 1986). In studies of semantic focusing in Dutch, Nootseboom and Kruyt (1987) operationalized the notion of semantic focus by designing news stories consisting of sentence pairs. Below is shown one of the first sentence contexts, and all four possible second sentences, in which the intonation was systematically varied. Three additional first sentences established contexts that yielded all permutations of given and new for the target words in the second sentences.

Subjects were asked to rate the acceptability of each sentence pair. The overall result was that subjects preferred new information to be intonationally focused, while given information could be focused or not. Focusing of given information was licensed by the occurrence of a constituent later in the sentence that was new, or by the use of an alternate referring expression, such as the town when Gouda appeared in the first sentence. Nootseboom and Kruyt analyzed this phenomenon as one of thematic-focusing. Lexical repetitions under these conditions, however, proved to be less acceptably accented than alternate referring expressions. Nootseboom and Kruyt (1987) mentioned one acceptability judgment that they could not account for, namely the complete deaccenting of the mayor of Gouda, when the mayor was new and Gouda given in the discourse context established by the first sentence. This finding can be explained by the result of Needham (1990) on
(1) de stad sc Gouda staat onder WATER.
The city of Gouda is flooded.

(2a) [Gouda/Gouda] is GETEISTERD door een WOLKBREUK.
Gouda has been afflicted by a cloudburst.

(2b) [DE GEMEENTE/de gemeente] is GETEISTERD door een WOLKBREUK.
The town has been afflicted by a cloudburst.

(2c) [DE BURGEMMESTER/debergemmester] van [Gouda/Gouda] heeft de OVERHEID om
HULP gevraagd.
The mayor of Gouda has asked the government for help.

Figure 2.2: Stimuli used in Nooteboom and Kruyt’s (1987) psycholinguistic study of focus and prominence.

... the deaccentuation of inferrable referents.

Terken and Nooteboom (1987) evaluated the discourse focusing nature of accentuation in comprehension experiments on multi-sentence discourses describing the movements of geometric objects. They found that accentuation appropriate to the discourse status of referents, namely accenting of new information and deaccenting of given information, speeded the comprehension of the utterances in which they occurred.

To measure comprehension, subjects were asked to verify whether the spoken description of the geometric objects correctly described a visual display of the same objects. When given information was accented, or when new information was deaccented, verification latencies were higher. Previous studies showed that new information may be most efficiently processed when accented, but this study was the first to provide evidence that accenting on given information could have detrimental effects on sentence comprehension.

To explain these results, Terken and Nooteboom concluded that reference resolution proceeds differently for accented and unaccented expressions, hypothesizing that listeners assume that an unaccented expression refers to a member of a “restricted set of activated entities” in the discourse context, while the interpretation of an accented expression is not constrained in this manner (Terken and Nooteboom, 1987, p. 148). They further speculated that mental representations of entities were constructed differently for accented and unaccented referring expressions. They proposed that representations were built bottom-up for accented items, allowing the hearer to use the content of the expression to resolve the reference. In contrast, unaccented expressions were said to be resolved top-down, by taking as candidates for reference resolution the restricted set of activated entities.

Finally, following the concrete approach to prominence perception, researchers have substituted a notion of INTELLIGIBILITY for accent. Intelligibility is measured by various acoustic parameters. In this paradigm, the first occurrence of a word or referring expression is taken to be a baseline. Subsequent tokens with reduced duration, amplitude, vowel clarity and so forth are said to be relatively less intelligible than the original token or earlier tokens. Researchers have suggested that the marking and interpretation of given and new information in spoken language is directly related to intelligibility (Fowler and Housum, 1987). Bard and colleagues (1995), for example, claim that the unintelligibility of words leads to their resolution with respect to the discourse context. This claim seems to be contradicted by Cutler’s (1976) experiment on predicted accents, in which less intelligible, neutral words were as easily processed as accented words in prosodic context.

To summarize, cognitive studies have established some basic claims about the nature of prominence,
but many details and elaborations of hypotheses concerning the role of prominence in language understanding require much more experimentation. The abstract versus concrete distinction leads to important and interesting differences in processing models, but absent detailed proposals concerning the representations and algorithms entailed by each experimental technique in each experimental design, it is difficult to judge between the two accounts. What is possible to conclude is that notwithstanding differences, the two views also share in common some main results that reinforce the findings in descriptive linguistic literature. While explanatory principles may differ, it seems that as a whole the experimental literature as well points to a crucial role for intonational prominence in lexical processing, reference resolution and semantic and discourse focusing. Specific experimental findings in each of these areas, however, have yet to be brought together in a unitary processing model.

2.3.2 Text-to-speech Systems

The majority of practical text-to-speech synthesis systems utilize lexical category information to determine the accentuation of words in text. The simplest strategy assigns prominence or accent to each content or open class word (i.e. nouns, verbs, adverbs and adjectives), while deaccenting function or closed class words. Based on corpus analysis, Altenberg (1987), Hirschberg (1990a) and Ross and colleagues (1992) proposed refinements in rules for accenting function words. Using fine-grained lexical categorizations, Altenberg defined closed-accented and closed-deaccented subclasses, and Hirschberg identified closed-accented, closed-deaccented and closed-cliticized classes, each including several to a dozen function word classes. Using a coarse-grained system, Ross and colleagues found for their corpus that quantifiers and negatives were often accented. Results of these corpus-based studies of lexical category information have been incorporated into pitch accent assignment systems for TTS. Several of these systems, reviewed below, consider additional factors affecting accentuation, ranging from information status to metrical constraints.

The experimental systems developed by Horne and colleagues (Horne, 1987; Horne et al., 1993) have tested various hand-crafted linguistic rules for assigning accent using syntactic, semantic and discourse information. In Horne’s (1987) system for English synthesis, a hierarchy of accentability for grammatical functions was established: object > subject > predicate. For each sentence, the grammatical positions were visited in the order given by the hierarchy. Accenting was determined for each constituent based on whether the root morph of the lexical head occurred previously in the discourse. If it did occur previously, the constituent was deaccented. If it did not, it was accented, and the height of the next accent to be potentially assigned was decreased by a set ratio. In later work on a Swedish synthesis system, Horne and colleagues (1993) extended the lexical identity notion of discourse-given/discourse-new to include relations of hyponomy, part-whole, situationally given items in the domain and synonymous relationships. These additional sources for determining given/new information status were determined by human analysis of the system domain of stock market reports. Relations found within a window of 60 words were coded and passed on to the prosodic processor for accent and phrase assignment. Function words were not eligible to receive accent. Otherwise, the coding of given information restricted the placement of focal accents in Swedish, which may fall only on new information, and guided the placement of non-focal accents. Specifically, synonymous referring expressions were eligible to receive non-focal accent, as was given information in certain phrasal configurations. These rules accord with cited psycholinguistic results for English and Dutch, although the presence of focal accents in Swedish alters the semantic focusing function of prominence compared to English and Dutch.

In the NewSpeak prosodic synthesis system (Hirschberg, 1990a; Hirschberg, 1993), pitch accent was assigned using a decision-tree system automatically trained on large corpora. Shallow text-processing techniques were used to automatically label lexical category, surface position, discourse given/new status and syntactically marked cases of contrast and semantic focus in the corpora. In particular, given/new status
was determined within a set window using identity of morphological stems as the criterion for givenness, similar to Horne’s system (1993). Hirschberg’s given/new modeling was novel in its use of two focusing mechanisms to determine information status. Hirschberg defined a static GLOBAL FOCUS SPACE that was said to contain the equivalent of Horne’s situationally given items, which were predicted to be deaccented throughout the discourse. Extending Silverman’s (1987) proposal to keep track of mentioned open-class word lemmas within paragraphs, Hirschberg also defined a dynamic LOCAL FOCUS SPACE that was updated as the discourse proceeded. In different experiments, the window size for determining both the global and local focus spaces was set to the size of the orthographic phrase, sentence or paragraph instead of a fixed number of words. The presence of cue phrases triggered an update of the local focus space as well. In this regard, Hirschberg’s focusing model was the first implementation of the discourse focusing mechanisms of Grosz (1977) in an accent assignment system. Experimentation showed that the linguistically based units enabled more accurate accenting using the given/new distinction: performance was best when the global focus space was set to the first sentence, and the local focus space was updated after every paragraph, which can be said to roughly correspond to a discourse segment. Features of the focusing model were combined with the other features in the CART learning system. Experimental results from this study are discussed in detail in Chapter 5. The decision trees for accent assignment produced by the CART system have been directly incorporated into a full text-to-speech system, the Bell Laboratories NewTTS system (Sproat and Olive, 1997).

Finally, Dirksen (1992) incorporated F-marking and focus projection into prosodic phrase structure representations, and modeled the effects of prosodic well-formedness rules, such as the rhythm rule governing stress shift, on these prosodic structures. Dirksen appealed to a simple notion of focus as discourse-given; his work is notable rather for its attempt to integrate focus-marking into a computational grammar of prosody that implements focus projection rules as well as metrical rules for deaccenting. His system, PROS-3, which was part of the POLYGLOT ESPRIT-project on multi-lingual text-to-speech, was written in the unification grammar formalism, using two separate grammatical processing phases. In the first phase, a parser identified the argument structure of sentences. Then, the syntactic parse was translated into a metrical structure annotated with F-markers. Prosodic rules then assigned prominence to the leaf nodes of the metrical tree, taking into account metrical rules and focus-marking.

### 2.3.3 Message-to-speech Systems

Message-to-speech synthesis has received considerably less attention than text-to-speech synthesis, despite the fact that many facets of intonational meaning have been theorized to exist at the pragmatic and discourse levels of language interpretation. While early message-to-speech synthesis systems mainly explored syntax constraints on prosody (Young and Fallside, 1979), later systems incorporated discourse modeling components and pragmatic information. Below are reviewed a handful of systems that address the linguistic properties of prominence in novel ways.

In the Direction Assistant system developed by Davis and Hirschberg (1988), a task grammar for describing routes was used to specify the hierarchical structure of synthesized driving directions. Certain contrastive cases of accenting, such as left versus right, were stipulated for the domain. The Direction Assistant relied on the task structure to determine the given/new status of discourse entities. Entities that occurred in the current or previous discourse segment, as determined by the hierarchical nature of the route grammar, were considered given and were marked for deaccenting. Lexical category information was used in default accent assignment rules.

Footnote 4: In Hirschberg’s (1993) study, the term, global focus space, corresponds to Grosz’ notion of a top-level focus space containing entities that are permanently accessible throughout the discourse (Grosz, 1994), and the term, local focus space, corresponds to Grosz’s (1977) term, focus space, in which the salient entities and relations in a segment are dynamically recorded.
House and Youd (1990) incorporated numerous higher-level effects on prominence into their SUNDIAL travel information system. In particular, they used a discourse model to represent discourse entities, events and relations, and roles connecting them. Discourse objects in the logical form expressing the message structure were analyzed as mutually known, belonging to the current task space or belonging to the same type of object as a more recently mentioned referent or another referent in the task space. The first class of objects was assigned negative focus; the second class, referring focus; and the final class, either referring or emphatic focus, depending on the dialogue act. Referring focus was used to trigger cases of accent marking contrast in the comparative sense. Dialogue act information was used, on the other hand, to determine corrective situations, or repeated attempts to elicit parameter values (i.e. emphasis as the highlighting of information already known to the speaker and hearer). Both of these situations warranted emphatic, focus accentuation in SUNDIAL, involving the accenting of the corrected information or parameter that was sought, and the deaccenting of the rest of the words in the sentence. The following is a case of correction,

You can’t fly from \textit{GATWICK} \textbf{emphasis} on Sunday.—

while the following is a case of emphatic repetition,

(1a) Where are you traveling to \ldots and where are you traveling from \ldots

(1b) And where are you traveling \textbf{[FROM]} \textit{emphasis}

Finally, fine-grained lexical category information was used to refine the assignment of default accentuation.

In the BRIDGE project on intonation in dialogue generation, Zacharski and colleagues (1992) proposed to account for four major factors influencing accentability: linear order, lexical category, ‘semantic weight’ and givenness. The notion of semantic weight was intended to explain the frequent deaccenting of words such as \textit{person}, \textit{thing}, \textit{matter}, \textit{place}, \textit{do} and \textit{go}. Also, the givenness hierarchy of Gundel and colleagues (1989) was proposed as the system for representing discourse status information. Gundel’s system defines strict correspondences between certain forms of referring expressions and givenness status. In an implementation of the proposals by Zacharski and colleagues (1992), Monaghan (1994) computed an accentability score by assigning integer values for each of three features — focus, givenness and word class — and multiplying these scores together. The determination of focus was based on the topic/comment structure of a sentence. Words in the topic were not in focus and got a score of 0, while words in the comment were in focus and received a 1. To score givenness, the givenness hierarchy of Gundel and colleagues (1989) was divided into four intervals, and integer values were associated with each interval starting with 1 for the most given interval. Finally, word class scores were as follows: nouns scored a 4; verbs, adjectives and adverbs, 3; prepositions, 2; all other items plus all semantically light words, 1. The accentability scores were computed and passed on to the accent placement module. This module assigned nuclear accent in any domain to the rightmost word with the highest accentability score. Pre-nuclear accents were assigned to a number of words with high accentability scores, depending on the speech rate setting. Metrical well-formedness rules were said to be involved in the assignment of pre-nuclear accents, but details were not described. Thus, the BRIDGE system implemented in a very different way the insights concerning lexical category information and attentional status captured in other systems. It also highlighted the notion of semantic weight as a contributing factor, which uniquely provides for the deaccenting of light content words in message-to-speech synthesis.

As can be seen from this survey of language processing research, there are areas of theoretical interest that have barely been tapped from cognitive and computational perspectives. Progress has been made in implementing many theoretical concepts in speech synthesis systems, and testing the cognitive bases for theories of prominence. However, there is a noticeable lack of computational work on the interpretation of prominence, which would encompass representations and algorithms for processing prominence. A central
outstanding issue is the integration of prominence processing with other NLP components in a speech understanding architecture. The speech synthesis work has begun to address this issue, by relating prominence to additional linguistic features, such as grammatical function and lexical category information. In addition, it has explored basic discourse processing models for determining discourse focus, and certain methods for modeling emphasis and contrast in a domain have been tested. Less progress has been made in modeling semantic focus information, which is stipulated in all of the synthesis systems, and operationalized as recent discourse mention or question focus in cognitive studies of prominence interpretation. As a result, in most studies, the line between semantic focusing and discourse focusing has been blurred.

2.3.4 Summary

As can be seen from this survey, linguistic and speech modeling research has provided a solid basis for integrating prominence into speech understanding systems. Prominence and prosodic recognition methods can supply a speech transcription with prominence locations with over 80% accuracy for professionally read as well as spontaneous speech. As for language processing research, on the generation side, several theoretical models for synthesizing prominence have been developed and tested in speech synthesis systems. Various message-to-speech synthesis systems have among them modeled the relationship of prominence to all of its associated linguistic factors, namely emphasis, contrast, semantic focus and given/new status. The SUNDIAL system (House and Youd, 1990), for example, used dialogue and discourse modeling to separately model contrast as comparison, contrast as correction and emphasis as the repetition of mutually known information. Accenting and deaccenting rules were defined for each of these discourse situations. Research in text-to-speech has tried to make use of correlations between the higher-level linguistic concepts associated with prominence and features of lexical, syntactic and surface structure, which are easier to compute. For example, the notion of given is often operationalized in TTS systems as the repetition of an identical lemma or morph within a certain text window, defined by a predetermined number of words or by a linguistic unit; words that are given are candidates for deaccenting.

In sum, based on the linguistic and speech modeling work, language processing research has contributed many findings on separate linguistic phenomena, but the pieces of a complete picture of prominence processing remains to be put together in models that rigorously define cognitive or computational representations and algorithms.

2.4 A Functional Prosody Perspective

This thesis develops an explanatory, computational language processing account of the discourse focusing nature of prominence, as a marker of the attentional status of discourse referents. This aspect of prominence interpretation has been investigated in theoretical linguistics, psycholinguistic experiments and synthesis systems, yet the findings from this research have not been fully integrated in computational models for prosodic interpretation and accent assignment in speech synthesis. In addition to the computational perspective, the problem of prominence interpretation is approached from a FUNCTIONAL PROSODY perspective. In the field of functional syntax, formal linguists develop grammars, rules and principles to explain the effects of discourse and pragmatic levels of meaning on the syntactic realization of a proposition or message. As noted by Kuno (1987), functional explanations must be integrated into existing grammatical theories of syntax, but functional syntacticians focus less on the properly syntactic and more on exploring phenomena that pose difficulties for the theories or are currently ignored or overlooked in theories despite empirical evidence to the contrary. The functional prosody perspective can be defined on analogy with functional syntax, where prosody — the grammatical structure of the sound of language, is substituted for syntax, the grammatical structure of the lexemes of language. Two major assumptions are involved in the framework
of functional prosody: (1) the functional nature of prominence can be examined directly through its effects on pragmatic and discourse interpretation, and (2) prominence can be viewed as a grammatical entity, belonging to prosodic structure. Although discourse focusing is one of many proposed functions of accenting, it is perhaps the most pervasive, constant communicative function served by accentuation. Emphasis and contrast, for certain, arise in a limited number of discourse contexts, although they may occur more regularly in dialogue speech where clashes in beliefs or discourse intentions among the discourse participants can give rise to the need to contradict information believed by other parties in the conversation, or to repeat propositionally given information to refocus attention on salient aspects of a speaker’s argument (Walker, 1993).

The characterization of the discourse focusing function of accent as pervasive and constant can be elaborated by considering how it underlies or supports its semantic focusing function. In this thesis, a view of the discourse focusing function of accent as an independent contributor to discourse modeling processes, such as resolving anaphoric references and updating the discourse model to reflect the dynamically changing attentional focus of the discourse participants, is developed in a manner that is compatible with Rooth’s (1992) theory of semantic focus interpretation and intonational focus. On Rooth’s theory, the discourse processes of anaphora resolution and presupposition formation are viewed as supplying features of the context model in which the interpretation of focus constructions takes place. In particular, the discourse model can be used to establish or rule out various existential presuppositions, or even propositional presuppositions if the discourse model records events and relations, as well as roles that connect them, as outlined in the global focusing model of Grosz (1977). Rooth’s (1992) theory, therefore, offers one possible framework for understanding the relationship between semantic focusing and discourse focusing via the contribution of the latter processing to the former.

Rooth’s theory also draws a boundary between the role of intonational focus information, which he views as the phonological realization of the F-feature considered in focus projection theories (Gussenhoven, 1984; Selkirk, 1993), and discourse focusing. Specifically, Rooth interprets F features or F-marking within his theory of focus interpretation, separating them from discourse focusing entirely, contra Selkirk (1993). Rooth leaves aside the question of how to formulate rules describing the exact acoustic-prosodic nature of intonational focus-marking but implies that his concept of intonational focus may correspond to the nuclear accent. What is important is that his theory does not constrain intonational focus-marking based on the given/new status of discourse entities REALIZED\(^5\) in the sentence. Rather, there is a role for discourse focusing to play in supplying anaphoric and presuppositional information as part of a formal semantic model of discourse context. Finally, Rooth also claims that his theory subsumes accounts of focus as contrast, which would accord with theoretical proposals on contrastive stress as NARROW (versus BROAD) focus (Ladd, 1979b) and with results from Bartel and Kingston’s (1994) perceptual studies, which failed to show that listeners reliably associate a contrastive interpretation to words marked by the L+H* (Pierrehumbert, 1980) or SCOOP accent. It has been suggested in previous work that this accent conveys contrastiveness (Pierrehumbert and Hirschberg, 1990).

The discussion of Rooth’s theory shows how most of the communicative functions proposed for prominence can be interrelated through semantic focus interpretation processes. It is clear, however, that Rooth’s theory of focus interpretation does not explain how discourse focusing is to be modeled, or how prominence contributes to its interpretation. The semantic and discourse notions of focus, while distinct, are also intimately related, because what has already been said in a discourse often becomes the background, sentence topic or theme that is paired with a new semantic focus in a new discourse contribution. A handful of proposals have addressed the dual functions of prominence as a marker of both semantic and discourse

\(^5\)Throughout this thesis, the term, realize, is used in a technical sense to mean the expression of lexical information that can be said to refer to a discourse entity in a discourse model (Sidner, 1979; Grosz, Joshi, and Weinstein, 1995). Thus, realization of a discourse entity in a discourse model is to be distinguished from reference to an actual entity in the real world.
focus (Hirschberg and Pierrehumbert, 1986; Selkirk, 1993). One approach to reconciling these functions, taken by Selkirk, for example, is to view prominence as an F-marker of words in a syntactic structure and to project F-marking from the words to phrases and dominating heads according to focus projection rules, while appealing to information status to constrain the possibilities for F-marking at the word level. This approach has potential, but in its current formulation, it simply pushes the problem from the derivation of focus from prominence cues, to the interpretation of semantic focus from intonationally licensed F-marking.

Another way of achieving reconciliation is to explore whether properties of prominence itself reveal the specific function or functions that it may serve. This approach is suggested by Hirschberg and Pierrehumbert’s theory of intonational meaning (1986; 1990). They hypothesize that the absence or presence of accent may mark discourse focus, while the tonal properties of the accent itself convey semantic focus properties. Defining discourse focus as the attentional state representation of discourse entities in a discourse model, they claimed that accent placement must be determined based on the local discourse context or segment in which the accentable item occurs (Hirschberg and Pierrehumbert, 1986). In later work, they proposed that intonational prominence marks information that the speaker intends to be predicated, or added to the mutual beliefs held between speaker and hearer (Pierrehumbert and Hirschberg, 1990). Specifically, they hypothesized that the low-star accents, L*, L*+H and H+L*, mark information that is intended to not be predicated in the mutual belief model shared between speaker and hearer, while the high-star accents, H* and L+H*, convey that the accented information is to be predicated. They commented that the H*+L accent is an exception and can be used to mark given information, or at least information that is expected by the hearer. Finally, they allowed that given information may be accented to convey semantic properties such as contrastiveness. Unlike Selkirk’s theory (1993), this theory makes no attempt to define formal syntactic constraints on the interpretation of accent function. Rather, it embeds hypotheses about the function of prominence within a compositional theory of intonational meaning. Reflecting the formal hierarchical structure of Pierrehumbert’s (1980) intonational grammar, the compositional theory relates the meaning of three levels of prosodic structure — pitch accents, phrase accents and boundary tones — in a unified framework of discourse interpretation based on Grosz and Sidner’s (1986) theory of discourse structure.

Hirschberg and Pierrehumbert’s (1986) theory of intonational meaning was the first to hypothesize a specific association between accent placement and Grosz and Sidner’s attentional state model. Since this work, several theoretical proposals for treating prominence in local focusing algorithms have been made (Cahn, 1990; Kameyama, 1994). These proposals, however, lack empirical validation. This thesis represents the first extensive empirical study of how prominence may be interpreted in a computational discourse model. In addition, it develops original principles of prominence interpretation that generalize across the two levels of discourse focusing defined in the Grosz and Sidner model, namely global focusing at the segment level and local focusing at the utterance level. While no prior study has used both focusing models to explain the discourse focusing function of prominence, discourse processing research on anaphora, cue phrase disambiguation and many other phenomena has demonstrated the need for both top-down or segment-based and bottom-up or utterance-based systems of discourse processing (Moore and Pollack, 1992; Hobbs, 1995). As this thesis demonstrates, without examining the interactions of these two qualitatively distinct focusing mechanisms, much empirical data on prominence cannot be explained.

Finally, following the approach of functional syntax, this thesis attempts to integrate hypotheses concerning the discourse focusing function of accent into accounts of other linguistic factors that influence prominence. Two features that play a significant role in attentional focusing algorithms, grammatical function and form of referring expression, are considered at length in an empirical study of prominence in the spontaneous narrative monologue, reported in Chapter 3. Chapters 4 through 7 investigate the relation of prominence to a range of lower-level linguistic features, such as lexical category and surface position, as well as higher-level features, such as the previously studied grammatical function and form of referring expression features as well as new syntactic and discourse features. The contributions of these diverse features
to accent prediction for read and spontaneous direction-giving monologues are evaluated using machine learning techniques to build classification models for various prominence classes.

In sum, this study represents an original attempt to integrate prominence directly into the discourse structure theory of Grosz and Sidner, contributing to a better understanding of how to process prominence in the language understanding component of speech understanding systems. The accent prediction experiments in this thesis represent a first attempt to integrate the new account of the discourse functions of prominence with existing work on numerous other linguistic factors known to affect accent placement. Research in this direction can improve the modeling of prominence for speech synthesis.
Chapter 3

Narrative Study

This chapter develops a new understanding of the discourse focusing nature of prominence by drawing on distributional data from a spontaneous narrative monologue. Results from two distributional studies on the spontaneous narrative monologue motivate the formulation of a new principle defining the discourse focusing role of prominence. In the first study, the given/new study presented in Section 3.1, several existing hypotheses relating syntactic and lexical factors to the given/new interpretation of prominence are explored. Results show that accent cannot be accurately predicted by either grammatical function or form of referring expression, or the combination of the two features. The second study, the attentional modeling study presented in Section 3.2, builds on the findings of the first study but examines discourse focusing factors directly. The study uses an independently motivated taxonomy of given/new information status, based on the attentional state model from the computational theory of discourse structure proposed by Grosz and Sidner (1986). The attentional state model, which is described in Section 3.2, provides an enriched taxonomy of the given/new information status of discourse entities in a dynamic discourse model. Using the new taxonomy, reanalysis of the role of prominence in conveying discourse focusing information is undertaken for the narrative data in Section 3.3. An account of the discourse focusing role of accent function is developed based on detailed distributional analysis of the discourse functions of accent in six linguistic configurations. The new, discourse-based interpretation of prominence identifies separate but interacting contributions of grammatical function, form of referring expression and prominence in conveying the attentional status of a discourse referent. Section 3.4 formulates principles expressing the contributions of each factor, and illustrates how the principles may be implemented in high level discourse processing algorithms for reference resolution and the modeling of attentional state. The algorithms show that the interpretation of prominence in discourse is not essentially separate from that of other nonprosodic linguistic factors. All of the examined grammatical factors serve to cue inferences in discourse processing, such as marking changes in attentional state and establishing relations among referents. As discussed in Section 3.5, the new account explains more empirical data and offers a more general interpretation of the discourse focusing nature of prominence than do previous accounts. Further, this study represents the first attempt to apply the complete attentional state model of Grosz and Sidner. A review of related findings and hypotheses shows that the prominence interpretation principle unifies and reconciles many previous findings in the theoretical, psycholinguistic and computational literature. Finally, directions in which to extend the findings of the narrative study are discussed in Section 3.6.
3.1 The Given/New Study

3.1.1 Hypotheses

As reviewed in Chapter 2, corpus-based studies of prominence have provided evidence for the claims that grammatical function, form of referring expression, lexical category and surface position influence the assignment of prominence in discourse. This first study of the spontaneous narrative monologue examines the interactions of several of these factors through their roles in conveying given/new information status. Previous research has shown there is a general tendency for given information to be unaccented and new information to be accented (Brown, 1983; Terken, 1984). It is also generally thought to be the case that the information status of pronoun referents is given, and that the information status of proper names and full noun phrases is new (but see Section 3.3). Finally, there is a general tendency for grammatical subjects to represent given information and grammatical direct objects, new information (Prince, 1988).

In both computational and cognitive models of prominence processing, reviewed in Chapter 2, there are few explanatory accounts of how factors such as grammatical function and form of referring expression should be weighed together to determine accentuation. While statistical (Hirschberg, 1993) and other quantitative (Monaghan, 1994) methods for combining the contributions of these and other factors have been implemented in speech synthesis systems, attempts to relate these two factors in a theory of prominence interpretation have achieved limited results. Brown (1983) directly investigated the correlations of prominence with grammatical function and with form of referring expression and concluded:

*If syntactic and intonational forms are both regarded as criteria for ‘givenness’, these forms may supply contradictory information to the hearer . . . We considered whether, in principle, it is possible to establish a taxonomy of information status independently of the forms of expression used by speakers, and we concluded that it is not possible. It seems that our only safe access to information status is provided by the form of the expressions* (Brown and Yule, 1983).

The first study in this chapter reexamines the issue raised by Brown and Yule (1983) concerning the interpretation of conflicting cues to information status. Specifically, the study tests the general claims about the accentuation of given/new information that predict that (1) pronouns, as given information, will be unaccented, and full noun phrases and proper names, as new information, will be accented, and that (2) subjects, which tend to convey given items, will tend to be unaccented, and direct objects, which tend to convey new items, will tend to be accented. The narrative study provides data challenging both of these basic hypotheses, which will be referred to as the form of expression hypothesis and the grammatical function hypothesis respectively.

3.1.2 Analysis

The spontaneous narrative speech examined in this chapter consists of twenty minutes of American English speech, obtained from an adult male speaker using sociolinguistic interview techniques. An excerpt of the spontaneous narrative monologue appears in Figure 3.1.

From the narrative speech, 481 animate noun phrase referring expressions were analyzed for accentuation, grammatical function (e.g., subject, direct object, object of preposition) and form of referring expression (e.g., proper name, pronoun, definite/indefinite noun phrase). Two accentuation classes, PROMINENT and NONPROMINENT, are defined. The accentuation of an expression is said to be prominent if its head and/or an internal modifier is marked by either H* or a complex pitch accent in Pierrehumbert’s (1980)

---

1The narrative was collected by Virginia Merlini for the purpose of studying American gay male speech and was made available by Mark Liberman at the University of Pennsylvania Phonetics Laboratory.
...and Masson was a con man
he’s one of those slick greasy kind of people who
weasel their way into your life
and Eissler—being an old man—wanted to unload this
cuz he wanted to retire and die okay
analysts are very very strange people
they’re very closely knit in other words
they will not talk to you if you have not been analyzed
they will not talk to anybody who is not an analyst alright...

Figure 3.1: Excerpt from the spontaneous narrative monologue.

system of English intonation. An expression not meeting these criteria is said to be nonprominent.² For
the present analyses, therefore, L* accented expressions and deaccented expressions are conflated in the
nonprominent category.³ Accentuation judgments were made by the author by listening to the speech and
by examining the pitch tracks, amplitude waveform, and spectrographic information in some cases. All
speech analysis was performed using Entropic Research Laboratories WAVES+ software (Talkin, 1989) on
SUN workstations.

3.1.3 Results

Distributional analyses consider the role of accentuation in relation to grammatical function and lexical form
of referring expression. The analyses focus on the two most frequent classes of grammatical function for all
of the noun phrase referring expressions, subject versus direct object, and the two largest classes of lexical
forms, pronouns versus explicit forms, based on distributions in our narrative. Explicit forms, also referred
to as full forms, include proper names, definite NPs and indefinite NPs.

Overall results in Table 3.1 show that explicit forms are generally prominent and pronouns generally non-
prominent, providing indirect support for the hypothesis that new information (conveyed by explicit forms)
is generally prominent and given information (conveyed by pronouns) generally nonprominent. Although
this trend is significant (p<.00001, χ²=69.5, df=1), the hypothesized given/new information conveyed by
accentuation on 20% (40/200) of the referring expressions in Table 3.1 conflicts with predictions of the form
of referring expression hypothesis.

Overall results also show that subjects are somewhat more likely to be prominent than direct objects,
although this trend is not significant (p<.3, χ²=1.3, df=1). These distributions nonetheless run counter to
the grammatical function hypothesis that subject position, which is the preferred grammatical function for
old information, is less likely to hold a prominent expression than is direct object position.

In fact, closer examination of the narrative data reveals an asymmetric interaction of grammatical func-
tion and lexical form. As shown in Table 3.1, grammatical subjects are prominent only 23% (25/111) of

² In analyzing proper names, an additional complication arises due to the independent lexical status of first and last names. For
the study, a proper name was considered prominent if either the first or last name bore a H* or complex pitch accent. Similarly,
issues concerning NP-internal accent placement were not addressed as the narrative contains relatively few internally modified
complex NPs.

³ A more systematic study of the possible discourse uses of L* versus deaccenting remains as future work. A total of six tokens,
all explicit forms, were labeled with the L* accent. Based on this small sample, clear distinctions could not be made between the
discourse functions of L* and deaccenting.
Table 3.1: Percentage of prominent referring expressions.

<table>
<thead>
<tr>
<th>PERCENT PROMINENT</th>
<th>Subject</th>
<th>Direct Object</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pronouns</td>
<td>23%</td>
<td>7%</td>
<td>21%</td>
</tr>
<tr>
<td>Explicit forms</td>
<td>91%</td>
<td>55%</td>
<td>81%</td>
</tr>
<tr>
<td>TOTAL</td>
<td>45%</td>
<td>34%</td>
<td>43%</td>
</tr>
</tbody>
</table>

Table 3.2: Table of subject expressions, showing relation between prominence and form of expression.

<table>
<thead>
<tr>
<th>SUBJECT EXPRESSIONS</th>
<th>Prominent</th>
<th>Nonprominent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pronouns</td>
<td>25</td>
<td>86</td>
</tr>
<tr>
<td>Explicit forms</td>
<td>49</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 3.3: Table of direct object expressions, showing relation between prominence and form of expression.

<table>
<thead>
<tr>
<th>DIRECT OBJECT EXPRESSIONS</th>
<th>Prominent</th>
<th>Nonprominent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pronouns</td>
<td>1</td>
<td>14</td>
</tr>
<tr>
<td>Explicit forms</td>
<td>11</td>
<td>9</td>
</tr>
</tbody>
</table>

Terken (1984) investigated some of the same distributional trends in Dutch task-oriented monologues. He reported that 11% of 544 referring expressions exhibited conflicting accentuation and form of referring expression (i.e. accented pronouns and deaccented full forms), and the interaction of form of expression and accentuation in his data was significant. Further, he reported a significant interaction of grammatical function...
(subject versus non-subject) with accentuation. However, additional analysis showed that for sentence-initial position, there was no such significant interaction. Because Dutch is a scrambling (‘free’ word-order) language and English is not, the interactions of grammatical function and accentuation may differ in the two languages. The interaction of grammatical function, form of expression and accentuation was not analyzed by Terken and has not been noted in any other previous studies. Additional factors need to be studied to make sense of this new observation.

3.2 The Attentional Modeling Study

3.2.1 Hypotheses

The results of the first narrative study refute the conjecture that form of referring expression provides safe access to information status, and that accentuation is a redundant marker of information status (Brown and Yule, 1983). The discovery of systematic interactions between accentuation, grammatical function and form of referring expression raises two new questions: (1) what factors govern the accenting of subject pronominal expressions and why are they more frequently accented than direct object pronouns? and (2) what factors govern the deaccenting of direct object full forms and why are they more frequently deaccented than subject full forms? The notion of discourse-old/discourse-new status is of little value in predicting accentuation for the problematic classes of accented subject pronouns and deaccented direct object full forms, since all of the pronouns in the narrative, subject or otherwise, refer to discourse-old entities, while fewer than one half of the prominent direct object explicit forms refer to discourse-new entities.

To shed light on these questions, the taxonomy of information status provided by Grosz and Sidner’s (1986) attentional state model is applied in a reanalysis of the distributional data. This second study of the narrative speech addresses two issues. One is that the hierarchical structure of discourse influences accent decisions in systematic ways. Previous studies have shown, for example, that the accentuation of referring expressions can be correlated with topic or discourse structural properties (Terken, 1984), and that taking discourse structure into account can improve the performance of pitch accent assignment algorithms (Hirschberg, 1993). While the Grosz and Sidner attentional state model has been hypothesized by Hirschberg and Pierrehumbert (1986) to be relevant to understanding accentuation and has been used to determine accent placement in speech synthesis systems (Davis and Hirschberg, 1988; Hirschberg, 1993), no previous study has employed this taxonomy in distributional analysis of prominence in a speech corpus. The second issue addressed in this study is how to relate the influence of hierarchical structure on accentuation to the influence of local structure on accentuation. In this study, local structure is investigated by continued examination of the utterance feature, grammatical function.

3.2.2 Discourse Structure Analysis of the Narrative

The hierarchical, segment-based structure of discourse and the local, utterance-based structure are treated by two separate attentional focusing models in the Grosz and Sidner (1986) theory of discourse structure. Grosz and Sidner (1986) claim that discourse structure is made up of at least three interrelated components: INTENTIONAL STRUCTURE, LINGUISTIC STRUCTURE, and ATTENTIONAL STATE. In their theory, a discourse is comprised of DISCOURSE SEGMENTS whose hierarchical relationships are determined by intentional structure and realized by linguistic structure. Given a hierarchical discourse structure, attentional focusing processes are modeled at two qualitatively distinct levels. At one level, the GLOBAL level, the relationships between segments affect the accessibility of discourse entities realized in different segments of a discourse. At another level, the LOCAL level, the relationships between utterances affect the accessibility of discourse entities realized in successive utterances within a single discourse segment. An overview of
the Grosz and Sidner discourse structure theory and discussion of its application in discourse analysis of the narrative are presented below.

**Intentional and Linguistic Structure**

A discourse can be divided into hierarchically structured discourse segments, each of which can be assigned a **DISCOURSE SEGMENT PURPOSE** (DSP) that describes the communicative goal for that segment. The DSP is intended by the speaker to be recognized by the hearer (Grosz and Sidner, 1986). Relationships among segments in the linguistic structure can be analyzed based on the relationships of their DSPs within a collaborative plan (Lochbaum, 1994). Two DSPs can be related by the **DOMINANCE** or **SATISFACTION**-**PRECEDENCE** relation. The dominance relation expresses that the achievement of one DSP contributes to the achievement of another DSP, while the second relation expresses that the achievement of one DSP enables the achievement of another.

Discourse structure analysis of the narrative sample was undertaken based on speaker intention (Grosz and Sidner, 1986), following procedures defined and utilized by Grosz and Hirschberg (1992). The intentional structure of the spontaneous narrative was analyzed by specifying a hierarchy of DSPs. The narrative was divided into a hierarchy of discourse segments based on these DSPs. Cue phrases or discourse markers also served as cues to segmental structure. The narrative is essentially a purposeful monologue, in which the speaker tries to persuade the interviewer to read a biography of Sigmund Freud. The speaker describes the life stories of various personae from this biography, explaining how they became involved in the Freudian circle of analysts and their activities within it. The genre of narrative may exhibit complex turns of topics and events, many of which the listener may not be able to predict before the time of hearing. While structural choices of syntactic configuration and form of referring expression can be readily detected and widely agreed upon, analysis of the topic or discourse structure poses greater difficulties and uncertainties. Nevertheless, analysis of a completed narrative often reveals a clear intentional structure.

**Attentional State**

The attentional status of the referring expressions in the narrative was determined based on the discourse segmentation and the global and local focusing models defined by Grosz and Sidner (1986). As noted, in Grosz and Sidner’s discourse model, discourse processing proceeds at two levels, global and local. The global level concerns the hierarchical structuring of discourse segments, while the local level concerns the linear structuring of utterances within a single segment. In this framework, the notion of salience is formalized by attentional focusing mechanisms that are claimed to underlie discourse processing in general. The attentional state component dynamically records the entities and relations that are salient at the two distinct levels at any point in the discourse.

The global level of attentional state is modeled as a last-in first-out **FOCUS STACK** of **FOCUS SPACES**, each containing representations of the entities and relations salient within the discourse segment corresponding to the focus space (Grosz, 1977). The focus space at the top of the focus stack is termed the **IMMEDIATE FOCUS SPACE**. Pushes and pops of focus spaces obey the hierarchical segmental structure of the discourse. An empty focus space is pushed onto the stack when a segment begins; entities are recorded in the immediate focus space as the discourse advances until the discourse segment closes and its focus space is popped from the focus stack. Those entities represented in the top focus space on the stack are in **IMMEDIATE GLOBAL FOCUS**. Entities represented elsewhere on the stack are in **NON-IMMEDIATE GLOBAL FOCUS** and are said to be less accessible as referents than entities in the immediate focus space.

It is important to note that there are three discourse segment boundary types that each entail specific manipulations of the focus stack. First, there is the initiation of an embedded segment, or subsegment, corresponding to the push-only move of a focus space onto the focus stack. When an embedded segment

30
Figure 3.2: Illustration of global focusing mechanisms.

opens, a new focus space is pushed onto the stack, on top of the focus space of its embedding segment. Second, there is the completion of an embedded segment, which entails the pop-only move. Upon the close of an embedded segment, the focus space of the embedded segment is popped from the top of the focus stack. The focus space of the embedding segment becomes the immediate focus space, and entities within it are said to be immediately accessible. Finally, there is the transition between sister segments corresponding to the pop-push move. In this case, one segment ends and its focus space is popped; immediately, a new focus space is pushed for the next sister segment.

Figure 3.2 illustrates the push-only and pop-only manipulations of global focus on an excerpt from the spontaneous narrative. At the end of the fourth line, “he was married to Martha and knocked up his sister-in-law”, the global focus stack contains one focus space with contents as shown in the left diagram. When Segment B begins, a new focus space is pushed as represented in the center diagram. At the end of Segment B, its focus space is popped and the stack holds only the previous focus space for Segment A as shown in the right diagram. When the pronoun he is encountered after the pop, the entities in the preceding embedded segment are no longer on the focus stack and are therefore not available for pronominal reference. The pronoun instead refers to an entity in the focus space of the outer segment that is resumed, namely Freud.

The local level of attentional state is modeled by CENTERING mechanisms (Sidner, 1979; Joshi and Weinstein, 1981; Grosz, Joshi, and Weinstein, 1983; Kameyama, 1985; Brennan, Friedman, and Pollard, 1987; Grosz, Joshi, and Weinstein, 1995). All of the salient discourse entities realized in an utterance are recorded in the partially ordered set of FORWARD-LOOKING CENTERS (Cf list) for that utterance. The Cf list specifies a set of possible links forward to the next utterance. The ranking of the Cf list members reflects their relative salience in the local discourse context at the time of the uttering of that particular utterance. Each utterance also has a single BACKWARD-LOOKING CENTER (Cb), which is the most central and salient discourse entity that links the current utterance to the previous utterance. The Cb of an utterance $U_n$, $Cb(U_n)$, is defined as the highest-ranking member of the Cf list of the prior utterance, $Cf(U_{n-1})$, that is realized in $U_n$. Centering theory currently stipulates that for English the Cf list members are ordered based
on grammatical function (subject > direct object > indirect object > other arguments) (Grosz, Joshi, and Weinstein, 1995) and surface position (Gordon, Grosz, and Gilliom, 1993). The highest ranking member of the Cf list is called the PREFERRED CENTER (Cp), and is the most likely candidate to become the Cb of the following utterance (Brennan, Friedman, and Pollard, 1987).

Figure 3.3: Illustration of local focusing (centering) mechanisms.

Figure 3.3 illustrates how centering operates on a sequence of utterances from the narrative. The narrative excerpt is centrally about Freud, who remains the Cb throughout. Centering constructs are computed within each discourse segment.
Analysis of the narrative study data requires rudimentary definitions of the four attentional statuses provided by the Grosz and Sidner model: primary local focus or Cb, secondary local focus or Cf members (exclusive of the Cb), immediate global focus, and non-immediate global focus. To review, a discourse entity may be globally salient by virtue of its being represented on the focus stack in either the immediate focus space or a non-immediate focus space. Entities in the immediate focus space are claimed to be relatively more accessible than those in focus spaces deeper in the stack. At the local level, the Cb is claimed to be more salient than non-Cb members of the Cf list. Psychological evidence for this claim is provided by anaphor resolution experiments (Hudson-D’Zmura, 1988; Gordon, Grosz, and Gilliom, 1993). These constraints on accessibility may be exploited in reference resolution algorithms that search through the most accessible referents to the least accessible referents to resolve reference (Sidner, 1979; Grosz, 1977). For the narrative study, the attentional status of referring expressions was determined by hand at the global and local levels, using the discourse segmentation and the centering rules as defined in (Grosz, Joshi, and Weinstein, 1995).

It should be emphasized that most existing theories of information status or focus of attention do not assume a fundamental qualitative distinction between global and local salience. Thus, Grosz and Sidner’s attentional model contrasts with theories of given/new information status that rest upon the notion of shared knowledge of discourse participants (e.g., Prince’s (1981) taxonomy of shared knowledge), that stipulate successive grades on a single scale of givenness (e.g., Gundel and colleagues’ (1989) scale of familiarity) or that treat the entire discourse as a monolithic global focusing space (e.g. Prince’s (1988) reexamination of her taxonomy of shared knowledge in terms of the notions of discourse-new and discourse-old). By defining segment-to-segment manipulations of global attentional state and utterance-to-utterance manipulations of local attentional focus, the attentional model of Grosz and Sidner characterizes notions of discourse salience relative to a dynamically unfolding and folding record of mutual beliefs about the discourse at two qualitatively distinct levels of linguistic structure.

### 3.3 A Discourse-Based Interpretation of Accent Function

It is proposed that the distributions of pitch accent in the narrative study can be explained by the communicative functions that accent serves in various linguistic configurations, as summarized in Table 3.4. The proposed functions of accent have in common that each serves to manipulate the dynamic record of the activated or salient entities in a discourse model. Unlike previous accounts, in the proposed account, accent functions are defined at both the local and global levels of discourse structure. This discourse-based interpretation of accent function provides an overarching framework that unifies and reconciles previous findings on interpreting prominence, while contributing original analyses of the problematic accentuation classes uncovered in the given/new study. This analysis of the discourse functions of prominence makes two basic claims: first, the meanings conveyed by choices of form of referring expression and syntactic structure are separate but interacting; and second, the role of accentuation must be interpreted against the background of these choices in linguistic expression. The principles of accent interpretation are supported by detailed analyses of the discourse structural properties of referring expressions in the six linguistic configurations specified in Table 3.4.

In addition to the six configurations in Table 3.4, two more configurations are logically possible, namely nonprominent subject explicit forms and prominent direct object pronouns. The narrative contained six cases of the former (out of 54 subject explicit forms total) and one of the latter (out of 15 pronoun direct object forms total). These distributions led to the formulation of the following constraint in earlier work (Nakatani, 1993):

- If grammatical function and form of referring expression convey conflicting GIVEN/NEW statuses,
Table 3.4: Discourse functions of prominence.

<table>
<thead>
<tr>
<th>Functions of Prominence</th>
<th>Discourse Function</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Linguistic Factors</strong></td>
<td><strong>Discourse Function</strong></td>
</tr>
<tr>
<td>SBJ pronoun</td>
<td>Shift local attention to new $C_b$</td>
</tr>
<tr>
<td>SBJ explicit form</td>
<td>Introduce new global referent as $C_p$</td>
</tr>
<tr>
<td>DOBJ explicit form</td>
<td>Introduce new global referent</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Functions of Nonprominence</th>
<th>Discourse Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOBJ explicit form</td>
<td>Maintain referent in global focus</td>
</tr>
<tr>
<td>DOBJ pronoun</td>
<td>Maintain non-$C_b$ referent in local focus in $Cf$ list)</td>
</tr>
<tr>
<td>SBJ pronoun</td>
<td>Maintain $C_b$ referent in local focus as current $C_b$</td>
</tr>
</tbody>
</table>

then accentuation must “reinforce”, or agree with the (preferred) GIVEN/NEW status conveyed by the form of referring expression.

This constraint echoes the mentioned proposal by Brown and Yule (Brown and Yule, 1983). However, while Brown and Yule’s remark was simply an observation, the attentional modeling analysis in this chapter constitutes an explanatory account of how this constraint arises. In particular, using the principles of this study, the configurations that are dispreferred by this constraint can be analyzed as exhibiting anomalous attentional properties. That is, independent psycholinguistic findings have identified the REPEATED NAME PENALTY effect, namely that proper names in subject position are read more slowly than pronouns when they corefer with the $C_b$ of the previous sentence (Gordon, Grosz, and Gilliom, 1993). On the analysis stated in Table 3.4, the presence of prominence on subject explicit forms signals the introduction of a new candidate for $C_b$. It is consistent with the finding of the repeated name penalty to conclude that this shift in attention would not be strongly cued by the use of nonprominence on the subject explicit form, leading to the anomalous nature of this class of expressions.

Similarly, prominent direct object pronouns could be analyzed as attentionally anomalous by combining a principle of centering theory concerning the realization of locally focused expressions with the analyzed discourse functions of prominence on pronominal forms. The so-called Rule 1 of centering theory (Grosz, Joshi, and Weinstein, 1995) is stated:

**RULE 1**: If any element of the $Cf$ list of the preceding utterance, $Cf(U_n)$, is realized by a pronoun in the current utterance, $U_{n+1}$, then the $C_b$ of the current utterance, $C_b(U_{n+1})$, must be realized by a pronoun also.

It follows from this rule that if there is a single pronoun in the current utterance (realizing an entity already in local focus), it must be the $C_b$. The discourse functions of prominence on pronominal expressions in Table 3.4 state that prominence on subject pronouns indicates a shift in local attention to a new $C_b$, while nonprominence on either subject or direct object pronouns indicates that the realized entity is already in local focus. Given the strong preference for placing the $C_b$ in subject position (Gordon, Grosz, and Gilliom, 1993), it can be argued that when shifting centers, placing the new $C_b$ in subject position is especially helpful to listeners who must identify the new $C_b$. The use of prominence is an additional cue for local attentional processing signaling to listeners that a shift in centers has occurred. The anomalous nature of prominent direct object pronouns arises from the fact that speakers do not often realize local shifts in $C_b$s, which are often cued by intonational prominence on reduced forms, by placing the new $C_b$ in direct object position.
Further empirical investigation is needed to test the pragmatic and discourse properties of tokens that do arise in the two allegedly attentionally anomalous configurations. It is suggested by the narrative data that semantic uses of prominence such as contrast and discourse phenomena such as top-level global focus may be related to the occurrences of prominent direct object pronouns and nonprominent subject explicit forms respectively.

The discourse functions of prominence in six linguistic configurations have been fully analyzed. The detailed distributional analyses show that accentuation serves to signal changes in the attentional state of a discourse in systematic ways. The analyses examine the role of prominence first in local attentional modeling and then in global attentional modeling.

3.3.1 Manipulations of Local Focus: Pronouns

Subject pronouns

Of the 25 cases of prominent subject pronouns, sixteen cases can be viewed within the framework of attentional state modeling as falling into two major classes of center shifts. Of the remaining nine cases, six were marked by emphasis or contrast and three required limited inference to determine the pronoun referent.

For the first class of center shifts (seven cases), accentuation signaled reference to a previous discourse center that was not salient in the immediate discourse context. For this class, intonation marks a global center shift, which accompanies a focus space pop upon the completion of an embedded discourse segment or sister segment. Accent in these cases signals reference to a previous center that was not realized in the immediately preceding utterance, and therefore was not a member of the Cf list of the immediately preceding utterance. An excerpt of the narrative displayed in Figure 3.4 illustrates this global center shift signaled by accent on the pronoun, she, in line 2. In all of the figures displaying narrative excerpts in this chapter, segmentation structure is indicated by white space at segment boundaries and indentation showing the hierarchical relationships among segments. The accented pronoun in line 2 refers to Anna, who was not

Anna was confused because Swales was feeding her misinformation about Masson saying Mrs. you know, ah Professor Freud . . .

uh this Masson is no good
he’s going to discredit your father [line 1]

okay and SHE was very confused [line 2]
she was an old woman

Figure 3.4: Shifting centers: Example of prominent subject pronoun (in capitals) signaling global shift in centers.

mentioned in the immediately preceding utterance, line 1, but was the Cb at the beginning of the segment continuing line 1. The shift in attention back to Anna at the beginning of the sister segment in line 2 is partially cued by the accented pronoun. Prominent pronouns signaling global shifts occurred as the first grammatical subject following the completion of an embedded or sister discourse segment.

For the other class of center shifts signaled by accented subject pronouns (nine cases), accentuation marked a local shift in attention away from the current discourse center to a new discourse entity that was indeed salient, but not centrally so, in the immediate discourse context. In contrast to the first class, the pronoun referent always occurred in the Cf list of the immediately preceding utterance, but never as the
Cb. An excerpt of the narrative displayed in Figure 3.5 illustrates this kind of local center shift. In line 2

*So Masson became the new curator.*

*He flies to London and, you know,*

*he’s already met Anna Freud and therefore*

*he has access to the secret cupboard of Freudian letters*

*and naturally Anna assumed that uh [line 1]*

*SHE was a brilliant woman too – [line 2]*

*she did more a lot of work in child psy- psychiatry and psychoanalysis*

---

Figure 3.5: Shifting centers: Example of prominent subject pronoun (in capitals) signaling local shift in centers.

of Figure 3.5, the accented pronoun, *she*, realizes an entity that is in the local context given in line 1. The accented subject pronoun helps establish Anna Freud as the center of the embedded subsegment beginning at line 2.

In contrast to the class of prominent subject pronouns, the majority of nonprominent subject pronouns referred to the Cb of the last utterance (73%). Several examples of nonprominent subject pronouns continuing the Cb occur in the excerpt in Figure 3.6. A smaller subclass of nonprominent subject pronouns

*so MASSON became the new curator*

*he flies to London and, you know,*

*he’s already met Anna Freud and therefore*

*he has access to the secret cupboard of Freudian letters*

---

Figure 3.6: Maintaining Cb focus: Example of nonprominent subject pronoun (in boldface) continuing the Cb.

(14%) realized the Cb of a **NEIGHBORING SEGMENT** on the focus stack, which is defined in this study as the segment whose focus space was in immediate global focus at the time the focus space for the current segment was pushed onto the focus stack. The neighboring segment, therefore, can be either a sister segment or an embedded segment, whose focus space has just been popped off of the stack, or an embedding segment, whose focus space was in immediate focus at the start of the current segment and remains on the stack beneath the current focus space. The **NEIGHBORING FOCUS SPACE** is defined as the focus space of the neighboring segment. Interestingly, eight out of the 12 cases in this subclass of nonprominent subject pronouns continued the Cb of an outer segment upon the closing of an embedded segment. Therefore, while there may be a tendency to signal with accent a global shift in centers as defined above, it is clearly not necessary to do so. Psychological testing is needed to determine whether accent plays a facilitative role in cuing center shifts of this sort. For the remaining 13% (11 cases) of nonprominent subject pronouns, which do not continue or resume a Cb, three tokens occur in repairs, two in dialogue tags, and one in response to an interruption from the interviewer.

**Direct object pronouns**

The nonprominent occurrences of direct object pronouns (14/15) fell into three subclasses: inter-sentential anaphora (three cases), where pronominalization of an NP in a subordinate clause is licensed by established
syntactic rules; multiple pronouns (five cases), where the direct object pronoun occurs in addition to a subject pronoun realizing the Cb; and objects of verbs of perception (five cases), where it is hypothesized that the lexical semantics of the verb (e.g. *see, approach*) actually renders the direct object as the Cp or most likely candidate to become the most salient entity in the continuation of the discourse (Grosz, Joshi, and Weinstein, 1995). Figure 3.7 provides an example of nonprominent direct object pronouns in multiple pronoun sentences. In the excerpt, the plural pronoun, *they*, realizes the Cb throughout the segment,

```
so they got together and
they went “Mr. Masson?”
boom
they kicked him out
they ostracized him
```

Figure 3.7: Maintaining non-Cb focus: Example of nonprominent direct object pronouns (in boldface) in multiple pronoun sentences.

satisfying Rule 1 of centering theory, which states that the Cb must be realized by a pronoun if any other entity in the sentence is realized by a pronoun.

The final token of the nonprominent direct object pronouns occurs in a repair, while the single prominent direct object pronoun token, shown in Figure 3.8, occurs in a contrastive context. In the discourse context

```
inside the Freudian archives
she interviewed Masson and wrote a book
a very short one, I think I have it here, it’s called –
and now he’s suing her because he’s claiming
she misquoted me
but nooo I don’t think so
it sounds like HIM [line 1]
because I heard him on a talk show
```

Figure 3.8: Contrastive usage: Example of prominent direct object pronoun (in capitals).

in Figure 3.8, the speaker is asserting in line 1 that the person quoted sounds like Masson and not someone else, as Masson is claiming.

Given the distributions of direct object expressions, it can be concluded that centering theory, enhanced with inter-sentential rules and provisions for lexically marked classes of verbs, is sufficient to account for the attentional role of lack of accent on direct object pronouns. Accent on direct object pronouns may reflect semantic aspects of meaning such as contrast, as in the single example in the narrative, or perhaps emphasis. More data are needed to explore this hypothesis.

### 3.3.2 Manipulations of Global Focus: Explicit Forms

**Direct object explicit forms**

Analysis of the direct object explicit forms shows that accentuation is generally determined by the global focus state. Eighty percent (16/20) of these expressions are cases of first mention of an entity in the current
discourse segment, although three quarters of these (12/16) are also references to discourse-old entities.

In the narrative, intonational prominence marks the introduction of entities into a segment’s focus space; crucially, these entities do not occur in neighboring focus spaces, i.e. the immediately preceding sister segment or embedded segment, or the immediately embedding segment. An example of a prominent direct object explicit form appears in Figure 3.9. In Figure 3.9, a new character, Eissler, is introduced with a

You wanna hear the story about Paul Masson . . .
in his thirties, late thirties, he decided to be psychoanalyzed
and he was, and he became very interested in Freud
and so he read, and
he jumped into the Freud world, the Freudian world, like I jump into a pot of pasta, okay

And he approached Paul [sic] EiSSLER [line 1]
head of the Freudian archives based in New York...

Figure 3.9: Introduction of new global referent: Example of prominent direct object explicit form (in capitals).

prominent proper name in direct object position in line 1. Eissler is not referred to in the preceding segment and in fact is mentioned here for the first time in the narrative.

In contrast to introducing new referents with prominence, lack of intonational prominence marks the introduction or reintroduction of entities in RECENT GLOBAL FOCUS. Recent global focus is defined in this study as a new class of globally focused entities, namely the entities in either neighboring focus or immediate focus. To be in recent focus, the reintroduced entity must have been previously introduced in the immediate outer or embedding segment, the immediately preceding sister segment, the embedded segment that was recently popped or the current segment. References to entities that are either in the embedding focus space on the focus stack, or in the most recently popped focus space, do not require accentual prominence. References to entities that are not in recent global focus, on the other hand, are reintroduced into a focus space by prominent expressions. In short, the choice of accentuation on direct object proper names reflects the global salience of the corresponding entity.

An example of a nonprominent direct object explicit form appears in Figure 3.10. In Figure 3.10, the

Peter Swales
he’s another interesting character in this little drama
was a hippie in the sixties...
and all of a sudden he realized that his interest was in psychoanalysis
and again, like Masson, or Maason, some people call him Maason, jumped into it like I would
in a... I don’t know what, I said pasta before, I’ll say something else

and, but you see, Anna Freud didn’t like Peter Swales [line 1]
(Is she dead?) No she was sh- yes she’s dead...

Figure 3.10: Maintenance of immediate global referent: Example of nonprominent direct object explicit form (in boldface). An interruption from the interviewer occurs in parentheses.
central character of the preceding sister segment, Swales, is reintroduced in line 1 with a nonprominent proper name at the start of a new segment that centrally concerns Anna Freud.

Finally, the two cases that were not first mention since a discourse segment boundary both occurred during the speaker’s reading from a book; there is no account offered for these last two cases at this time.

Applying the principles of analysis described above and summarized in Table 3.4 for direct object explicit forms, the accentuation for 90% (18/20) of the direct object explicit forms occurring in the narrative is correctly predicted. These results surpass those of alternate strategies based on notions of given/new used in previous studies, such as accenting all and only cases of discourse-initial mentions (13/20 or 65% correct) or accenting all and only cases of segment-initial mentions (11/20 or 55% correct).

Subject explicit forms

The class of explicit forms in subject position shares the property that a majority of expressions (44/54 or 81%) are first-mentions in the current discourse segment. Of the ten out of 54 cases that are not segment-initial, three occur in repairs and three in quoted contexts. This lends support to the generalization that explicit lexical forms are used to signal the global newness of a discourse referent. However, fully 91% (49/54) of subject explicit forms are prominent, compared to 55% for direct object explicit forms. Intonational prominence in these cases signals not only the global newness of a referent, but also a shift in centers from that of the previous segment to the new subject referent.

Figure 3.11 shows several examples in a single excerpt of how a prominent subject explicit form introduces a new Cp, that often becomes the Cb for the utterance. This analysis is also consistent with the

...and MASSON was a con man
he’s one of those slick greasy kind of people who
weasel their way into your life

and EISSLER – being an old man – wanted to unload this
cuz he wanted to retire and die okay

ANALYSTS are very very strange people
they’re very closely knit in other words
they will not talk to you if you have not been analyzed
they will not talk to anybody who is not an analyst alright...

Figure 3.11: Introduction of global referent as new Cp: Examples of prominent subject explicit forms (in capitals).

mentioned repeated name penalty result showing a delay in reading times when a proper name is used in subject position to realize an established Cb (Gordon, Grosz, and Gilliom, 1993). Under this analysis, prominence on subject explicit forms, like prominence on subject pronouns, allows the hearer to infer a local attentional shift.

Given this discourse function of introducing a new global referent as a new preferred center, 44 out of 49 prominent subject explicit forms are explained. Of the five remaining prominent expressions, two occur in quoted contexts, one occurs in a repair, one occurs as an argument of a verb-of-telling and the final one
involves a set-member coreference relation. Of the five nonprominent explicit subject forms, one occurs in a quoted context, one in a repair and one in an interruption. The remaining two expressions refer to characters in the narrative, Freud and Masson, who may be said to be in top-level global focus.

Finally, independent evidence for this analysis of prominence and global focusing has been discovered in recent studies of Italian (Avesani, 1996) and of Japanese (Venditti, 1996).

### 3.3.3 Evaluation

The new theory of the discourse functions of prominence can be quantitatively evaluated against the narrative data. For ease of reference, the distributional results of the first, given/new study, are redisplayed in Table 3.5. The proposed discourse functions of prominence resulting from the second study are redisplayed in Table 3.6.

<table>
<thead>
<tr>
<th>PERCENT PROMINENT</th>
<th>Subject</th>
<th>Direct Object</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>%</td>
<td>N</td>
<td>%</td>
<td>N</td>
</tr>
<tr>
<td>Pronouns</td>
<td>23%</td>
<td>25/111</td>
<td>7%</td>
</tr>
<tr>
<td>Explicit forms</td>
<td>91%</td>
<td>49/54</td>
<td>55%</td>
</tr>
<tr>
<td>TOTAL</td>
<td>45%</td>
<td>74/165</td>
<td>34%</td>
</tr>
</tbody>
</table>

Table 3.5: Distributional results of given/new study.

<table>
<thead>
<tr>
<th>FUNCTIONS OF PROMINENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Linguistic Factors</strong></td>
</tr>
<tr>
<td>SBJ pronoun</td>
</tr>
<tr>
<td>SBJ explicit form</td>
</tr>
<tr>
<td>DOBJ explicit form</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FUNCTIONS OF NONPROMINENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Linguistic Factors</strong></td>
</tr>
<tr>
<td>DOBJ explicit form</td>
</tr>
<tr>
<td>DOBJ pronoun</td>
</tr>
<tr>
<td>SBJ pronoun</td>
</tr>
</tbody>
</table>

Table 3.6: Discourse functions of prominence.

Results of the analyses of the distributional data are summarized in Table 3.7, which gives breakdowns of the number of expressions accounted for by the prominence interpretation principles and by other factors. Considering all eight possible linguistic configurations, 86.5% (173/200) of the referring expressions in the spontaneous narrative are explained directly by the attentional modeling account of the discourse functions of prominence. If performance factors, such as repairs and interruptions, and special discourse situations, such as direct quotations and items in top-level global focus, are included in the number of expressions that are accounted for, then the analyses in this study account for 96.5% (193/200) of the referring expressions. For the targeted configurations of prominent subject pronouns and nonprominent direct object full forms,
which have posed problems for previous accounts, all of the data are accounted for, largely in terms of the attentional modeling definitions of the functions of prominence.

### 3.4 Algorithms for Processing Prominence in Discourse

This study of spontaneous narrative monologue speech provides evidence that intonational prominence is one source of linguistic information contributing to discourse processing decisions. The new attentional modeling theory of the discourse functions of prominence, summarized in Table 3.6, may be implemented in message-to-speech systems, in which discourse structure and meaning are directly encoded, for the purpose of conveying meaningful contrasts in prominence. The principles may also be implemented in text-to-speech synthesis systems that heuristically model attentional state. However, no previous theory of the meaning of prominence, has allowed for a principled integration of prominence with other non-prosodic grammatical features in both local and global attentional state modeling. It is instructive, therefore, to integrate the results of the attentional modeling study into a computational account of prominence interpretation. High level discourse processing algorithms, presented in this section, address this deficit in current speech understanding systems. They provide a partial specification for the attentional state processing component of a natural language understanding system that treats prominence information in tandem with syntactic and lexical information. These algorithms show how prominence information may be used for reference resolution and attentional state modeling, thereby motivating work on prominence recognition for speech understanding.

#### 3.4.1 Principles

The algorithms are based on several principles governing the discourse processing of referring expressions in spoken language. The first concerns the role of form of expression information in the attentional processing of referring expressions.

- **The Lexical Form of a Referring Expression** indicates the level of attentional processing, i.e., pronouns involve local focusing while full lexical forms involve global focusing.

This principle was formulated in earlier work on centering theory (Grosz, Joshi, and Weinstein, 1983) and is strongly supported by data in the narrative study. It reflects the primacy of lexical form of referring expression in determining attentional status (Brown, 1983; Gundel, Hedberg, and Zacharski, 1989; Prince, 1988), *inter alia*. It is also consistent with the spirit of Levy’s (1984) cognitive analysis of how surface forms create discourse cohesion or indicate discontinuity in ongoing discourse.

However, lexical form is not the sole determinant of attentional status. In addition to Prince’s (1988) finding that subject position tends to hold given information, while non-subject positions tend to hold new information, early work on centering theory (Grosz, Joshi, and Weinstein, 1983; Kameyama, 1985) has shown that grammatical function contributes to the ranking of the Cf list as well as to the identification of the Cb. This leads to the formulation of the second principle.

- **The Grammatical Function** of a referring expression reflects the local attentional status of the referent, i.e., subject position generally holds the highest ranking member of the Cf list, while direct object holds the next highest ranking member of the Cf list.

The third and final principle treats intonational prominence and represents an original contribution of this study.

- **The Intonational Prominence** of a referring expression serves as an inference cue to shift attention to a new Cb, or to mark the global (re)introduction of a referent; **Nonprominence** serves as an inference cue to maintain attentional focus on the Cb, Cf members or global referents.
While previous research has shown a relationship between intonational prominence and discourse salience, the current study is unique in its consideration of attentional state processing at both the global and local levels, leading to principles that nevertheless generalize across both levels of discourse structure.

3.4.2 Algorithms

The attentional modeling study suggests new and more specific ways in which accent information contributes to language understanding. These results are incorporated in two algorithms, one for processing full lexical forms, shown in Figure 3.12, and the other for processing pronominal expressions, shown in Figure 3.13. The algorithms embody the two major claims of this study, that accent serves as an inference cue, marking the preservation of and changes in attentional status, and that accent function must be interpreted against a dynamic background of the attentional state model as well as additional linguistic factors, following the aforementioned principles for lexical form, grammatical function and intonational prominence. The algorithms, which can be viewed as two parts of a single whole, also reflect the primacy of form of referring expression in determining the discourse processing of referring expressions. The different cases of referring expressions (e.g., intonationally prominent subject pronoun) receive different treatments with respect to referent search and update of the attentional state. The input that is assumed for the algorithms consists of a referring expression, marked for form of expression, grammatical function and intonational prominence; and the current attentional state, represented by the focus stack at the global level and the Cb and the Cf list at the local level. The output of the algorithm is an updated attentional state, with referential indices capturing the referential act of the processed expression. Recent focus spaces, defined in Section 3.3, include the current focus space as well as the neighboring focus space.

To demonstrate the application of the algorithms, the processing of a few examples of accented subject pronouns that occur in the narrative is explained. The first example appears in Figure 3.14. The pronoun he in line 3 of the text, alright HE was human too, bears a H* prominence. To interpret this pronoun, the processing steps for prominent subject pronouns given in Figure 3.13 are followed. The Cb of the previous utterance, line 2, is Minnie and the Cf list is \{Minnie, abortion\}. The first test, clause 1(a)i in Figure 3.13, fails due to the gender conflict between the pronoun he and the Cb of the previous utterance, Minnie. The second test, clause 1(a)ii, also fails because the Cf list of the previous utterance does not contain a male referent. So the third clause, 1(a)iii, is tried and applied successfully. The focus space for Segment B is popped and the pronoun refers to the entity that is the Cb of the immediately prior utterance of the embedding segment, namely Freud from line 1. Finally, the local focusing structures are updated, with Freud at the head of the Cf list.

Two more examples of H* accented subject pronouns are presented in Figure 3.15. The first accented pronoun, in line 2, exemplifies clause 1(a)ii in Figure 3.13. The Cf list of the prior utterance, line 1, contains Freud, but not as the Cb. So, the pronoun refers to an entity already in local focus but not primarily so. Prominence on the subject pronoun in line 2 shifts attention to Freud as the central character in a subsegment that elaborates on line 1. A new focus space is pushed for this subsegment and Freud is entered as the new Cb. The next subject pronoun in line 3 refers to the Cb of the previous utterance. As suggested by the clause 3.8:1(a)i, the context can be viewed as emphatic (corroborated by an increase in acoustic energy in this case). Indeed, analysis of the intentional structure shows that the asserted proposition, that Freud is considered infallible, is central to the speaker’s argumentation in this story and is expressed at several different points in the narrative.
If expression is FULL lexical form, then

1. If expression is SUBJECT, then
   (a) If expression is intonationally PROMINENT, then
      i. Create new discourse entity $e_i$
      ii. Push new focus space
      iii. Add entity $e_i$ and information predicated of it to new focus space
      iv. Add entity $e_i$ to Cf list as Cp
   (b) If expression is intonationally NONPROMINENT, then
      i. Search for referent $e_i$ in immediate and then neighboring focus spaces, checking the parallel
         subject referent of the previous utterance first.
      ii. Add entity $e_i$ and information predicated of it to immediate focus space (if not already
         there)
      iii. Add entity $e_i$ to Cf list as Cp

2. If expression is OBJECT, then
   (a) If expression is intonationally PROMINENT, then
      i. Create new discourse entity $e_i$
      ii. Add entity $e_i$ and information predicated of it to immediate focus space
      iii. Add entity $e_i$ to Cf list after entity realized by subject
   (b) If expression is intonationally NONPROMINENT, then
      i. Search for referent $e_i$ in immediate and then neighboring focus spaces
      ii. Add entity $e_i$ and information predicated of it to immediate focus space (if not already
         there)
      iii. Add entity to Cf list after entity realized by subject

Figure 3.12: Attentional state processing for full lexical forms.
If expression is PRONOMINAL form, then

1. If expression is SUBJECT, then

   (a) If expression is intonationally PROMINENT, then
       i. If expression refers to Cb of previous utterance, then check for emphatic or contrastive context and add entity $e_i$ to current Cf list as Cp
       ii. Else, search Cf list of previous utterance and if it contains referent $e_i$, then push new focus space and add entity $e_i$ to current Cf list as Cp
       iii. Else, pop focus space and check coreference of entity $e_i$ and Cb of new immediate focus space; add entity $e_i$ to current Cf list as Cp

   (b) If expression is intonationally NONPROMINENT, then
       i. If expression refers to Cb, or other member of Cf list of previous utterance, then add referent $e_i$ to current Cf list as Cp
       ii. Else, search Cf list and if it does not contain referent $e_i$, then pop focus space and check coreference of entity $e_i$ and Cb of new immediate focus space; add entity $e_i$ to current Cf list as Cp

2. If expression is OBJECT, then

   (a) If expression is intonationally PROMINENT, then
       i. Search Cf list of previous utterance for referent $e_i$
       ii. Add entity $e_i$ to current Cf list after entity realized by subject
       iii. Check for contrast relationship with parallel object referent, or inferable entity, in previous utterance

   (b) If expression is intonationally NONPROMINENT, then
       i. Search Cf list of previous utterance for referent $e_i$
       ii. Add entity to current Cf list after entity realized by subject

Figure 3.13: Attentional state processing for pronominal forms.
... so Freud had a few affairs with Fliese
so big deal you know what I’m saying
he knocked up Minnie Bernais
he was married to Martha and knocked up his sister-in-law [line 1]
CB=Freud, CF=Freud, Martha, Minnie

and they gave her hey-
she had an abortion in one of these clap [line 2]
CB=Minnie, CF={Minnie, abortion}

alright HE was human too [line 3]
CB=NONE, CF={Freud}

Figure 3.14: Processing example of prominent subject pronoun.

They all put Freud on a pedestal [line 1]
CB=ANALYSTS(=THEY), CF={ANALYSTS, Freud, pedestal}

HE is an icon okay [line 2]
CB=Freud, CF={Freud}
HE can do no wrong [line 3]
CB=Freud, CF={Freud}

Figure 3.15: Additional processing examples of prominent subject pronouns.

3.5 Related Work

The evaluation of the discourse-based interpretation of accent function in Section 3.3.3 offers a quantitative picture of how well the new proposals explain the discourse focusing nature of prominence. In this section, qualitative comparisons are made between the findings of this study and numerous results from previous empirical investigations of accent function and related phenomena. Discussion is organized around three major hypotheses from prior work, namely that accentuation communicates given/new information status, topic-hood and relative salience.

Given/New

In a study of task-oriented speech, Brown (1983) found that given information was rarely accented (less than 5% of tokens). She failed to identify any any principled role to grammatical function for determining accentuation. In a study of elicited speech in which the lexical and syntactic forms of utterances were controlled, Terken and Hirschberg (1994) demonstrated that structural properties, such as the persistence of grammatical function and surface position, contribute to the determination of accent for given information.
So, while the above two studies did not address the problem of how grammatical function and form of referring expression information interact with prominence, the proposed theory of prominence interpretation does not address surface position information. Further distributional analysis of utterances exhibiting word order variation in the studied corpora may shed light on this problem. These findings could be related to results from Terken and Hirschberg’s study and tested on new data.

**Topic-hood**

The role of accentuation in topic-marking was investigated by Terken (Terken, 1984). In his instruction-giving monologues, Terken found that the first introduction of a discourse topic was generally accented and subsequent references to the topic were often expressed as unaccented pronouns. In contrast, non-initial references to non-topical entities were often expressed as accented full forms. Terken’s first observation accords with the large distributions of prominent subject explicit forms (serving as topical or Cp introductions) and of nonprominent subject pronouns (generally continuing or resuming the Cb). His second observation accords with the proposed analyses of direct object explicit forms. The vast majority of these refer to entities that were previously introduced into the discourse, yet were not topics or Cbs.

The attentional modeling theory of accent function takes Terken’s account two steps further. First, a significant correlation is identified between grammatical position, form of expression and accentuation. Second, the theory offers more fine-grained proposals about factors determining accentuation on direct object explicit forms, specifying a set of conditions on the global attentional state that appears to license the accenting as well as deaccenting of direct object explicit forms in discourse. In short, the lack of intonational prominence on non-topical referents is determined by the recent global focusing history. These cases of nonprominent direct object explicit forms cannot be explained by Terken’s analysis, which predicts that non-topical expressions are prominent irrespective of recent global focusing history or grammatical function information.

Terken (1984), Fuchs (1984), Cahn (1990), Gundel and colleagues (Gundel, Hedberg, and Zacharski, 1989) and others have made observations about the topic shifting function of accented pronouns. For example, Cahn proposed that items may be H*-accented to reinstantiate the Cb after a discourse push or pop (Cahn, 1990, pp. 15-16); and Terken noted that “the use of accented pronouns immediately after topic introduction seems to be [a] way to confirm for the listener that there has been a topic shift” (Terken, 1984, p. 70). Gundel and colleagues (1989) proposed that across many languages, stressed pronouns can realize ‘activated’ entities, while unstressed pronouns can realize only ‘in focus’ entities at the center of attention. The new theory casts these notions more precisely by identifying two distinct sets of conditions on the local and global attentional state that appear to license the accenting of subject pronouns in narrative discourse.

**Relative Salience**

The association between accent placement and Grosz and Sidner’s attentional structure was first theorized by Hirschberg and Pierrehumbert (1986). In particular they claimed that,

> Just as salience is always determined relative to some particular context, accent placement must be determined with respect to the segment in which the accentable item appears.... [It] is the signaling of salience relative to the discourse segment that produces the secondary effects of given-new distinction, topic-hood or contrastiveness, and the favoring of one reference resolution over another (Hirschberg and Pierrehumbert, 1986, p. 14).

Later theoretical work more closely examined the relationship between accenting and local focusing (Cahn, 1990; Kameyama, 1994). These studies focused on the problem of accented pronouns from a theoretical perspective. So, the analyses of the role of accentuation in attentional modeling in these studies are not as
general as the one presented here and are not empirically validated. This study also builds on Hirschberg and Pierrehumbert’s original proposal by defining more detailed interactions among intonational prominence and other linguistic factors known to contribute to the determination of discourse salience, making use for the first time of both local and global focusing mechanisms to specify discourse salience.

In addition, an important distinction can be made between the proposals arising from the narrative study and Kameyama’s (1994) analysis of accented pronouns, called the Complementarity Hypothesis. The major claims of the Complementarity Hypothesis are that there are underlying preferences for the hypothetical unaccented version of a pronoun, and that these preferences can be used to derive the interpretation preferences for a given accented pronoun. In particular, attentional constraints apply during the computation of the interpretation preferences for the hypothetical unaccented pronoun. Once these preferences are determined, the preferences for the accented pronoun are simply computed by reversing the original preference list of the hypothetical unaccented version. This processing, it is important to note, takes place in Kameyama’s framework of TOTAL PRAGMATICS. This means that there are additional semantic, pragmatic, and world knowledge constraints contributing to the processing. While this richness is lacking in the proposals in this paper, it is clear that accent information plays a fundamentally different role in Kameyama’s account. In the relevant component of her algorithm, accenting information is considered subsequent to attentional constraints. In contrast, the current algorithms incorporate accent information directly into attentional state modeling at the same stage of processing as syntactic and lexical information.

In addition to pronoun studies, the accentuation of proper names has also been investigated by prosody researchers. It has been suggested that proper names should often be unaccented because their very use presumes familiarity with the named entity on the part of the hearer (see discussion by Ladd (1980, p. 91)). On the other hand, Hirschberg (1993) noted that the accenting of given proper names in a large speech corpus could be explained by the proposal by Sanford and colleagues (1988) that proper names may be used to refocus the speaker’s attention on previously established discourse entities that are not salient in the immediate discourse context. The preponderance of segment-initial mentions among explicit forms of all types seems to accord with this idea. Such reintroductions were found to often bear pitch accent in news speech (Hirschberg, 1993). The narrative data support an analysis in which the choice of lexical form is independent of, but related to, the choice of accentuation. Use of an explicit form indicates newness relative to the immediate focus space, while accentuation reflects newness relative to the recent focus spaces.

In this regard, the analysis of direct object proper names poses a problem for Grosz and Sidner’s stack model of global focusing, since the focus space of a sister segment is popped from the focus stack before its sibling focus space is pushed, and the focus space of an embedded segment is popped upon its completion. As mentioned in Section 3.2.2, entities in a focus space that is popped from the focus stack are claimed to be unavailable for reduced reference. However, the proposed analysis can be reconciled with the Grosz and Sidner model if this constraint is relaxed to allow entities in the most recent immediate focus space to remain globally salient. This precise modification has been implemented in the global focusing mechanisms of a message-to-speech system (Davis and Hirschberg, 1988, p. 142). Further research is required to determine whether the representation of the recent focusing history should include the focus space for only the linearly preceding discourse segment, as suggested by the narrative study, or should be extended further back in the discourse. These questions should also be considered in relation to an alternative model for capturing linear precedence effects in discourse focusing, the cache model proposed by Walker (1996).

The narrative study findings also extend psycholinguistic results on the processing of accent information. As discussed in Chapter 2, Terken and Nooteboom (1987) conducted psycholinguistic experiments to test the role of intonational prominence in discourse processing by humans. They concluded that reference resolution proceeds differently for accented and unaccented expressions, namely that listeners assume that an unaccented expression refers to a member of a “restricted set of activated entities” in the discourse context, while the interpretation of an accented expression is not constrained in this manner (Terken and
Nooteboom, 1987, p. 148). They further hypothesized that mental representations of entities were constructed differently for accented and unaccented referring expressions. They proposed that representations were built bottom-up for accented items, allowing the hearer to use the content of the expression to resolve the reference. In contrast, unaccented expressions were said to be resolved top-down, by taking as candidates for reference resolution the restricted set of activated entities. These two hypotheses contradict each other in the cases of accented pronominal forms and unaccented full forms (Terken, 1995). That is, for an accented pronoun, the limited semantic content of the pronoun makes reference resolution by strictly bottom-up processing implausible. Similarly, for unaccented full forms, the use of semantically contentful lexical items is unnecessary (and unexplained) if reference resolution proceeds strictly top-down.

The spirit of Terken and Nooteboom’s proposals is maintained in the prominence processing algorithms, while the remaining contradictions are removed by considering two levels of attention, each providing a distinct set of activated entities for reference resolution. For full forms, the relevant set of activated entities is formalized in terms of the structured contents of the global focus stack. For pronominal expressions, the relevant set of restricted entities is formally cast in terms of the centering constructs computed at the local level of discourse processing. That is, the lexical content of a referring expression signals its discourse structural level of salience, while accentual prominence conveys further shades of given/newness within the appropriate local and global focusing structures.

The discourse-based interpretation of accent function subsumes previous hypotheses concerning given/new information status, topic-hood and relative salience. To summarize, the narrative study contributes the following findings: first, accentuation cannot simply be associated with form of referring expression or grammatical function, but rather makes an independent contribution to the structuring of information in discourse, contra Brown (1983). Second, the distinction between local and global levels of attention can be used to make precise the linguistic principles governing the interpretation of prominence and nonprominence in discourse. The generalization emerges that intonational prominence marks the newness of discourse entities at either the local or global level, or both (as for prominent subject explicit forms), while nonprominence serves to maintain focus on entities in either local or global focus. Against the appropriate background of grammatical and lexical properties, accent may cue precise inferences for reference resolution and attentional state modeling.4

3.6 Discussion

In contrast to previous computational and cognitive processing proposals, the principle of prominence interpretation arising from the narrative study can be used directly for both assigning prominence in speech generation and interpreting prominence during discourse processing. As spoken language systems become more widespread, it is critical to develop models for processing prominence that make maximal use of prosody to interpret language in context and to convey meaningful spoken messages. To date, an adequate computational model of the discourse focusing nature of prominence has been lacking in theories of the meaning of prominence as well as in speech synthesis models for assigning prominence. The two studies presented in this chapter integrate prominence into a computational model of attentional processing based on the Grosz and Sidner theory of discourse structure. By analyzing the processing of prominence in a computational discourse modeling framework, a much richer, explanatory account of the discourse functions of prominence and nonprominence has been constructed.

---

4A similar inference-cuing role for accent was proposed in Hirschberg and Ward (1991a) for a very different reference phenomenon. Hirschberg and Ward found that accentuation patterns on anaphors in the source clause in VP-ellipsis constructions affected the interpretation of target clause anaphors in systematic ways: accenting served to flip preferences from either strict to sloppy or sloppy to strict readings of the target clause, against the background of underlying lexico-semantic preferences for either a strict or sloppy reading.
To build on the findings of this study, one could investigate additional linguistic factors related to prominence as well as additional aspects of prominence itself. The following issues must be addressed by further research. First, sparse data in the narrative corpus did not allow for thorough analyses of the cases of prominent object pronouns and nonprominent subject full forms. Thus, the treatment of these cases is based as much on observations in the literature as on the corpus analysis. It is hypothesized that these configurations may be associated with the phenomena of contrastive and emphatic accent, as well as special semantic focus phenomena deriving from parallel syntactic conditions. Similarly, it is hypothesized that emphatic or contrastive contexts license prominence marking for subject pronouns that refer to the Cb of the previous utterance instead of shifting attention to a new Cb. Second, whether pitch accent type further specifies the discourse functions of prominence is a long-standing yet intriguing question that should be empirically examined. The narrative monologue contained too few tokens with complex pitch accents to draw general conclusions. Finally, there are more factors to be studied and more fine-grained distinctions to be made among classes of referring expressions. For example, surface order and lexical category are two factors known to influence accentuation. Research on a new multi-speaker corpus of spontaneous task-oriented direction-giving monologues, described in the remainder of this thesis, addresses several of these outstanding issues.
<table>
<thead>
<tr>
<th>Subject Pronouns (N=111)</th>
</tr>
</thead>
<tbody>
<tr>
<td>25 prominent 23%</td>
</tr>
<tr>
<td>16 shift in Cb</td>
</tr>
<tr>
<td>6 contrast</td>
</tr>
<tr>
<td>3 emphasis</td>
</tr>
<tr>
<td>86 nonprominent 77%</td>
</tr>
<tr>
<td>75 continue or resume Cb</td>
</tr>
<tr>
<td>3 repair</td>
</tr>
<tr>
<td>2 dialogue tag</td>
</tr>
<tr>
<td>1 interruption from interviewer</td>
</tr>
<tr>
<td>5 unaccounted for</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Direct Object Pronouns (N=15)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 prominent 7%</td>
</tr>
<tr>
<td>1 contrast</td>
</tr>
<tr>
<td>14 nonprominent 93%</td>
</tr>
<tr>
<td>10 maintain non-Cb in Cf list</td>
</tr>
<tr>
<td>3 inter-sentential anaphora</td>
</tr>
<tr>
<td>1 repair</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Explicit Forms (N=54)</th>
</tr>
</thead>
<tbody>
<tr>
<td>49 prominent 91%</td>
</tr>
<tr>
<td>44 introduce new global referent as Cp</td>
</tr>
<tr>
<td>2 quoted context</td>
</tr>
<tr>
<td>1 repair</td>
</tr>
<tr>
<td>2 unaccounted for</td>
</tr>
<tr>
<td>5 nonprominent 9%</td>
</tr>
<tr>
<td>2 top-level global focus</td>
</tr>
<tr>
<td>1 quoted context</td>
</tr>
<tr>
<td>1 repair</td>
</tr>
<tr>
<td>1 interruption from interviewer</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Direct Object Explicit Forms (N=20)</th>
</tr>
</thead>
<tbody>
<tr>
<td>11 prominent 55%</td>
</tr>
<tr>
<td>11 introduce new global referent</td>
</tr>
<tr>
<td>9 nonprominent 45%</td>
</tr>
<tr>
<td>7 maintain referent in global focus</td>
</tr>
<tr>
<td>2 quoted context</td>
</tr>
</tbody>
</table>

Table 3.7: Coverage of narrative data.
Chapter 4

Task-oriented Monologue Study:
Part I

Various hypotheses emerged from the narrative study, including new generalizations about the discourse functions of accent and new interpretations of accentuation in specific linguistic configurations. However, these hypotheses were developed on a single-speaker narrative corpus, and need to be put to the test by applying them to more spoken language data. In particular, how do the accentuation principles apply to a wider variety of grammatical functions? Do they generalize to language with richer argument structures, such as sentences with multiple complements and adjuncts? Similarly, how do they apply to new forms of referring expressions, such as compound nouns, deictics and demonstrative noun phrases? How do the features studied — grammatical function, form of referring expression and discourse structural constraints — relate to other linguistic features that interact with accentuation, such as intonational phrasing and surface position information? Finally, do the narrative study results transfer to a new genre or reflect peculiarities of narrative, or perhaps, even more narrowly, of stories about people that are rich in animate noun phrase reference and pronoun usage? The claims of the narrative study can be exercised in all of these directions by investigating old and new linguistic features in speech of a different genre.

At another level, the findings of the narrative study can be made more robust by applying more rigorous empirical methodologies, at various stages from data collection to data analysis. Specifically, it would be advantageous to examine multiple speakers’ speech as well as to collect discourse segmentations from multiple subjects and ensure that the segmentation data are statistically reliable. Likewise, the robustness of prosodic data can be improved by having multiple labelers. Since the time the narrative analysis was undertaken, empirical methodology in both discourse and prosodic analysis has advanced. Various protocols for collecting segmentation data have been tested, and techniques for measuring their reliability explored (Hirschberg and Grosz, 1992; Flammia and Zue, 1995; Swerts, 1995; Carletta, 1996; Hirschberg and Nakatani, 1996; Pas-sonneau and Litman, 1997). Similarly, the ToBI prosodic transcription standard, described in Chapter 1, was recently developed by an interdisciplinary team of researchers interested in the creation and use of prosodically labeled speech databases (Beckman and Ayers, 1994; Beckman and Hirschberg, 1994). Inter-labeler reliability has been measured for test groups of ToBI labelers, which included labelers who learned the transcription standard exclusively from the devised tutorial and training materials (Silverman et al., 1992; Pitrelli, Beckman, and Hirschberg, 1994).

A third direction in which to extend the narrative study would be to integrate existing work on lower-level linguistic features with findings about how grammatical function, form of referring expression and attentional status contribute to pitch accent prediction. Features such as lexical class or part-of-speech information have long been identified as useful, but not perfect predictors of accentuation in spontaneous speech. The problem of how higher-level and lower-level linguistic features combine can be fruitfully
explored by experimentation with pitch accent assignment systems that utilize these various features. While important insights can be gained by building a pitch accent assignment system with hand-crafted linguistic knowledge sources and rules (Horne, 1987; Davis and Hirschberg, 1988; Zacharski et al., 1992; Horne et al., 1993), Hirschberg (1993) and others have shown that machine learning techniques can be applied productively to quantitatively study the interactions of accentuation and a wide array of linguistic features. Results from machine learning experiments not only offer direct guidance for constructing pitch accent systems, but also provide focus for distributional analyses and motivate algorithms and architectures for processing intonational prominence in speech understanding systems in general.

This chapter describes a new corpus study that provides the data needed to expand on the narrative study findings and to measure progress with regard to performance on the computational task of pitch accent prediction. This second investigation of intonational prominence begins with the design and collection of a new corpus of elicited task-oriented monologues, the Boston Directions Corpus, as discussed in Section 4.1. Next, basic text-based and speech-based processing of the corpus is described in Section 4.2. Results from the foundational text and speech analyses on a portion of the corpus are presented in Section 4.3. Analyses on this subcorpus provide data for various machine learning experiments on accent prediction, presented in Chapter 5. The prediction experiments utilize a wide range of predictors reported on in the literature, many of which were not taken into account in the narrative study. The experimental results establish performance benchmarks for the task of accent prediction on the Boston Directions Corpus, providing a realistic baseline of performance beyond which higher-level features may be deemed to aid in the determination of accentuation. The corpus analyses reviewed in Sections 4.2 and 4.3 and the experimental results from Chapter 5 serve as the foundation for further corpus analysis of higher-level linguistic features, presented in Chapter 6. The role of these features in accent prediction and their relationship to lower-level features are investigated in Chapter 7.

4.1 Corpus Design and Collection

4.1.1 Motivation

In addition to genre, many factors must be carefully controlled in a successful corpus collection. In large part, the collection of a corpus is dictated by the research problem in mind, and is shaped by the goal of maximizing the quality of the data that can be extracted from the resulting corpus to address the research issues at hand. Recently there has been a growing interest in elicitation methods for spontaneous speech, as opposed to read speech. The speech research community has shown concern that the exclusive study of LABORATORY SPEECH — speech recorded in highly controlled laboratory settings — may preclude the discovery of how speech, and especially prosody, is used naturally to achieve communicative goals (Beckman, 1997; Sagisaka, Campbell, and Higuchi, 1997). There is a practical need to study speech in communicative contexts as well, as the demand quickly grows for speech technologies embedded in interactive spoken language or multimodal systems.

So-called laboratory speech is defined by Beckman as “multiply repeated productions of relatively small corpora of sentences designed by the experimenter to vary only in certain dimensions of interest for the prosodic model” (Beckman, 1997, p. 8). Such laboratory speech often exhibits stylized prosody. Related to laboratory speech, the genre of READ SPEECH consists of read productions of preprepared written texts, as exemplified by oft-studied professionally produced newscaster speech. Although the texts that are read may be of considerable length, professionally recorded read speech also exhibits stylized prosody, due to aspects of professional speech training as well as to idiosyncracies of the genres of the texts themselves, which are often intended to engage or entertain an audience as well as to communicate information.

Over the years, researchers have explored the spectrum of speech styles, from laboratory and profes-
sional read speech to non-professional spontaneous speech. Several exemplars of the latter class of corpus have been successfully collected and prosodically analyzed. These include unrestricted, undirected conversations (Bruce et al., 1997), unrestricted, directed problem-solving human-human and human-machine dialogues, such as the ATIS (MADCOW, 1992), Voyager (Zue et al., 1989), Trains (Allen et al., 1995; Heeman and Allen, 1995), Map Task (Anderson et al., 1991) and the ATR travel dialogue (Ehara, Ogura, and Morimoto, 1990) corpora; semi-restricted, directed spontaneous speech elicited by means of visual prompts and displays, such as moving geometric figures (Terken and Hirschberg, 1994) or chains of geometric figures (Brown, 1983; Swerts, Gelyukens, and Terken, 1992); unrestricted, directed spontaneous speech elicited by means of visual prompts and displays, such as moving geometric figures (Terken and Hirschberg, 1994) or chains of geometric figures (Brown, 1983; Swerts, Gelyukens, and Terken, 1992); unrestricted, directed narratives such as descriptions of paintings (Swerts, 1995) and retellings of films and movies (Chaïé, 1980); and unrestricted, undirected narratives elicited using sociolinguistic interview techniques, such as the narrative analyzed in Chapter 3.

For the purposes of the present study, each of the above-mentioned spontaneous, non-professional, corpora has its drawbacks. First, dialogue corpora introduce complexities of speaker modeling that do not arise to the same degree in monologue speech. It has been hypothesized that certain prosodic phenomena in dialogue reflect turn-taking strategies, or even strategies for negotiating mutual beliefs and goals among conversational participants (Local and Kelly, 1986; Ayers, 1994; Bruce et al., 1997) . If one accepts the assumption that the understanding of monologue speech is a first step toward understanding dialogue speech, it would be reasonable as a research strategy to factor out dialogic prosodic phenomena as much as possible by concentrating initially on monologic discourse.

The cited dialogue corpora, nevertheless, capture quite natural task settings for conceivable spoken language systems. For example, the ATIS, Voyager and ATR travel corpora were collected in Wizard-of-Oz or simulated sessions with travel assistance systems. For the Map Task collection, speakers communicated routes on maps, while for the Trains collections, speakers jointly planned the transportation of commodities. This criterion of approaching realistic spoken language system input and output is what disqualifies the remaining corpora from consideration. The elicited speech corpora based on the close description of visual stimuli exhibit stylized syntax, lexical choice and arguably discourse structure. Such deliberate control enables the careful and clear testing of hypotheses, but the range of language occurring in more unconstrained settings may be only partially covered by the resulting analyses. As results on these corpora accumulate, experimenters may gather together many hypotheses and take the next step toward less controlled, more natural speech.

Be that as it may, researchers taking too large a leap can land beyond the bounds — or at least at the edges — of current theory. At this extreme lie the narrative corpora of painting descriptions and movie retellings. In such speech the train of thought of the speaker/reteller is sometimes lost in the recounting of factual details, and it is often unclear how the speaker’s intentions are related to the original communicative goal of the painter, film maker or story writer. In such discourses, the intentional structure can be subtle, or in the end, open to real choices in interpretation — in any event, surely too complex for practical AI/NLP systems to model. Indeed, for this kind of corpus, it has been found that obtaining reliable discourse segmentation data requires extracting a set of “majority” decisions from analyses by a large number of segmenters whose individual disagreements are many (Passonneau and Litman, 1993; Swerts, 1995).

Midway between controlled elicited speech and free-flowing narratives, there is much ground to explore. For example, Terken successfully collected and analysed a corpus of house-building instruction monologues. Speakers were presented with an assortment of colored pieces of a house, and recorded instructions on how to assemble the pieces into a single house. It was assumed that a discourse segment comprised a sequence of utterances on how to place a particular piece. The piece being placed was analyzed as the topic and other
pieces mentioned were non-topics. Thus, the task structure was used to break down the monologues into discourse segments and to assign degrees of saliency to the referents. This study yielded many insights into accentuation patterns and their relationship to topic structure (see the review in Chapter 3). To expand on the narrative study findings, however, it is necessary to design a corpus with a larger variety of referring expressions than elicited in the house-building monologues and with greater complexity in task structure, allowing for a hierarchical analysis of intentions.

4.1.2 Desiderata for a Spoken Language Corpus

The lack of a pre-existing corpus meeting certain desiderata, as outlined above, motivates the undertaking of the design and collection of a new corpus. If the corpus is to resemble the sort of speech that must be handled by SLS’s, the corpus should consist of spontaneous, purposeful productions by multiple non-professional speakers. On the other hand, there are practical limits on the sort of language that can be tackled using current theory, methods and tools. As noted, it is desirable for the moment to focus on monologues, factoring out prosodic aspects of dialogue control and thereby sacrificing a critical feature of SLS-style speech. In addition, to facilitate speech and text processing, the speech needs to be reasonably fluent. And finally, the corpus needs to be amenable to current text and speech processing techniques and intention-based discourse analysis methods. These last two qualities are discussed below.

When speech technology developed for read or prepared speech is applied to spontaneous speech, unforeseen obstacles arise and breakdowns often occur. Apart from greater variation in phrasing, pitch range and tempo relative to read speech, spontaneous speech is plagued by the presence of disfluencies. It is desirable to reduce disfluent phenomena to a minimum, since disfluent speech is an entire research subject in itself (Shriberg, 1994). Disfluent speech poses difficulties for prosodic transcription, although some progress towards standards has been made, e.g. (Nakatani and Shriberg, 1993). Similarly, NLP tools employed in this study for corpus analysis are robust enough to handle mildly disfluent or noisy language, but they are not designed to directly attack the problem of identifying and repairing disfluent streams of text. Some theories of language production presume that certain kinds of disfluencies act as time-killing, floor-gaining or floor-holding devices, continuing the flow of the speech stream even when the appropriate words, phrases or thoughts have not been entirely formed in the speaker’s mind (Goodwin, 1981; Local and Kelly, 1986; Levelt, 1983). It has been proposed in particular that disfluencies may accompany periods of high planning loads (Levelt, 1989). Thus, there arises a confluence of two goals: (1) minimizing disfluencies; and (2) limiting the complexity of linguistic form as well as message.

4.1.3 Boston Directions Corpus

The new corpus, the Boston Directions Corpus, is made up of spontaneous and read versions of elicited direction-giving monologues by multiple non-professional speakers. The genre of direction-giving was chosen for the new corpus collection with the discussed desiderata in mind. In previous investigations of discourse structure in Associated Press news stories, Grosz and Hirschberg (1992) encountered stylistic idiosyncracies of professionally read speech that influenced prosodic and discourse analyses, and although

---

2 A parser that can repair detected disfluencies was proposed by Hindle (1983). Disfluency detection, however, remains a difficult problem requiring multiple knowledge sources (Bear, Dowding, and Shriberg, 1992; Nakatani and Hirschberg, 1994; Heeman and Allen, 1997).

3 The original design for the Boston Directions Corpus was conceived of by Barbara Grosz and Julia Hirschberg in 1992. The author contributed to the testing and implementation of the design, and created the experimental protocols involved in subject recording. Actual recordings were made by the author and Charles Reiss, from May 1994 to May 1995 at Harvard University. The author, Alan Capil and Charles Reiss transcribed the speech. The author, Charles Reiss and Barbara Grosz processed the transcriptions for the read speech recordings. Subsequent speech processing was carried out at AT&T Labs - Research (formerly AT&T Bell Laboratories) and at Harvard University.
reliable segmentations were obtainable, the wide differences between news stories hampered analysis. Thus, in embarking on a new corpus collection they envisioned a corpus designed to yield not only reliable but also comparable discourse analyses. The directions domain satisfies these requirements because of its objective and limited domain knowledge, which allows the clear definition of explicit discourse goals. Also, there are many examples of fruitful spoken language research in the directions domain, such as the Direction Assistant (Davis, 1987; Davis and Hirschberg, 1988), the Voyager system (Zue et al., 1989) and the Map Task project (Anderson et al., 1991).

Experience with numerous pilot collections revealed that careful manipulation of planning complexity is the most important factor in achieving fluency and in constraining linguistic structures, from vocabulary size to discourse intentions. Pilot collections also showed that a secondary factor in maximizing fluency and facilitating language planning is the appropriate use of visual aids, such as maps and written notes.

Collection of the Boston Directions Corpus consisted of two phases, the recording of elicited spontaneous speech and the recording of read versions of the spontaneous speech. Speakers were solicited to participate in both phases with a print advertisement. Selection criteria contained in the advertisement were as follows: well-versed in Cambridge/Boston city life, good at giving directions, native speaker of American English and normal hearing. All speakers were graduate or advanced undergraduate students at Harvard University or at local universities and all were paid for their participation. Subjects had no knowledge of the purpose of the recordings, besides the information in the advertisements for subject recruiting.

In the spontaneous recording sessions, speakers were given written instructions to perform a series of nine increasingly complex direction-giving tasks. Speakers first explained simple subway routes such as getting from one station to another, and then more complex routes involving subway line transfers. Next, they recorded simple and complex walking routes. Then, they combined walking and subway routes into longer journeys, and gave tourist information such as restaurant recommendations. Thus, they progressed to the last and most complex task of planning a round-trip journey originating from Harvard Square in Cambridge to several tourist sights in Boston, providing tourist information along the way. This task progression allowed the speaker to become comfortable in the early tasks with the planning and production of short segments which became the building blocks for later tasks. This building-block approach significantly increased fluency over pilot approaches in which only complex tasks were collected. Another feature of this design was that it facilitated discourse analysis, because the explicit discourse goals of simple tasks became easier to identify when they were embedded as subgoals in more complicated tasks.

As mentioned, a second factor in controlling planning complexity is the use of visual aids. To ease the planning load and eliminate memory recall problems as much as possible, the speakers were provided with various maps and they were allowed to write notes (but only words and phrases, not full sentences) to themselves as well as trace routes on the maps. Speakers were allowed to look at the maps and their notepad during both the planning and recording of the spontaneous version. For the duration of the experiment, the speakers were in face-to-face contact with a silent experimental partner (a confederate) who traced on her map the routes described by the speakers. The controlled task environment simulated in the corpus collection came close to resembling a typical SLS scenario, where a user is usually enlisting the help of a system for limited problem-solving in a multimedia environment, whether sitting at a workstation or on the telephone with pen and paper.

The elicited speech was subsequently orthographically transcribed, with false starts and other speech errors repaired or omitted. The edited transcription was then printed with one sentence per line, since without sentence punctuation, readability is severely compromised. Subjects returned several weeks after their first recording to read the transcribed speech aloud. Speakers were instructed to read through the transcript one task at a time until they clearly comprehended the directions. At this time they were allowed

---

4 Sentence periods used for this purpose were determined during transcription following practices of office dictation. This sentence punctuation was then discarded and a new method for determining sentence boundaries employed, as discussed in Section 4.2.
to mark-up the transcriptions with punctuation if desired. Visual aids used in the spontaneous recording sessions were not available during the recording of the read speech. This collection technique of having the same speaker record spontaneous speech and then later read his or her own transcribed words has been used successfully as a control in studying the prosody of spontaneous speech (Gårding, 1967; Ayers, 1994; Bruce et al., 1997). Differences in accent patterns in read and spontaneous speech remain relatively unexplored and are therefore one focus of the current study.5

Appendix A contains the exact instructions to subjects for both the spontaneous and read recordings. Figure 4.1 shows two different speakers’ spontaneous versions of the fifth task, a moderately complex task combining walking and subway routes. The exact task instructions were:

What route would you take to get from the steps of the entrance to the MIT “Dome” building at 77 Massachusetts Avenue to the New England Aquarium in Boston?

Speaker H1: Spontaneous Version of Task Five

currently we are on Massachusetts Avenue / we have just left the building / we will take a right turn / and walk to the corner of Vassar Street / here we will take another right turn / and travel down toward Main Street / after getting to Main Street / we’ll take yet another right turn / and travel down / past / Ames Street / and past / Carleton Street / there’ll be a large square here / where the T is found / we will then get on the T / at the Kendall Square stop / and travel down the Red Line until we get to Park Street / at Park Street we will get off the Red Line and get on the Green Line / then we will take the Green Line northbound toward Lechmere Station / and we will get off {at the first stop / Government Cen(ter)- / the} at the next stop / Government Center / here we will get on the Blue Line / and take the Blue Line / to State Street and then to the Aquarium stop / we will then {get off the stat(ion)- uh} get off at the Aquarium stop / and exit the T station / we should be facing two large buildings / the Harbor Towers / we will walk toward them / and take our first right / out onto the wharf / here we will find the New England Aquarium

Speaker H3: Spontaneous Version of Task Five

from the MIT Dome / I would turn left / and walk / down / Memorial Drive / two blocks / to Wadsworth Street / at {Wadsworth-} Wadsworth Street / I turn left onto Wadsworth / and / walk two blocks / to Kendall Square / at Main Street / if you just {walk} turn left and walk a block to the T station / that’s the Kendall Square T stop / get on the Red Line / and ride the Red Line / to Park Street Station / at Park Street station you wanna take / the Green Line / to Government Center which is just one stop / at which time you have to make yet another / {uh} line switch / and this time you’ll get the Blue Line / {uh} headed / toward Wonderland ultimately / but / you’ll only ride / for two stops /and get off at the Aquarium

Figure 4.1: Two speakers’ spontaneous versions of Task Five. Forward slashes indicate intonational phrase boundaries and curly braces mark off the reparanda in disfluencies.

As can be seen, the speakers were unconstrained in what they were allowed to say. In the discourses in Figure 4.1, the two speakers chose entirely different paths from the starting point, the front steps of MIT, to the nearest subway station, Kendall Square Station. One went clockwise around the block, and the other went counterclockwise (while yet another speaker wound his way through the Infinite Corridor). Despite this

5Preliminary results on acoustic-prosodic correlates of discourse structural properties in read and spontaneous speech, and other prosodic analyses of the Boston Directions corpus, are reported elsewhere (Nakatani, Grosz, and Hirschberg, 1995; Hirschberg, 1995; Hirschberg, Nakatani, and Grosz, 1995; Hirschberg and Nakatani, 1996).
freedom in problem-solving, the domain of direction-giving and identical task specifications for all speakers led to the use of relatively straightforward syntax and a small vocabulary size for the corpus overall.

Both the spontaneous and read recordings were made using a Shure SM10A headset microphone. For all but two of the speakers, the speech was directly recorded onto a SONY DAT recorder. The speech of the two remaining speakers was recorded onto a SUN Sparc station. All speech recordings were subsequently transferred onto Silicon Graphics workstations for further speech processing. The speech was downsampled to 16 kilohertz, filtered and pitchtracked using Entropics software in addition to speech analysis software developed at AT&T Bell Laboratories. In total, the Boston Directions Corpus consists of sixteen sets of spontaneous speech, with corresponding read versions recorded for six speakers.

4.2 Foundational Analyses

Two sets of foundational analyses, one text-based and one speech-based, were applied to the Boston Directions Corpus. This basic level of processing not only provides data for baseline machine learning experiments, but also serves as a foundation for further data analysis. Basic text analysis consisted of orthographic transcription, sentence boundary placement and part-of-speech tagging, and was carried out with the use of NLP tools where applicable. Basic speech analysis consisted of orthographic boundary placement and intonational labeling, carried out by hand.

4.2.1 Text Analysis

The first phase of text-based processing consisted of three analyses: orthographic transcription, marking sentence boundaries and lexical or part-of-speech (POS) tagging. From these foundational analyses, a number of word features were derived for the learning experiments described in Chapter 5. These included sentence position information, POS tag for each word, neighboring POS tags and various classifications of POS tags, such as broad class and function versus content word.

The orthographic transcription was created first. Then sentence boundaries were placed in the transcription. Finally, the orthographic transcription and sentence boundaries were included as input to the POS tagger. In the next sections, the three processes are described in the order in which they were applied to the data.

Orthographic Transcription

The text-based processing started with the orthographic transcription of the speech. The transcription was checked for phonetic accuracy and aligned with the speech amplitude waveform. Orthographic transcription conventions outlined in the ToBI Guidelines (Beckman and Ayers, 1994), which were adapted from the ATIS conventions (MADCOW, 1992), were followed. ToBI standards for marking disfluencies (Nakatani and Shriberg, 1993) were also followed. It is important to note that in spoken language transcription there is no certain method for introducing punctuation.

Sentence Boundaries in Spoken Language

The performance of the NLP tools used for text-processing is improved when sentential units are marked in the input, motivating the placement of sentence boundaries in the transcriptions. While all non-final punctuation can be relatively safely omitted, it was necessary to define a consistent method for determining sentence boundaries in this corpus. The task of finding sentence boundaries in written text is often significantly aided by the presence of orthographic punctuation. In speech transcriptions, the beginnings and
endings of sentential units are often ambiguous. Since there are no generally accepted standards for marking sentence boundaries in spontaneous speech, new guidelines were developed to suit the purposes of this study. The main goals were to achieve consistency and to execute the task so as to optimize the performance of the chosen NLP tools.

Based on the text transcription alone, sentence boundaries were placed before coordinating conjunctions and cue phrases when the two resulting sentences without the conjunction or cue phrase did not differ syntactically from the conjoined structure. This resulted in shorter sentences, which improves the performance of the tagger and parser. For example, the following two sentences were divided at the conjunction: First enter the Harvard Square T stop. And buy a token. Sentences were divided at cue phrases regardless of whether the cue phrase was used in its discourse or sentential sense, as in: Any of the different Green Line cars will take you to Copley Station. So simply board one. The placement of the sentence boundary in such a case arguably alters the nuances of the meaning of the sentence, but the criterion of preserving syntactic structure is met. Conjoined verb phrases were marked as separate sentences when they did not share any modals, arguments or adjuncts in common. Otherwise, they were marked as a single sentence. The following phrases were therefore marked as a single sentence, since the two main predicates share the same subject and modal: So you will walk through Harvard Square up toward Johnston Gate and take a right turn into the Yard.

These rules reduced sentence length as much as possible without introducing artificial syntactic artifacts. When real syntactic ambiguity did arise, as for subordinate clauses that could syntactically attach to either the preceding or following sentence, the semantics of the discourse context was used to disambiguate the sentential structures. For example, although the subordinate clause, after entering Harvard Yard, is syntactically ambiguous in the context, And we can enter the Yard from a gate after entering Harvard Yard we will pass by a number of dorms, the semantics clearly indicates that the subordinate clause begins a new sentence, as in: And we can enter the Yard from a gate. After entering Harvard Yard we will pass by a number of dorms. When the choice in attachment for a subordinate phrase was both syntactically and semantically ambiguous, the rule was followed that the subordinate clause was assigned to the second sentence. Thus, for the following context, In addition Quincy Market contains a number of fine places to have lunch before we go to the museum there’s also some shopping down toward the further side of Quincy Market, a sentence boundary was placed by rule before the ambiguous phrase, before we go to the Museum, resulting in the following two sentences: In addition Quincy Market contains a number of fine places to have lunch. Before we go to the museum there’s also some shopping down toward the further side of Quincy Market. For these cases, intonational cues such as prosodic phrasing may influence syntactic interpretation, but since the main goal of sentence boundary analysis was to aid the performance of the text processing tools, no use of speech cues was made in this syntactic approach to placing sentence boundaries in spoken language text.

**Part-of-Speech (POS) Tags**

Lexical or part-of-speech tagging has become one of the most robust and useful NLP technologies for text processing. Besides improving the accuracy of syntactic analysis, POS information itself is highly correlated with accentuation and often plays a large role in accent placement in both text-to-speech and message-to-speech applications (see Chapter 2).

For this study, the Boston Directions monologues were tagged with the Penn Treebank POS tags (Santorini, 1991). This task was semi-automated by using a tagger that employs a transformation-based learning algorithm, devised and implemented by Brill (1995). The speech-aligned orthographic transcriptions were tokenized and tagged for POS using Brill’s transformation-based tagger (v. 1.14) in 1-best mode. Each lexical item was given a unique tag from the standard Penn Treebank set of lexical tags, with the following extensions: word fragments and disfluency markers, such as *uh* and *um*, were permanently assigned the
interjection tag, *UH*. This allowed for the simple filtering of disfluency markers and incomplete word productions in later analysis and experiments and also prevented the tagger from attempting to process these non-lexical tokens.

Brill’s tagger lends itself to small corpora due to its ability to quickly leverage performance from an initial system trained on a very large corpus with limited training cycles. Because the rules learned by the tagger are in the form of familiar linguistic rewrite rules, outstanding bugs can often be eliminated by error analysis and rule-editing. This is in contrast to purely stochastic language-modeling approaches, in which the system learns probabilities that are uninterpretable outside of the model, e.g. (Church, 1988). One bootstraps on Brill’s initial system by iterating through relatively small training sets of labeled data from the target corpus. On each cycle, the training process can be facilitated by hand-editing of both the lexicon and two sorts of transformation-based rule sets, lexical and contextual. These rules express different kinds of linguistic environments that trigger a change in an assigned tag. The majority of these transformation-based rules are learned automatically, but user intervention is particularly helpful to encode modifications in rules that arise from a change of domain. For example, the word *train* in the *Wall Street Journal* is most likely to be used as a verb, while in the *Boston Directions* domain, it is a noun used to refer to a subway. Likewise, new lexico-syntactic constructions may arise when switching domains. These problems are often intertwined, as in the case of the light verb construction *to make a left turn*, which should be tagged *to/TO, make/VB, a/DT, left/JJ, turn/NN* in the directions domain. However, in the *Wall Street Journal* corpus, *left* is tagged as a noun and *turn* as a verb. Such cases can be readily fixed by adding a new contextual rule. When bootstrapping a system for a small corpus, it is quite possible to acquire new rules for unseen language, but sparse data makes it unlikely that rules that fit the initial large training corpus will be overturned.

When an “out-of-the-box” version of Brill’s tagger, trained on roughly three million tagged words of the *Wall Street Journal* from the Penn Treebank, was run on a subset of the directions, namely the complete set of read speech from one speaker (referred to hereafter as H1), the resulting system achieved 87.3% accuracy (286 errors out of 2,244 words). The automatically assigned tags for the H1 read set were then hand-corrected and added to the training corpus; the tagger was then retrained on the expanded training corpus. In addition, a small number of “patches” or transformational rules, formed by inspection of errors, were added by hand to the lexical and contextual rule files and a small number of lexicon entries were tuned to the domain. This retrained version of the tagger achieved 92.6% accuracy (174 errors out of 2,359 words) on the spontaneous version of H1, whose transcription is almost identical to the read transcription. This represented a 5.4% decrease in errors, or a 42% reduction in the error rate over the original version of the tagger. On a completely blind test set, the spontaneous speech from speaker H3, the retrained tagger with no further tuning performed with 92.5% accuracy (121 errors out of 1,616 words). Given the comparable performance on the two spontaneous data sets, no further training was undertaken.

### 4.2.2 Prosodic Analysis

In this study, the prosodic transcription, a more abstract representation of the intonational prominences, phrasing, and melodic contours, was obtained by hand-labeling. In particular, prosodic transcription of the Boston Directions Corpus adhered to the ToBI (Tones and Break Indices) standard (Silverman et al., 1992; Beckman and Ayers, 1994; Beckman and Hirschberg, 1994; Pitrelli, Beckman, and Hirschberg, 1994). As described in Chapter 1, a ToBI transcription is made up of time-aligned transcriptions of speech events along several dimensions or tiers, namely orthographic, tonal, junctural and paralinguistic (i.e. a ‘miscellaneous’

---

6 Automation of word alignment was attempted, but due to disfluencies and perhaps other aspects of spontaneous speech, such automation was deemed less time-efficient than word segmentation by human.

7 The ToBI Guidelines, Annotation Conventions and accompanying speech materials are obtainable by writing to tobi@ling.ohio-state.edu.
ToBI transcription, including the orthographic labeling discussed in Section 4.2.1, was carried out using the Entropic \textit{waves}+ software package (Talkin, 1989) and the ToBI labeling tools on Silicon Graphics workstations.

Tonal transcription in ToBI is based upon Pierrehumbert’s (1980) theory of American English intonation. The tonal transcription supplies a labeling of pitch prominences as well as a breakdown of the speech sample into intermediate and intonational phrases, with phrase-final contours labeled for phrase accent and boundary tone. The junctural level of analysis consists of the placement of a \textit{BREAK INDEX} (BI) level between each pair of words, indicating the temporal unity of the given word pair (Ostendorf et al., 1990). Generally speaking, a BI of level 3 corresponds to an intermediate phrase boundary, while a BI of level 4 corresponds to an intonational phrase boundary. (BIs 0, 1 and 2 describe various levels of word connectivity within a prosodic phrase.) For the most part, BI labeling at levels 3 and 4 is redundant with phrase boundary labeling in the tonal tier and BI labels are therefore not used in the machine learning experiments.

Each set of speech was analyzed by two labelers trained in the ToBI system in a two-pass system to ensure higher reliability of prosodic transcription. The ToBI group conducted a study of inter-transcriber reliability on a diverse group of twenty-six transcribers. Consistency was measured in terms of the percentage of all possible pairs of transcribers who agreed on the prosodic transcription associated with each word. Pitrelli and colleagues (1994) report inter-transcriber agreement of 88% as to the presence or absence of a particular category of tonal element (i.e. pitch accent, phrase accent or boundary tone), 81% as to the exact tonal label and 92% as to break index level when differences of one level were permitted.

### 4.3 Subcorpus

The foundational speech and text analyses described in Section 4.2 were carried out on three sets of speech from the Boston Directions Corpus. This subcorpus included the read and spontaneous speech of one male speaker (speaker H1), and the spontaneous speech of one female speaker (speaker H3). These sets will be referred to as H1-read, H1-spon and H3-spon, respectively. Each consisted of nine direction-giving monologues, as described in Section 4.1.

As previously mentioned, the effect of speaking style on prosody is one research interest. This issue can be explored by comparing the read and spontaneous speech of speaker H1. A second research interest is speaker differences. To this end, comparing the spontaneous speech of speakers H1 and H3 allows for tentative insights into individual speaker differences. Of course, a larger sample of speakers should be investigated to draw reliable conclusions in this area.

To provide a quantitative as well as qualitative picture of the subcorpus, a number of corpus measures are given in the tables below. First, text-based characteristics for each set in the subcorpus are given in Table 4.1, including total number of words, total number of sentences and average sentence length. Differences in word counts between H1-read and H1-spon arise from the occurrence of disfluencies, in either the read or spontaneous recordings. Speaker H3 completed the direction-giving tasks with fewer words and sentences, but on average her sentences were much longer than those of speaker H1.

Next, speech-based characteristics for each set are given in Table 4.2, including total amount of speech in terms of number of intermediate and intonational phrases. Speaker H1 shows similar phrase lengths for both read and spontaneous speaking styles, while speaker H3 uses shorter intermediate and intonational phrases than does speaker H1 in the spontaneous speaking style.

Table 4.3 displays information on disfluencies. These statistics are based only on the occurrences of tokens labeled with the \textit{UH} part-of-speech tag, as discussed in Section 4.2. As such, they are only approxi-
Table 4.1: Corpus measures from text-based analysis.

<table>
<thead>
<tr>
<th>Directions set</th>
<th>Total no. of words</th>
<th>Total no. of sentences</th>
<th>Average no. of words/sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1-read</td>
<td>2,244</td>
<td>184</td>
<td>12.2</td>
</tr>
<tr>
<td>H1-spon</td>
<td>2,359</td>
<td>185</td>
<td>12.8</td>
</tr>
<tr>
<td>H3-spon</td>
<td>1,616</td>
<td>99</td>
<td>16.3</td>
</tr>
</tbody>
</table>

Table 4.2: Corpus measures from speech-based analysis.

<table>
<thead>
<tr>
<th>Directions set</th>
<th>Total no. of intermediate phrases</th>
<th>Average no. of words per intermediate phrase</th>
<th>Total no. of intonational phrases</th>
<th>Average no. of words per intonational phrase</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1-read</td>
<td>503</td>
<td>4.5</td>
<td>340</td>
<td>6.6</td>
</tr>
<tr>
<td>H1-spon</td>
<td>561</td>
<td>4.2</td>
<td>365</td>
<td>6.5</td>
</tr>
<tr>
<td>H3-spon</td>
<td>570</td>
<td>2.8</td>
<td>365</td>
<td>4.4</td>
</tr>
</tbody>
</table>

Table 4.3: Corpus measures regarding disfluencies.

<table>
<thead>
<tr>
<th>Directions set</th>
<th>Total no. of repair tokens</th>
<th>Overall repair rate</th>
<th>Average no. of repair tokens/sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1-read</td>
<td>14</td>
<td>.006</td>
<td>.075</td>
</tr>
<tr>
<td>H1-spon</td>
<td>68</td>
<td>.029</td>
<td>.364</td>
</tr>
<tr>
<td>H3-spon</td>
<td>70</td>
<td>.043</td>
<td>.707</td>
</tr>
</tbody>
</table>

mate measures, and do not, for instance, express the total number of words that occur in the *reparandum* or corrected material in a restart repair. Nevertheless, since the presence of a disfluency is frequently marked by a filled pause or word fragment, the statistics provide a reasonable basis for comparison between speakers and with other corpora. Differences in overall repair rates arise between read and spontaneous speech for speaker H1, as would be predicted. While six in one thousand tokens are disfluent in H1-read, nearly five times as many, or three in one hundred, are disfluent in H1-spon. The overall repair rate for H3-spon is somewhat higher than that of H1-spon. Partly because of H3’s longer sentences, it is almost twice as likely that a repair token will occur in a sentence in H3-spon than in H1-spon, as indicated by the statistics on the average number of repair tokens per sentence.

Finally, Table 4.4 provides a breakdown on the number of words and phrases per task, illustrating the general progression in length from the simple to complex tasks. It is interesting to examine the numbers in Table 4.4 with reference to Table 4.5, which gives a taxonomic description of the direction-giving instructions for each task. Careful inspection shows that monologue length is well correlated with planning difficulty as conceived of in the corpus design.

As noted, results and further analyses described in the remainder of this thesis derive from investigation of the spontaneous speech for speakers H1 and H3 as well as the read speech for speaker H1, on all nine
<table>
<thead>
<tr>
<th>Task No.</th>
<th>Task</th>
<th>Intermediate Phrases</th>
<th>Intonational Phrases</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Simple subway route (no transfers)</td>
<td>H1r: 42, H1s: 43, H3s: 50</td>
<td>H1r: 10, H1s: 16, H3s: 25</td>
</tr>
<tr>
<td>2</td>
<td>Moderate subway route (one transfer)</td>
<td>H1r: 90, H1s: 98, H3s: 70</td>
<td>H1r: 22, H1s: 24, H3s: 27</td>
</tr>
<tr>
<td>3</td>
<td>Walking route</td>
<td>H1r: 142, H1s: 142, H3s: 88</td>
<td>H1r: 32, H1s: 31, H3s: 35</td>
</tr>
<tr>
<td>4</td>
<td>Walking route</td>
<td>H1r: 115, H1s: 120, H3s: 73</td>
<td>H1r: 27, H1s: 31, H3s: 30</td>
</tr>
<tr>
<td>5</td>
<td>Simple walking and subway routes</td>
<td>H1r: 205, H1s: 215, H3s: 132</td>
<td>H1r: 42, H1s: 45, H3s: 53</td>
</tr>
<tr>
<td>6</td>
<td>Moderate walking and subway routes</td>
<td>H1r: 330, H1s: 345, H3s: 232</td>
<td>H1r: 67, H1s: 70, H3s: 91</td>
</tr>
<tr>
<td>7</td>
<td>Walking route</td>
<td>H1r: 301, H1s: 341, H3s: 212</td>
<td>H1r: 64, H1s: 83, H3s: 66</td>
</tr>
<tr>
<td>8</td>
<td>Moderate walking, subway, and tourist information combined</td>
<td>H1r: 408, H1s: 424, H3s: 388</td>
<td>H1r: 88, H1s: 99, H3s: 123</td>
</tr>
<tr>
<td>9</td>
<td>Complex walking, subway, and tourist information combined</td>
<td>H1r: 611, H1s: 631, H3s: 371</td>
<td>H1r: 151, H1s: 162, H3s: 120</td>
</tr>
</tbody>
</table>

**Table 4.4:** Length by word and phrasal units, per task.

<table>
<thead>
<tr>
<th>Task No.</th>
<th>Taxonomic Description of Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Simple subway route (no transfers)</td>
</tr>
<tr>
<td>2</td>
<td>Moderate subway route (one transfer)</td>
</tr>
<tr>
<td>3</td>
<td>Walking route</td>
</tr>
<tr>
<td>4</td>
<td>Walking route</td>
</tr>
<tr>
<td>5</td>
<td>Simple walking and subway routes combined</td>
</tr>
<tr>
<td>6</td>
<td>Moderate walking and subway routes combined</td>
</tr>
<tr>
<td>7</td>
<td>Tourist information</td>
</tr>
<tr>
<td>8</td>
<td>Moderate walking, subway, and tourist information combined</td>
</tr>
<tr>
<td>9</td>
<td>Complex walking, subway, and tourist information combined</td>
</tr>
</tbody>
</table>

**Table 4.5:** Taxonomic description of tasks, describing level of planning difficulty.

direction-giving tasks.
Chapter 5

Baseline Accent Prediction Experiments

As noted in Chapter 2, many practical speech synthesis systems rely on lexical class or POS information to determine accent placement, while more sophisticated systems incorporate surface order information, syntactic information such as preposing, discourse information such as given/new status and semantic information such as contrast, focus and semantic weight. There are many interactions between the features that complicate attempts to model accentuation. For example, consider the following sentence:

At the museum there are many fine exhibits to see.

The prepositional phrase in this sentence, At the museum, is in sentence-initial position, is preposed from its underlying sentence-final position and is a candidate for contrastive focus marking. This constellation of properties may appear to mark this item for accenting, but with a deeper understanding of the semantics and pragmatics of preposing, this decision must be reconsidered. Suppose that the immediate context of the above sentence is the following:

Walk across the street toward the Museum of Fine Arts.  
At the museum there are many fine exhibits to see.

In this context, the preposed prepositional phrase is in fact not a contrastive focus but a discourse topic (Ward, 1985) or scene-setting locative phrase (Kuno, 1975), and is therefore more likely to be deaccented on account of its givenness (Turken, 1984).

The diverse, copious amounts of data needed to make accenting decisions can be mined using machine learning techniques on a labeled speech corpus. For the purposes of this study, machine learning is viewed as a data exploration tool. Machine learning experiments can shed light on pockets of difficult phenomena that the learned system cannot adequately model, which might be more closely examined in distributional analysis studies. Ultimately, such investigations can provide empirical validation or refutation of proposed architectures for pitch accent assignment and offer performance-based assessments of critical areas for improvement.

The alternatives to data mining techniques have been pushed to their limits. Many hand-crafted systems have come to rely on intuitive notions of semantic and pragmatic properties that are neither rigorously defined nor integrated with other areas of linguistic knowledge. Because of a lack of empirical testing, it is often difficult to assess the efficacy of these hand-crafted systems, and there arises the obvious risk of incomplete coverage of the data. This can be partially remedied by conducting distributional analyses, but the enormous space of viable hypotheses to test makes exhaustive analysis using classic statistical inference techniques nearly infeasible.

This chapter reports results from machine learning experiments on accent prediction, using the labelled subcorpus described in Section 4.3. The experiments test how well a wide range of linguistic features reported on in the literature predict accentuation in the Boston Directions Corpus. Previous work on accent
prediction, reviewed in Section 5.1, serves as the departure point for new machine learning experiments using a rule-based induction system called RIPPER (Cohen, 1995), which is described in Section 5.2. The classification schemes and predictors used in the RIPPER experiments are defined in Sections 5.3 and 5.4 respectively. Section 5.5 presents experimental results from an initial set of baseline experiments that duplicate and slightly extend previous work on classifying tokens as accented versus unaccented. These experiments establish performance benchmarks for the task of accent prediction on the Boston Directions Corpus and provide a realistic baseline for assessing the gains to be had by modeling higher-level features for accent prediction.

5.1 Previous Work

While language modeling techniques have been extensively developed for speech recognition, the use of machine learning techniques for prosodic modeling has been pursued more recently (Riley, 1989). In an early study, Wang and Hirschberg (1992) used classification and regression trees (CART) (Breiman et al., 1984) to predict intonational phrase boundaries in a speech corpus. Since then, machine learning has been applied to prosodic data to probabilistically score parses using prosodic phrase structure (Veilleux and Ostendorf, 1993), to predict discourse segment boundaries based solely on intonational features (Hirschberg and Grosz, 1992) or on intonational features in combination with text-based features (Passonneau and Litman, 1993), to detect disfluencies in spontaneous speech (Hirschberg and Nakatani, 1993), to disambiguate cue phrase usage as sentential or discourse structural based partially on phrasing and accent information (Litman, 1996) and to assign tonal contours to annotated dialogues (Black, 1997).

Most relevant to the present study is Hirschberg’s extensive study of pitch accent prediction (Hirschberg, 1990a; Hirschberg, 1990b; Hirschberg, 1993), which provided comparative performance data not only for hand-crafted versus automatically learned systems but also for different types of speech corpora. This study also was the first to apply machine learning techniques to investigate the relative contributions of lower-level (i.e. lexical, syntactic and surface position) and higher-level linguistic features (i.e. contrast, semantic focus and global focusing). Hirschberg’s aim was to develop an automatically trainable system with robust performance over a variety of speech corpora. These requirements were dictated by the ultimate goal of developing a pitch accent assignment module for unrestricted TTS synthesis. To this end, automatic labeling methods for each predictor were employed, including original methods for heuristic semantic and discourse labeling based on shallow text-processing.

To summarize Hirschberg’s (1993) results, evaluation of hand-crafted rule systems on a portion of the Boston University Radio News Corpus (Ostendorf, Price, and Shattuck-Hufnagel, 1995) yielded 68% correct accentuation (i.e. classification as accented, deaccented but not cliticized, or cliticized) utilizing the function versus content word distinction; 77% using broad class POS categories; and 82.4% when given/new distinctions, focus and contrast information were heuristically modeled. The Boston University Radio News Corpus (Ostendorf, Price, and Shattuck-Hufnagel, 1995) consists of mainly professionally read speech with short segments from spontaneous interviews interspersed. The best hand-crafted rule systems achieved 98.3% correct classification on a corpus of citation-form sentences. Automatically learned systems were trained and tested on two additional corpora using classification and regression tree techniques (Breiman et al., 1984). Additional features, such as phrase position, POS context and speaker identity, were examined for these two corpora. Correct classifications were automatically learned for 80% of the tokens in professionally read news speech from the Audix corpus and 85% of the tokens in spontaneous dialogue speech from the ATIS corpus (MADCOW, 1992). The decision tree learned for the Audix corpus was summarized as follows: “If an item’s POS tag indicates it is closed-cliticized or closed-deaccented, or if it represents given information, or if it is predicted to be deaccented in complex nominal citation form, or if it is a nominative pronoun, predict that it is deaccented; otherwise, predict accented” (Hirschberg, 1993). Decision
trees for the ATIS corpus were similar, except that additional factors such as phrase position, neighboring POS information and speaker identity proved useful. Hirschberg conjectured that word lemma information might offer further improvements, but did not test this hypothesis. In sum, Hirschberg’s studies established that accent prediction in citation-style speech can be very accurately modeled, while complex interactions of factors determine accentuation in extended segments of professionally read speech as well as short segments of spontaneous speech. Given the difference in performance for citation-style speech on the one hand, and longer segments of read speech and spontaneous dialogue speech on the other hand, it may also be concluded that some factors influencing accentuation in discourse context remain to be discovered.

Later experiments (Ross, Ostendorf, and Shattuck-Hufnagel, 1992) investigated syllable-based and word-based accent prediction using similar features as well as new ones, such as factors associated with multiple pitch-accented words, also on the Boston University Radio News Corpus. Correct accent prediction was scored by comparing the predictions of the learned system against the accenting in four read versions of the same texts; if any speaker accented a word that the system accented, the system prediction was judged correct. Using CART on a feature set derived from Hirschberg’s study, 97% of words predicted to be accented were accented by at least one of the four speakers, while 15% of words not accented by any speaker were incorrectly predicted to accented, for a total error rate of 18%. The accentuation data analyzed in this manner confirmed that speakers produce variability in accent assignment that has yet to be fully modeled, since there were many words to which speakers did not all assign the same accentuation.

5.2 A Rule Learning System for Data Exploration

Various machine learning techniques such as neural networks, decision tree learning and rule learning, have been applied to language learning problems. The accent prediction experiments in this thesis make use of a rule induction system called RIPPER (Reduced Incremental Pruning to Produce Error Reduction), developed by Cohen (1995). RIPPER is an implemented supervised learning algorithm similar to the decision tree learning programs such as CART (Breiman et al., 1984) and C4.5 (Quinlan, 1994) in that it learns classification models by optimizing a measure of information gain in an iterative, greedy search process. Critical differences between RIPPER and the other systems are that RIPPER uses a separate-and-conquer strategy (as opposed to divide-and-conquer) to develop a set of IF-THEN rules (as opposed to decision trees) by utilizing incremental error reduction techniques in combination with novel rule optimization strategies that emulate the behavior of non-incremental or conventional reduced error pruning. In addition, RIPPER operates efficiently and effectively on very large and noisy data sets, scaling nearly linearly with the number of training examples, while preserving expressivity in rule-writing and feature representation. It has been argued that if-then rules are more readily understood by the human experimenter than decision-tree output and that rule induction systems are thus an appropriate tool for data exploration. However, RIPPER is the first implemented rule-based system whose performance, in terms of computational efficiency and accuracy, rivals that of the well-studied decision-tree systems.

In machine learning experiments on cue phrase disambiguation, Litman (1996) utilized both C4.5 and a precursor to RIPPER named CGRENDDEL (Cohen, 1993) and reported comparable performance for the two systems. Cohen (1995) tested RIPPER and C4.5 on a suite of benchmark learning problems, and reported that with 93% confidence, the probability is 0.5 that RIPPER’s measured error rate will be less than or equal to that of C4.5rules, which is a program that outputs rule set translations of C4.5 decision trees (Quinlan, 1994). With 95% confidence, the probability is 0.488, suggesting that the two systems deliver comparable performance measured by the accuracy of the learned concept. Pilot accent prediction experiments with C4.5 and RIPPER on the Boston Directions Corpus gave very similar results. For reasons of efficiency, expressivity of feature representations, and economy of rule representation (arising from the use of a minimum description length metric during rule optimization), RIPPER was chosen as the machine
5.3 Classification of Accent Types

Nine fundamental accentuation classes were derived from the ToBI prosodic transcription: cliticized (cl), deaccented (-), accented but with unknown tonal properties (*) and six classes of pitch-accent (H*, L*, H+L*, H*+L, L*+H, L+H*) (Pierrehumbert, 1980). The experiments reported below employ a simple grouping of these nine fundamental accentuation classes into two categories, UNACCENTED and ACCENTED. The clitic class and deaccented class together form the unaccented class. The remaining items — the four complex pitch accents, the two simple pitch accents and pitch accents with unknown tonal properties — make up the accented class.

5.4 Predictors

The experiments reported in this chapter incorporated all of the features utilized in Hirschberg’s experiments (1993) except for given/new discourse information, which is investigated in Chapters 6 and 7), and focus and contrast. Included as well were several new features that are hypothesized to capture larger aspects of prosodic and discourse context, namely tonal contour properties and discourse position. Finally, following Litman (1996), word lemma information was added in an attempt to capture lexical “exceptions” or behaviors strongly correlated with particular lexical tokens.

5.4.1 Part-of-Speech Tags

The fundamental predictor in these experiments is lexical class or POS tags. Hand-corrected output from Brill’s tagger was used (see Section 4.2.1).

5.4.2 Part-of-Speech Classes

Different text-to-speech synthesis systems productively employ generalized POS information. Two levels of POS classes are widely used, namely broad class and function versus content word.

Broad Class

For the experiments, broad classes include: NOUN, VERB, ADJECTIVE, ADVERB, PREPOSITION, WH-WORD, CONJUNCTION, DETERMINER, PRONOUN, (other) FUNCTION WORD and DISFLUENCY. Since disfluent tokens belong to many lexical classes, they were classified by themselves. The remaining broad classes were chosen based on categories that Hirschberg (1993), following Altenberg (1987), used to distinguish between closed class items that are typically deaccented (e.g. existential there, conjunctions, accusative pronouns and modals) and those that are typically accented (e.g. negative modals, wh-words, nominal pronouns). In Hirschberg’s work (1993), Altenberg’s tag set was translated into Church’s (1988) tag set. For the present study, Hirschberg’s broad class categories built from Church’s tag set were reconstructed using the Penn Treebank tag set (Santorini, 1991), as presented in Table 5.1.

The broad classes used by Altenberg and Hirschberg can only be approximately reconstructed using the much smaller Penn Treebank tag set. For example, the Penn Treebank tag set does not mark case for pronouns, or separately tag negative modals (which are tagged as adverbs). The latter situation can be addressed by considering word lemma information in conjunction with POS tags in machine learning experiments. The former requires additional processing to label case. A general, syntactic approach for labeling grammatical
Table 5.1: Assignment of Penn Treebank POS tags to broad classes.

<table>
<thead>
<tr>
<th>Broad Class</th>
<th>Penn Treebank Tags in Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>adjective (adj)</td>
<td>JJ, JJR, JJS</td>
</tr>
<tr>
<td>adverb (adv)</td>
<td>RB, RBR, RBS</td>
</tr>
<tr>
<td>noun (n)</td>
<td>CD, NN, NNP, NNPS, NNS</td>
</tr>
<tr>
<td>verb (v)</td>
<td>VB, VBD, VBG, VBN, VBP, VBZ</td>
</tr>
<tr>
<td>conjunction (conj)</td>
<td>CC, IN (subordinating conjunction)</td>
</tr>
<tr>
<td>determiner (det)</td>
<td>DT, PDT</td>
</tr>
<tr>
<td>preposition (prep)</td>
<td>IN (except subordinating conj.), TO (except infinitival)</td>
</tr>
<tr>
<td>pronoun (pro)</td>
<td>PRP, PRP$</td>
</tr>
<tr>
<td>wh-word (wh)</td>
<td>WDT, WP, WPS, WRB</td>
</tr>
<tr>
<td>function word (fn)</td>
<td>EX, MD, RP, POS, TO (infinitival)</td>
</tr>
<tr>
<td>disfluency (disfl)</td>
<td>UH</td>
</tr>
</tbody>
</table>

function is described in Chapter 6. Finally, although the assignment of a Penn Treebank tag to its broad class is in most cases rather direct, there are several instances in which a single Penn Treebank tag is assigned to tokens that belong to two quite distinct broad classes. Two of these instances were disambiguated for this study by automatic processing of lexical and POS context. First, for the tag class IN, subordinating conjunctions (e.g. before, because) were distinguished from other prepositions. Also, infinitival instances of to were distinguished from other cases, which were assigned to the class of prepositions, doing away with the ad hoc tag class TO.

**Function versus Content Word Classes**

Given the above broad class system, the grouping of broad classes into function word classes and content word classes is rather straightforward. Table 5.2 provides the category assignments for broad classes used in this study. Again, disfluencies were kept distinct since disfluent tokens may be either function or content words.

Table 5.2: Function versus content assignment of broad classes.

<table>
<thead>
<tr>
<th>Part-of-Speech Category</th>
<th>Assignment of Broad Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>content</td>
<td>adj, adv, n, v</td>
</tr>
<tr>
<td>function</td>
<td>conj, det, prep, pro, wh, fn</td>
</tr>
<tr>
<td>disfluency</td>
<td>disfl</td>
</tr>
</tbody>
</table>

**5.4.3 Phrase, Sentence and Task Position**

Surface position information seems to play a similar role in different grammatical theories. In syntax and pragmatics, word order is said to express information about sentence topic, or givenness and newness of semantic entities. Functional syntax theories state that information is presented from old to new, with the exception that sentence-initial position may be marked in sentences that diverge from underlying word order. At the discourse level, the relative newness and givenness of discourse referents is said to affect accentuation.
(Terken and Hirschberg, 1994). The position of a word in a task or discourse serves as an indirect way to capture this notion of discourse salience. Finally, in some prosodic theories, phrase-final constituents are privileged, bearing NUCLEAR STRESS in citation-form sentences. Metrical stress theory (Liberman, 1975; Liberman and Prince, 1977) greatly elaborates on this notion, defining hierarchical, rhythmic constraints on the placement of pitch accents whose careful examination lies beyond the scope of this thesis. To capture primitive aspects of these prosodic constraints, each word is assigned features based on its position within the intermediate phrase and also within the intonational phrase. Following Hirschberg (1993), positional information for intermediate phrase, intonational phrase, sentence and task units was measured in number of words. In each case, relative distance from unit boundaries and total distance of a unit were represented by three features: distance from beginning of unit, distance from end of unit and total length of unit.

5.4.4 Tonal Contours

Two features of tonal contours were derived from the ToBI prosodic labeling. Each word was assigned a phrase accent feature, which was simply the phrase accent labeling (H-, !H- or L-) for the intermediate phrase to which the word belongs. Additionally, each word was assigned a boundary tone feature, which was the boundary tone labeling (H% or L%) for the intonational phrase to which the word belongs. These features have not been previously tested in pitch accent prediction studies. They were included in this study because of their hypothesized relationship to discourse connectivity (Pierrehumbert and Hirschberg, 1990).

5.4.5 NP-accent Assignment Using TTS

The problem of complex nominal accent assignment has received considerable attention from TTS researchers and phonologists. Hirschberg (1993) used a dedicated complex-nominal accent assignment module developed by Sproat (1990) to obtain information about the predicted accenting of complex nominals for her CART experiments on accent prediction. The complex-nominal accenting module was used to code whether the token was part of a simple noun phrase, or if part of a complex nominal, whether the token would be accented or deaccented by the complex-nominal TTS module. For the present study, similar features were derived using the NP-accenting module (Sproat, 1994) of the Bell Laboratories New Text-to-speech system (NewTTS) (Sproat and Olive, 1997). However, it was not possible to ascertain based on the intermediate representations open to inspection by the experimenter when NewTTS had identified complex nominals. Thus, a slight modification of Hirschberg’s NP-accenting features (1993) was made as follows: tokens were coded as either part of a noun phrase or outside of a noun phrase based on NewTTS’s syntactic bracketing, and if part of a noun phrase, the accenting prediction made by NewTTS for the token was recorded. Also, for this study, all accent prediction information was obtained by processing sentences through the NewTTS system in isolation. For longer discourses, NewTTS may apply contextual accentuation rules, while it is generally the case that single sentences in isolation are assigned citation-form accentuation.

5.4.6 Word Lemma

Finally, word lemma information was provided by the morphological processor of the NewTTS system. No hand-correcting of these data was carried out, although errors such as giving inbind as the word lemma for inbound arose infrequently. As noted, word lemma information allows the system to learn lexical “exceptions” or behaviors strongly correlated with particular lexical tokens. Litman (1996) showed that word lemma or token information improves the performance of automatically learned systems for cue phrase disambiguation. Hand-crafted lists of lexical exceptions have often been made use of in accent assignment systems. For example, lexical information has been used to capture exceptions to POS category accentuation rules (Altenberg, 1987; Hirschberg, 1993), to trigger the deaccentuation of situationally given words in the
domain (Horne et al., 1993) and to list semantically “light” words that are commonly deaccented (Zacharski et al., 1992). Word lemma information has not been previously examined in machine learning studies on accent prediction.

5.5 Baseline Experiments

Experiments were run on the three sets of labeled speech data described in Section 4.3, namely H1-read (2,244 tokens), H1-spon (2,359 tokens) and H3-spon (1,616 tokens). First, POS and word lemma information were explored. A basic lexical model, referred to as Lex, was chosen. New classes of features were added to the Lex model, singly and collectively, to better determine the contributions of each feature class and their interactions.

All RIPPER experiments were run using ten-fold cross-validation, with two rule optimization passes in noisy data mode. In \( k \)-fold cross-validation, one complete training set is randomly partitioned into \( k \) equal data sets. Training and testing is then performed \( k \) times, each time holding out one of the ten data sets for testing, and training on the remaining \( k - 1 \) data sets. The testing results are then averaged over the \( k \) runs. This technique enables reasonable error rate estimation for small to moderately sized corpora. Results from RIPPER experiments are reported in terms of average classification success rate and standard deviation for the ten cross-validation runs.

Statistically significant differences in performance between two classification systems can be inferred by computing the 95% confidence interval for each system. Following Cohen’s proposal adopted by Litman (1996), it is inferred with \( p < .05 \) that a classification system, S1, performs statistically significantly better than another, S2, if the lower bound for S1’s performance for the 95% confidence interval does not overlap with the upper bound performance for S2. For classification systems whose performance is measured by a single error rate on a test set, a normal approximation to the binomial distribution is assumed and the 95% confidence interval is computed by adding and subtracting the standard error of 2.0 standard deviations from the observed performance figure. For cross-validated performance measures, the assumption that the experimental results can be modeled by a normal curve is discarded and a more conservative estimate of performance is computed using the Student’s curve approximation. Given 9 degrees of freedom for the 10-fold cross-validation experiments, the Student’s t-table value is 2.26 standard deviations for the standard error at \( p < .05 \), and 3.25 standard deviations at \( p < .01 \).\(^1\)

5.5.1 Part-of-Speech Tags and Classes

The Penn Treebank POS tags, broad class assignments and function versus content word class assignments were tested separately as sole predictors of accentuation. Table 5.3 gives performance figures. Results for H1 are comparable to Hirschberg’s (1993) result of 77% correct for hand-crafted classifications of broad class POS categories on the radio news corpus. Her result of 68% correct for the function-content word distinction approaches the performance of RIPPER on H3-spon, using either function versus content, broad class or POS tag information. The most striking result on the Boston Directions Corpus data is that all three POS models performed consistently on the three sets of speech. Comparing results on the three sets, two trends emerge. First, while classification rates for H3-spon are close to Hirschberg’s results for function versus content word classification systems, they are also approximately 10% lower than the classification

\(^1\)For classification systems that are learned by training and testing on separate data sets, Cohen (1995) has proposed another measure of significance. A paired test is done on the two sets of predictions generated by the different systems on the same test set. Significance is then measured as how far the probability of one system “winning”, i.e. generating a correct prediction when the other system generates an incorrect prediction, is from 0.5.
Table 5.3: Classification rates and standard deviations for POS experiments.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>H1-read</th>
<th>H1-spon</th>
<th>H3-spon</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS tags</td>
<td>78.42% ± 1.13%</td>
<td>77.45% ± 0.37%</td>
<td>68.38% ± 0.88%</td>
</tr>
<tr>
<td>Broad class</td>
<td>78.56% ± 0.98%</td>
<td>79.05% ± 0.66%</td>
<td>67.21% ± 0.72%</td>
</tr>
<tr>
<td>Fn vs. content</td>
<td>78.56% ± 0.98%</td>
<td>79.05% ± 0.66%</td>
<td>68.25% ± 0.97%</td>
</tr>
</tbody>
</table>

Table 5.4: Classification rates and standard deviations for POS plus word lemma experiments.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>H1-read</th>
<th>H1-spon</th>
<th>H3-spon</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS tags/lemma</td>
<td>79.58% ± 1.05%</td>
<td>80.88% ± 0.48%</td>
<td>69.50% ± 0.95%</td>
</tr>
<tr>
<td>Broad/lemma</td>
<td>79.36% ± 1.01%</td>
<td>80.70% ± 0.80%</td>
<td>68.76% ± 1.26%</td>
</tr>
<tr>
<td>Fn-content/lemma</td>
<td>79.18% ± 1.01%</td>
<td>80.91% ± 0.83%</td>
<td>69.43% ± 1.12%</td>
</tr>
</tbody>
</table>

rates for both H1 data sets. It is possible that the smaller number of tokens for training in H3-spon is lowering performance somewhat, but it is unlikely that this factor alone could so drastically reduce prediction accuracy. A second observation is that the prediction results for H1-read and H1-spon are not statistically different. This demonstrates that for at least this speaker, accent prediction for spontaneous speech is not more difficult than for read speech. This represents an original finding for speech that is spontaneously elicited from and read by the same speaker, and suggests that differences in speaking style may be speaker dependent, cf. (Ayers, 1994).

5.5.2 Adding Word Lemma

As noted, word lemma information has not been previously examined in machine learning studies on accent prediction although it has been thought to be useful especially for identifying deaccented words. Word lemma information was tested in RIPPER in combination with POS tag information. Performance figures are shown in Table 5.4. The results indicate that adding word lemma information significantly increases accuracy for the H1-spon POS tag model, although the improvement is slight — approximately one to two percent. The upper bound performance for the H1-spon POS tag model at $p < .01$ is 78.65% using Student’s curve. The lower bound performance at $p < .01$ for the H1-spon POS tag plus lemma model is 79.32% using Student’s curve. Since the upper bound performance for the H1-spon POS tag model is lower than the lower bound performance figure for the H1-spon POS tag plus lemma model, it can be concluded with 99% confidence that the addition of the word lemma feature aids accent prediction based on POS tags. Accuracy is generally increased for both H1-read and H3-spon as well, but these trends are not significant. The absence of other significant gains from word lemma information suggests that generalizations cannot be drawn, although it is most often the case that word lemma information does not significantly improve accent prediction.

Nevertheless, it is interesting to probe beyond these quantitative findings and study qualitative differences in the kind of word lemma information used in modeling accentuation for H1-spon versus H3-spon. Lemma rules learned for H1-spon are limited to accenting rules for verbal prepositions (i.e. get off) and semantically heavy prepositions (i.e. past, toward, down). The only additional lemma rules learned concern typically accented verbal tokens (i.e. travel, go, get). While similar rules are learned for H3-spon (i.e. accenting of particles down and out and potentially contrastive verbs such as walk), several other kinds of
lexically triggered rules arise. For example, a variety of deaccenting rules were learned for H3-spon for words that often occur in complex nominals, such as Line and Street, which occur repeatedly in compounds such as the Blue Line and Cowperwaithe Street. Deaccentuation is also learned for a number of specific function words and semantically light words, including the tokens and, be, come, just, the, which and you. These findings suggest that while previous work on accentuation has focused on POS information, seemingly subtle syntactic and semantic phenomena related to accentuation can be uncovered by considering word lemma information in addition to POS tags.

5.5.3 Adding Linear Position and Tonal Contour Features

Because the results for different POS and word lemma models were not significant, one model was chosen as a basic model for further testing. This model, which will be referred to as LEX, included broad class POS and word lemma features. Sentence position, (intermediate and intonational) phrase position, task position and tonal contour features were added to the basic Lex model, both singly and all together. The model including the Lex model plus all features together is referred to as the LEX ENHANCED model. Table 5.5 summarizes performance figures. The only significant finding from these experiments was that accent assignment in H3-spon is extremely sensitive to prosodic phrasing information. The model, Lex plus phrase position, gives the best performance of all models tested for H3-spon. It is significantly better than the Lex model alone at \( p < .01 \). Given that the t-table value for nine degrees of freedom is 3.25 at \( p < .01 \), the upper bound performance for the Lex model on H3-spon, is 72.86%, while the lower bound performance for the Lex plus phrase position model on H3-spon is 73.93%. Performance of the Lex enhanced model for H3-spon is also significantly better than the Lex model at \( p < .05 \), but it appears that the availability of phrase position information in the Lex enhanced model is responsible for the improvement.

Inspection of the rule sets built on each of the ten cross-validation runs for the Lex plus phrase position model revealed that speaker H3 shows a clear preference for placing nuclear stress at the end of intermediate phrases. The two best rule sets that were output by RIPPER are shown in Figure 5.1. Both return a classification rate on heldout data of 78.26%, which is comparable to the classification rates for H1. In Figure 5.1, each rule, in Prolog-style Horn clause form, is followed in parentheses by the total number of tokens in the training data that satisfy the conditions on the right-hand side of the rule and then the number of tokens incorrectly predicted by the rule to belong to the class on the left-hand side of the rule. The rule, \texttt{acc :- broadclass=n}, for example, can be read, \textit{If the broad class POS category for the token is NOUN, then the token is accented}. In the training set, this rule was applied correctly to 79 out of 148 tokens, and incorrectly applied to the remaining 69 tokens in the noun broad class category.

Although the rule sets appear to be very different, the single overriding principle at work in both rule sets is that intermediate phrase final tokens are accented. In the rule set from run 2, the first rule, \texttt{acc :- interfromend \leq 1}, states that a token that is in the final position of an intermediate phrase is classified as accented. The following rules in the rule set classify tokens based on POS information. In run 5, a

<table>
<thead>
<tr>
<th>Experiment</th>
<th>H1-read</th>
<th>H1-spon</th>
<th>H3-spon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lex/sentence pos</td>
<td>78.87% ± 1.04%</td>
<td>80.40% ± 0.80%</td>
<td>68.32% ± 0.95%</td>
</tr>
<tr>
<td>Lex/phrase pos</td>
<td>79.98% ± 1.26%</td>
<td>81.00% ± 0.67%</td>
<td>75.81% ± 0.58%</td>
</tr>
<tr>
<td>Lex/task pos</td>
<td>79.49% ± 1.00%</td>
<td>80.83% ± 0.75%</td>
<td>69.25% ± 0.80%</td>
</tr>
<tr>
<td>Lex/tonal contour</td>
<td>78.96% ± 1.08%</td>
<td>80.70% ± 0.70%</td>
<td>67.94% ± 0.78%</td>
</tr>
<tr>
<td>Lex enhanced</td>
<td>79.59% ± 1.30%</td>
<td>80.02% ± 0.86%</td>
<td>74.19% ± 0.88%</td>
</tr>
</tbody>
</table>

Table 5.5: Classification rates and standard deviations for enhanced Lex models.
complementary set of rules is learned to predict the unaccented class of tokens. A number of rules are learned that contain the condition that unaccented tokens must not occur in the intermediate phrase final position, \( \text{interfromend} \geq 2 \). No such dominating principle concerning the accenting of intermediate phrase final tokens is learned for speaker H1, whose intermediate phrases are fewer in number and greater in length (see Table 4.2). Similarly, no overriding effect of intermediate phrase final position was found in Hirschberg’s CART experiments (1993) for either the single-speaker read news corpus or the multi-speaker ATIS corpus. However, more subtle boundary effects for intermediate and intonational phrases as well as sentences did arise in Hirschberg’s experiments, often interacting with POS information. These findings suggest that accent assignment algorithms that ignore the intermediate phrase unit may fail to capture certain strategies for accentuation, but these strategies are speaker-dependent.

5.5.4 Adding NP-accent Assignment

Next, the NP-accent predictions from the TTS system were added to both the basic Lex model, and to the enhanced Lex model with all of the features mentioned in Section 5.5.3. Table 5.6 summarizes performance figures. NP-accent information from the TTS system improves performance more than any individual feature class added to the Lex model in Section 5.5.3 (excepting the contribution of prosodic phrase features to H3-spon), but not significantly so. Further, no significant performance gain is to be had by adding NP-accent information to the enhanced Lex model, whose classification success rates are repeated in the last line of Table 5.6 for ease of reference. Hirschberg (1993) reported that complex-nominal accent prediction was correct 59% of the time, contributing 2% of correct predictions and 6% of incorrect predictions to the overall result of 82.4% for the hand-crafted system.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>H1-read</th>
<th>H1-spon</th>
<th>H3-spon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lex/TTS</td>
<td>80.52%</td>
<td>81.58%</td>
<td>69.87%</td>
</tr>
<tr>
<td>Lex enhanced/TTS</td>
<td>80.87%</td>
<td>81.08%</td>
<td>75.20%</td>
</tr>
<tr>
<td>Lex enhanced</td>
<td>79.59%</td>
<td>80.02%</td>
<td>74.19%</td>
</tr>
</tbody>
</table>

Table 5.6: Classification rates and standard deviations for NP-accent feature experiments.

5.6 Assessment

Based on Hirschberg’s corpus studies, it is clear that pitch accent assignment for short samples of citation speech is for practical purposes a solved problem (Hirschberg, 1993). However, Hirschberg’s results on spontaneous dialogue speech and extended radio news stories show that room for improvement exists in the modeling of accentuation in both short segments of spontaneous speech and longer discourses. Discovery of further useful cues is of course the motivation for the research at hand, although research progress can only be assessed by building solidly on previous work.

Machine learning experiments reported in this chapter comprehensively tested linguistic features that are already known to perform an excellent job of predicting accentuation for citation-form speech. Results using the rule learning system called RIPPER show that, for the one speaker examined, spontaneous speech performance is not significantly better or worse than read. Further, drastic individual differences in overall prediction performance as well as learned rule sets arise between the two speakers whose speech is examined. For one speaker, adding in any one extra feature to the basic Lex model fails to significantly
decrease the error rate. However, for the second speaker, including prosodic phrasing information gives a performance boost of approximately 7%.

That being said, studies on speaker H1 give results for accent prediction of approximately 80% correct on extended, non-professional spontaneous discourses up to 631 words in length. This performance is comparable to that reported by Hirschberg (1993) for professionally read news speech of similar average length. Experiments on the Boston Directions Corpus thus demonstrate the feasibility of using current accent prediction techniques on long discourses in the spontaneous speaking style. The present study also shows that broader contextual features, such as task position and tonal markings at higher prosodic levels, do not significantly aid accent prediction. Thus, experimental performance has reached a plateau not far beyond that of current TTS technology for accent assignment. Climbing beyond this plateau is the aim of the next set of empirical studies on the Boston Directions Corpus, reported in Chapters 6 and 7.
H3-spon rule set for run 2: Lex plus phrase position

\[
\begin{align*}
\text{acc} &: \text{interfromend} \leq 1 \ (386/120), \\
\text{acc} &: \text{broadclass} = \text{n} \ (148/69), \\
\text{acc} &: \text{broadclass} = \text{adj} \ (27/9), \\
\text{acc} &: \text{broadclass} = \text{adv} \ (22/17), \\
\text{default unacc} &: (521/136).
\end{align*}
\]

H3-spon rule set for run 5: Lex plus phrase position

\[
\begin{align*}
\text{unacc} &: \text{interfromend} \geq 2, \text{broadclass} = \text{det} \ (121/11), \\
\text{unacc} &: \text{interfromend} \geq 2, \text{interfromend} \geq 3, \text{lemma} = \text{you} \ (47/1), \\
\text{unacc} &: \text{interfromend} \geq 2, \text{intonfromend} \geq 5, \text{intertotal} \leq 5, \text{intertotal} \geq 4, \text{interfrombeg} \leq 1 \ (33/5), \\
\text{unacc} &: \text{interfromend} \geq 2, \text{broadclass} = \text{fn} \ (58/3), \\
\text{unacc} &: \text{interfromend} \geq 2, \text{broadclass} = \text{prep} \ (94/33), \\
\text{unacc} &: \text{interfromend} \geq 2, \text{broadclass} = \text{conj} \ (31/5), \\
\text{unacc} &: \text{interfromend} \geq 2, \text{broadclass} = \text{v}, \text{lemma} = \text{be} \ (18/0), \\
\text{unacc} &: \text{interfromend} \geq 2, \text{intonfromend} \geq 5, \text{broadclass} = \text{v} \ (38/20), \\
\text{unacc} &: \text{interfromend} \geq 2, \text{broadclass} = \text{pro}, \text{interfrombeg} \geq 2 \ (22/2), \\
\text{unacc} &: \text{intertotal} \geq 2, \text{lemma} = \text{Line} \ (21/3), \\
\text{unacc} &: \text{interfromend} \geq 2, \text{broadclass} = \text{disfl} \ (12/3), \\
\text{unacc} &: \text{intertotal} \geq 2, \text{lemma} = \text{Street} \ (17/3), \\
\text{unacc} &: \text{intertotal} \geq 2, \text{lemma} = \text{ll} \ (10/0), \\
\text{default acc} &: (639/205).
\end{align*}
\]

<table>
<thead>
<tr>
<th>Code</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>broadclass</td>
<td>broad class POS category</td>
</tr>
<tr>
<td>interfrombeg</td>
<td>distance in words from beginning of intermediate phrase</td>
</tr>
<tr>
<td>interfromend</td>
<td>distance in words from end of intermediate phrase</td>
</tr>
<tr>
<td>intonfromend</td>
<td>distance in words from end of intonational phrase</td>
</tr>
<tr>
<td>intertotal</td>
<td>total length in words of intermediate phrase</td>
</tr>
<tr>
<td>lemma</td>
<td>word lemma</td>
</tr>
</tbody>
</table>

Figure 5.1: Learned rule sets for H3-spon, Lex plus phrase features.
Chapter 6

Task-oriented Monologue Study: Part II

This chapter describes the analysis of syntactic, lexico-syntactic, lexico-semantic, semantic and discourse features on the Boston Directions subcorpus. These higher-level features build on the lower-level features derived from basic speech and text processing, which was described in Chapter 4. Their contributions to accent prediction will be tested in machine learning experiments presented in Chapter 7.

6.1 Syntactic Analysis

In this study, the major function served by syntactic parsing is to identify NP constituents for further analysis. For this reason, a parser was chosen based on its ability to produce parsimonious structures exhibiting consistent NP bracketing. A secondary consideration was that the parser be trained on the Penn Treebank POS tag set, to take full advantage of the hand-corrected POS labeling of the Boston Directions subcorpus reported in Section 4.2.1.

The best reported constituent parsing results on the Penn Treebank Wall Street Journal corpus to date have been achieved by Collins’ lexical dependency parser (Collins, 1996). Collins’ parser is novel in that it trains on bigram lexical dependencies between head-words in a parse tree. A second interesting feature of Collins’ parser is that it uses the notion of BASENP in its statistical model to delimit head-word domains within complex NPs, and it preserves baseNP bracketings in its parse trees. A baseNP is a minimal NP which is non-recursive, meaning none of its child constituents are NPs. Collins illustrates this notion with the following example, in which baseNPs are in square brackets, and heads of baseNPs are in capitalized font (Collins, 1996, p. 185):

[John SMITH] [the PRESIDENT] of [IBM] announced [his RESIGNATION.]

For conjoined NPs, it is stipulated that the first conjunct contains the head-word for the larger conjoined NP (Collins, 1996).

This notion of baseNP is implicit in certain text-to-speech accent assignment rules, such as Horne’s (1987). Horne’s rules, for example, assign prominence to baseNP units within complex NPs in a top-down recursive manner. What is important to observe about such systems is that decisions on whether to accent a baseNP within a complex NP are independent of other accent decisions for baseNPs in the same complex NP.\(^1\) It seems worthwhile to explore whether reasonable accent prediction can be achieved based on baseNP

\(^1\) The order in which accent is assigned does matter for determining how to accent a particular word in Horne’s system (1987), since the prominence ratio (accent height relative to top-line) is decreased each time accent is assigned within a sentence. This issue of relative prominence does not relate to the present concern with predicting the presence or absence of prominence.
units, because this approach makes fewer demands of syntax than do full implementations of theories of accenting based on syntactic focus projection, e.g. (Rochemont, 1986; Selkirk, 1993).

Collins’ parser was applied to the tagged versions of the analyzed sentences from the spontaneous speech from speakers H1 and H3, H1-spon and H3-spon. Before parsing, syntactic disfluencies were hand-corrected by deletion of the reparandum and of all filled pauses and hesitation phenomena marked by the tag UH, following independently devised guidelines (Nakatani and Shriberg, 1993). The bigram lexical dependency model for the parser was trained by Collins on sections 02-21 of the Penn Treebank Wall Street Journal corpus, which is approximately 40,000 sentences (Collins, 1996). The parser was also configured to ignore punctuation information, since the Boston Directions transcriptions lack sentence-internal punctuation. To give a flavor of the output from Collins’ parser, a parse for a sentence from the corpus is given below: This

```
TOP S CC And
ADVP RB here
NP PRP you
VP MD will
VP VB see
NP NP NNP Paine
NP NNP Hall
NP DT a
NN music
NN building
```

Figure 6.1: Parse tree produced by Collins’ parser for sentence in the Boston Directions Corpus.

parse tree contains three baseNPs, you, Paine Hall and a music building. The two NPs, Paine Hall and a music building, together make up one complex NP.

The strategy taken in using this parser output was to take the baseNP as the minimal unit for accent prediction, but also to label baseNPs for headedness and constituency information. If accent assignment in complex NPs is more complex than is modeled by Horne (1987), it would be necessary to explore fuller representations of the syntactic structure of complex NPs and to more closely examine the accentuation patterns in simple-baseNPs versus head-baseNPs and child-baseNPs in complex NPs.

### 6.2 Lexico-syntactic and Lexico-semantic Analysis of Noun Phrase Referring Expressions

#### 6.2.1 Grammatical Function

The grammatical function of all arguments and adjuncts of verbal predicates was coded as one of the following: subject (SBJ), direct object (DO), indirect object (IO), adjunct (ADJ) or predicate complement (PC). Objects of verbal prepositions (e.g. get off the Red Line) were classified as direct objects, since they function grammatically as verb complements rather than adjuncts. That is, the sentence, The Red Line is where I get off, is grammatical, whereas prepositional objects in adjuncts cannot usually be preposed (e.g. *Sunday is when I go to the movies on). Finally, all baseNPs in complex NP arguments inherited the grammatical function label of their root NP in the syntactic parse tree.
6.2.2 Form of Referring Expression

The Boston Directions Corpus contained a greater variety of forms of referring expression than did the spontaneous narrative. These forms were coded exhaustively as shown in Table 6.1. Several classes of referring expression, such as adverbial NP, verbal noun, and cardinal expression, did not arise in the spontaneous narrative corpus.

<table>
<thead>
<tr>
<th>Code</th>
<th>Form of Referring Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>adv</td>
<td>adverbial NP (e.g. here)</td>
</tr>
<tr>
<td>ord</td>
<td>ordinal expression (e.g. June third)</td>
</tr>
<tr>
<td>def</td>
<td>definite NP</td>
</tr>
<tr>
<td>dem</td>
<td>demonstrative NP</td>
</tr>
<tr>
<td>dempro</td>
<td>demonstrative pronoun</td>
</tr>
<tr>
<td>expl</td>
<td>expletive there</td>
</tr>
<tr>
<td>gen</td>
<td>generic</td>
</tr>
<tr>
<td>indef</td>
<td>indefinite NP</td>
</tr>
<tr>
<td>name</td>
<td>proper name</td>
</tr>
<tr>
<td>poss</td>
<td>possessive NP</td>
</tr>
<tr>
<td>pro1</td>
<td>first person pronoun</td>
</tr>
<tr>
<td>pro2</td>
<td>second person pronoun</td>
</tr>
<tr>
<td>pro3</td>
<td>third person pronoun</td>
</tr>
<tr>
<td>quant</td>
<td>bare quantifier (e.g. any, most, some)</td>
</tr>
<tr>
<td>vn</td>
<td>verbal noun (e.g. make a turn)</td>
</tr>
</tbody>
</table>

Table 6.1: Coding system for form of referring expression.

6.3 Coreference Coding of Noun Phrase Referring Expressions

Coreferential relations for the Boston Directions Corpus were coded using NPs as referential units. Four kinds of coreference relations were analyzed: strict coreference, inferential links, modificational relations and a special relation between conjoined NPs and individual NP conjuncts.

Two NPs are in a strict coreference relationship, represented as “IDENT(np1, np2)”, when np2 occurs after np1 in the discourse and realizes the same discourse entity that is realized by np1. In addition to strict coreference, two NPs can be related by inferential links. This occurs when the discourse entities realized by the two NPs are semantically related by means of limited inference. For the Boston Directions Corpus, four inferential relations were recorded: “Part-Whole(np1, np2)”, “Whole-Part(np1, np2)”, “Set-Member(np1, np2)” and “Member-Set(np1, np2)”. Examples of these relations from the corpus are: Part-Whole(Green Line, the T) and Set-Member(three restaurants, Chef Chow’s). Finally, certain modificational relationships between two NPs are noted as “Modificational(np1, np2)”. These occur when np2 serves to help identify np1, as is the case for appositives (e.g. the store, my favorite coffee shop) and copular relative clauses (e.g. the store which is my favorite coffee shop). Modificational relationships also occur when np1 (or np2) is a modifier in a compound NP, and np2 (or np1) corefers not with the entire compound NP but with np1 (or np2). An example in the directions corpus is the relationship between the Museum stop and the Museum in the following excerpt:
When you get off at the **Museum stop**
you will be able to see the **Museum** off to your right.

The NP, *the Museum*, provides a name for the subway stop, but arguably subway stops are given names that help identify them, and so these cases involve the same identifying purposes as do appositives and copular relative clauses. Finally, the treatment of conjoined NPs involves one more relation. Conjoined NPs were counted as a single NP referential unit when they filled one NP argument position. When a NP conjunct, *np2*, from a previously mentioned conjoined NP, *np1*, is referred to independently of *np1*, the relation, “List-Item(*np1, np2*)” applies. The following sentences contain an example.

You will pass through *Kendall, Charles and Park Street Stations*. At *Park Street Station* you can change to the Green Line, but you want to continue on the Red Line until you get to South Station.

The first italicized conjoined NP, *Kendall, Charles and Park Street Stations*, is in a List-Item relation with the second italicized NP, *Park Street Station*. The inverse relation, Item-List, occurs as well in the corpus.

Coreference coding was carried out on H1-spon and H3-spon by hand-labeling using DTT (Discourse Tagging Tool), a GUI coreference annotation tool developed by Aone and Bennett (1995).²

### 6.4 Discourse Segmentation Analysis

In the Boston Directions Study, the Grosz and Sidner theory of discourse structure (1986), hereafter G&S, provides a foundation for segmenting discourses into constituent parts. While intention-based segmentation was carried out on the spontaneous narrative, the length of the narrative — over 2,000 words — posed difficulties. In longer discourses, there is a greater need to assess the reliability of discourse structure analyses. The AP news stories analyzed by Grosz and Hirschberg (1992) averaged 450 words in length, while the task-oriented speech segments studied by Brown (1983) averaged a few hundred. The Boston Directions discourses from H1-spon and H3-spon range from 43 to 631 words in length. To obtain reliable segmentation data for discourses of this length, it was necessary to analyze segmentations from multiple subjects, along the lines of previous work by Grosz and Hirschberg (1992). The method of collection and reliability of the data obtained for the Boston Directions Corpus are reported in Section 6.4.1 and Section 6.4.2.

#### 6.4.1 Methods

To investigate intonational correlates of discourse structure, Grosz and Hirschberg (1992) devised a set of instructions based on G&S for labeling the intentional and linguistic structures at both the local and global levels. Two modes of discourse segmentation were employed by subjects, all of whom were experts in the G&S theory. Three subjects labeled from text alone and three labeled from text and speech. Other than this difference in the availability of speech, all subjects received identical instructions. The text for each task was presented with one intermediate phrase per line, where intermediate phrase boundaries were equated with ToBI break index labels of level 3 or higher (Pitrelli, Beckman, and Hirschberg, 1994). Subjects were thus constrained to place discourse segment boundaries only at the boundaries between intermediate phrases.³

²Coreference coding was done by David Ahn under the supervision of the author. Discussions with David Ahn helped refine the original coding scheme.

³It has been observed that the intonational phrase is too large a unit for doing discourse segmentation (Hirschberg and Grosz, 1992; Swerts, 1995). In Swerts’ study (1995), subjects were allowed to place discourse segment boundaries between any pair of adjacent words. There was an abundance of cases where human subjects judged discourse segment boundaries to occur in the middle of intonational phrases, while this was largely not the case for intermediate phrases.
In the instructions, subjects were asked to analyze the linguistic and intentional structures by segmenting the discourse and specifying the hierarchical relationships among segments. In addition, local aspects of discourse structure, such as parenthetical expressions and referential information, were also labeled. Significant inter-labeler agreement was obtained using percent agreement as the reliability measure. These segmentations were carried out in a text editor, with speech playback capabilities provided in the Entropics wave+ environment (Talkin, 1989).

These instructions were substantially modified for use on the Boston Directions Corpus (Hirschberg and Nakatani, 1996), but were still geared toward analysts who were versed in the terminology and theory of G&S. One crucial enhancement was that analysts were required to write down explicit DSPs for each segment. These so-called expert segmentations from an earlier study (Hirschberg and Nakatani, 1996), comprise the expert data for this study. Appendix B contains the exact instructions used for expert segmentations on the Boston Directions.

6.4.2 Segmentation Results on Expert Data

In previous work, expert segmentation data on speaker H1 were extensively analyzed and were found to be reliable for text-and-speech labeling mode (Hirschberg and Nakatani, 1996). Following Grosz and Hirschberg’s procedures discussed in Section 6.4.1, two modes of segmentation were studied, namely text-alone and text-and-speech. For each mode, segmentations were obtained from three labelers. The reliability of the expert segmentation data for H1-read and H1-spon is discussed below.

Raw Agreement

Raw agreement among all three labelers provides a stricter measure of labeler agreement than does percent agreement, which has been reported in other segmentation studies (Hirschberg and Grosz, 1992; Passonneau and Litman, 1997). Labels on which all labelers in the group agreed are termed the consensus labels. The consensus labels for segment-initial (SBEG), segment-final (SF), and segment-medial (SCONT, defined as neither SBEG nor SF) phrase labels are given in Table 6.4.2.

Text-alone (group T) and text-and-speech (group S) segmentations differ significantly, in contrast to earlier findings by Grosz and Hirschberg (1992) on a corpus of read-aloud news stories and in support of Swert’s informal findings (1995). Table 6.4.2 shows that group S produced significantly more consensus boundaries for both read (p<.001, χ²=58.8, df=1) and spontaneous (p<.001, χ²=55.4, df=1) speech than did group T. When the read and spontaneous data were pooled, group S agreed upon significantly more SBEG boundaries (p<.05, χ²=4.7, df=1) as well as SF boundaries (p<.05, χ²=4.4, df=1) than group T. Further, it is not the case that text-alone segmenters simply chose to place fewer boundaries in the discourse; if this were so, then we would expect a high percentage of SCONT consensus labels where no SBEGs or

---

4 The author thanks the following segmenters for their donated time and expertise: Nancy Chang, Andy Kehler, Karen Lochbaum, Candy Sidner, Lisa Stifelman and Gregory Ward.

5 The analyses of expert segmentation data reported in this section were carried out in collaboration with Julia Hirschberg and have been previously published (Hirschberg and Nakatani, 1996).

6 Use of consensus labels is a conservative measure of labeler agreement. Other studies (Passonneau and Litman, 1997; Swerts, 1995) have shown that with a larger number of labelers, notions of boundary strength can be employed by computing the percentage of labelers who marked a discourse boundary at each potential boundary site. Boundaries marked by a higher percentage of labelers are considered stronger than those marked by a lower percentage of labelers. Various statistical measures can be used to determine a significant threshold of labeler agreement for identifying strong boundaries (Carletta, 1996; Passonneau and Litman, 1997).

7 Consensus percentages for the three types in Table 6.4.2 do not necessarily sum to the total consensus agreement percentage, since phrases may be both segment-initial and segment-final.
SFs were identified. Instead, we find that the number of consensus SCONTs was significantly higher for the labelings from text and speech than for labelings from text alone ($p<.001$, $\chi^2=49.1$, df=1).

So, it appears that the speech signal can help disambiguate among alternate segmentations of the same text. Furthermore, the data in Table 6.4.2 show that spontaneous speech can be segmented as reliably as read, contrary to Ayers’ previous results (1994).

### Inter-labeler Reliability

Comparisons of inter-labeler reliability, that is, the reproducibility of a coding scheme given multiple labelers, provide another perspective on the segmentation data. How best to measure inter-labeler reliability for discourse segmentation tasks, especially for hierarchical segmentation, is an open research question (Passonneau and Litman, 1997; Carletta, 1996; Flammia and Zue, 1995; Swerts, 1995). For comparative purposes, several measures proposed in the literature, namely COCHRAN’S $Q$ and Siegel and Castellan’s $\kappa$ COEFFICIENT (Siegel and Castellan, 1988), were explored. Cochran’s $Q$, originally proposed in (Hirschberg and Grosz, 1992) to measure the likelihood that similarity among labelings was due to chance, was not useful in the current study; all tests of similarity using this metric (pairwise, or comparing all labelers) gave probability near zero. This statistic did not serve, for example, to capture the differences observed between labelings of text alone versus labelings from text and speech.

Recent discourse annotation studies (Isard and Carletta, 1995; Flammia and Zue, 1995) have measured reliability using the $\kappa$ coefficient, which factors out chance agreement taking the expected distribution of categories into account. This coefficient is defined as

$$\kappa = \frac{P_O - P_E}{1 - P_E}$$

where $P_O$ represents the observed agreement and $P_E$ represents the expected agreement. Typically, values of .7 or higher for this measure are indicative of good reliability, with values of .8 or greater indicating high reliability. Isard and Carletta (1995) report pairwise $\kappa$ scores ranging from .43 to .68 in a study of naive and expert classifications of types of ‘moves’ in the Map Task dialogues (Anderson et al., 1991). For theory-neutral discourse segmentations of information-seeking dialogues, Flammia and Zue (1995) report an average pairwise $\kappa$ of .45 for five labelers and of .68 for the three most similar labelers.

An important issue in applying the $\kappa$ coefficient is how one calculates the expected agreement using prior distributions of categories. For the Boston Directions data, the prior probabilities were calculated based simply on the distribution of SBEG versus non-SBEG consensus labels by all labelers for one of the nine direction-giving tasks in this study, with separate calculations for the read and spontaneous versions. This task, which represented about 8% of the data for both speaking styles, was chosen because it was midway in complexity and in length among all the tasks. Using these distributions, $\kappa$ coefficients were calculated for each pair of labelers in each condition for the remaining eight tasks in the corpus. The observed percentage

### Table 6.2: Percentage of Consensus Labels by Segment Boundary Type

<table>
<thead>
<tr>
<th></th>
<th>SBEG</th>
<th>SF</th>
<th>SCONT</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>READ (N=494)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Text alone (T)</td>
<td>14%</td>
<td>11%</td>
<td>32%</td>
<td>57%</td>
</tr>
<tr>
<td>Text &amp; Speech (S)</td>
<td>18%</td>
<td>14%</td>
<td>49%</td>
<td>80%</td>
</tr>
<tr>
<td>SPON (N=552)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Text alone (T)</td>
<td>13%</td>
<td>10%</td>
<td>40%</td>
<td>61%</td>
</tr>
<tr>
<td>Text &amp; Speech (S)</td>
<td>15%</td>
<td>13%</td>
<td>54%</td>
<td>81%</td>
</tr>
</tbody>
</table>
of SBEG labels, prior distribution for SBEG, averaged pairwise $\kappa$ scores and standard deviations for those scores are presented in Table 6.4.2. The average $\kappa$ scores for group T segmenters indicate weak inter-labeler reliability. In contrast, average $\kappa$ scores for group S are .8 or better, indicating a high degree of inter-labeler reliability. Thus, application of this somewhat stricter reliability metric confirms that the availability of speech critically influences how listeners perceive discourse structure.

The calculation of reliability for SBEG location in effect tests the similarity of linearized segmentations and does not speak to the issue of how similar the labelers’ hierarchical labeling is. That is, if it is assumed that a new segment begins at each consensus SBEG label, then a flat consensus segmentation structure for the discourse can be constructed from multiple labelers’ segmentations. Although this CONSENSUS LINEAR SEGMENTATION provides useful data for the machine learning experiments in this thesis, it is only a crude reflection of the hierarchical structure provided by each individual segmentation. Flammia has proposed a method for generalizing the use of the $\kappa$ coefficient for hierarchical labeling, providing an upper-bound estimate on inter-labeler agreement.

This metric was applied to the Boston Directions segmentation data, to better determine the reliability of the discourse annotation conventions. Weighted averages of pairwise $\kappa$ scores were calculated from the three pairwise scores computed for each task. Results for each condition, together with the lowest and highest average $\kappa$ scores over the tasks, are presented in Table 6.4.2. Once again, averaged scores of .7 or better for text and speech labelings, for both speaking styles, indicate markedly higher inter-labeler reliability than do scores for text-alone labelings, which averaged .51 and .53.

In should be noted that all of the high-level factors investigated in the narrative study are represented in some manner in the new set of higher-level features. Some factors, such as form of referring expression and grammatical function, are represented directly. Others, such as global focusing status, can be derived from coreference coding and discourse segmentation analyses considered together. The exact representation of these features and their use in machine learning experiments on the role of higher-level linguistic features in accent prediction are described in the following chapter.

---

8 Flammia uses a flexible definition of segment match to calculate pairwise observed agreement: a segment in one segmentation is matched if both its SBEG and SF correspond roughly to segment boundary locations in the other segmentation.
Chapter 7

Accent Prediction Experiments on Higher-level Features

As noted in Chapter 5, near-perfect automatic accent assignment is attainable for citation-style speech (Hirschberg, 1993), but for extended, spontaneous discourses, further work is needed to develop computational models for highly natural accentuation. This chapter reports on machine learning experiments that explore the contributions of numerous higher-level linguistic features to accent prediction, including many that have not been previously tested in empirical studies on accenting. The experiments test how well the linguistic features investigated in the narrative study, namely grammatical function, form of referring expression and attentional status, predict accentuation in the Boston Directions Corpus. They also show how much predictive power can be gained by combining these lexico-syntactic, lexico-semantic and discourse features with key lower-level features examined in the baseline machine learning experiments reported in Chapter 5.

For these experiments on higher-level features, accent prediction is redefined as the prediction of patterns of deviation from citation form accentuation. Crucially, these deviations are captured at the noun phrase constituent level, which is the natural level at which to represent higher-level linguistic features such as form of referring expression, grammatical function and information status. This task redefinition has two advantages: (1) it bootstraps on the underlying knowledge about citation form or so-called context-independent prosody that is presently embodied in speech synthesis technology; and (2) the abstraction from word accentuation to constituent accentuation allows for the integration of lexico-syntactic, lexico-semantic and discourse features into the prediction methods.

The machine learning experiments again use Cohen’s rule induction system, RIPPER, which was described in Chapter 5. The classification schemes and predictors used in the higher-level experiments are defined in Sections 7.1 and 7.2 respectively. Section 7.3 presents results from several sets of experiments testing the effects of individual features and groups of features on accent prediction for NP constituents. These experiments establish the usefulness of viewing the task of accent prediction as the modeling of divergences from citation form accentuation. By taking this approach, it is possible to factor out the principles governing citation form accentuation and focus directly on uncovering the contributions of higher-level linguistic features, including features of the discourse context, to accenting decisions. Together with the baseline experimental results, the results of these new experiments reveal some productive features and principles to incorporate into accenting algorithms for extended, spontaneous discourses. The results reported in this chapter also shed light on difficult cases of accentuation that cannot be reasonably predicted using the tested features and that must be further examined for their salient linguistic and communicative properties.
7.1 Classification of Constituent Accentuation

The accentuation of baseNPs was coded according to the relationship of the actual accenting (i.e. accented versus unaccented) on each word in the baseNP to the accenting predicted by a TTS system that receives each sentence in the corpus in isolation. For these experiments, word accenting predictions of accented versus unaccented were provided by the Bell Laboratories NewTTS system (Sproat and Olive, 1997), which was discussed in Chapter 5. For each baseNP, one of four accenting patterns was assigned. If the actual and TTS-assigned accentuation of each word matched, then the baseNP was assigned to the CITATION FORM accent class. If one or more accented words in the baseNP were unaccented by TTS, and no words that were actually unaccented were assigned accent by TTS, then the baseNP was said to be in the SUPRA accent class. If the converse applied — one or more unaccented words in the baseNP were assigned accent by TTS, and no words that were actually accented were assigned no accent by TTS, then the baseNP was said to be in the REDUCED accent class. Finally, if there were mismatches between the TTS-assigned accenting and the actual accenting in both directions, then the baseNP was assigned to the SHIFT accent class. The four classes are summarized as follows:

- **CITATION FORM**: exact match between actual and TTS-assigned word accenting.
- **SUPRA**: the actual accentuation includes at least one additional accented word than the TTS prediction, and no accented words are predicted unaccented by TTS.
- **REDUCED**: the actual accentuation includes at least one unaccented word that is accented in the TTS prediction, and no unaccented words are predicted accented by TTS.
- **SHIFT**: At least one accented word is predicted unaccented by TTS, and at least one unaccented word is predicted accented by TTS.

Table 7.1 illustrates this accent classification system with several examples from the Boston Directions subcorpus.

<table>
<thead>
<tr>
<th>Accent Class</th>
<th>TTS-assigned Accenting</th>
<th>Actual Accenting</th>
</tr>
</thead>
<tbody>
<tr>
<td>citation</td>
<td><em>a LITTLE SHOPPING AREA</em> the SUBWAY we</td>
<td><em>a LITTLE SHOPPING AREA</em> the SUBWAY we</td>
</tr>
<tr>
<td>supra</td>
<td><em>one</em> the HARVARD SQUARE exit <em>a PRETTY nice AMBIANCE</em></td>
<td><em>ONE</em> the HARVARD SQUARE EXIT <em>a PRETTY NICE AMBIANCE</em></td>
</tr>
<tr>
<td>reduced</td>
<td><em>the GREEN LINE SUBWAY</em> yet ANOTHER RIGHT TURN CHARLES M G H</td>
<td><em>the GREEN Line SUBWAY</em> yet ANOTHER RIGHT TURN CHARLES M G H</td>
</tr>
<tr>
<td>shift</td>
<td><em>a VERY FAST FIVE MINUTE lunch</em></td>
<td><em>a VERY FAST FIVE minute LUNCH</em></td>
</tr>
</tbody>
</table>

Table 7.1: Examples of Accent Classes. Accented items appear in small capitalization.

BaseNPs assigned to the citation form accent class cannot be characterized as simply accented. Certain noun phrases, such as *we* or *one*, are predicted by TTS to be unaccented in citation form. If these words are similarly unaccented in the corpus, they are assigned to the citation form accent class. Citation form
noun phrases are neither more or less intonationally prominent than they would be if they occurred in isolated utterances in citation-style speech. Using citation form accentuation as a reference, the supra class contains constituents that are made more prominent than in normal citation-style speech, and the reduced class contains constituents that are less prominent. Constituents in the shift class contain baseNPs whose accent pattern has significantly changed from citation form accentuation. This accent classification system thus generalizes the notion of prominence and nonprominence to the constituent level, where changes in prominence are marked relative to citation form accentuation.

While this exact system of classification is original to this machine learning study, it accords with certain phonological categories of accenting phenomena. For example, citation form accenting has been studied as part of “default”, “context-independent” or “neutral” intonation. Reduced prominence is related to the phenomenon of “deaccenting”. Supra prominence is associated with the accentual marking of emphasis and intonational focus. And finally, certain cases of accent shift have been explained as arising from the interaction of prosodic well-formedness principles, such as the rhythm rule, with deaccenting (Liberman and Prince, 1977; Ladd, 1979b; Horne, 1991).

Finally, it should be noted that dedicated NP accent prediction systems have been previously explored, especially in work on complex nominal accenting (Liberman and Sproat, 1992; Sproat, 1994). However, the focus in earlier work was on producing citation form prosody. As discussed in Chapter 5, the NewTTS system used in the machine learning experiments incorporates complex nominal accenting rules (Sproat, 1994) as well as general, word-based accenting rules (Hirschberg, 1993). Thus, in the constituent-based machine learning experiments, both sources of accenting rules are being taken into account in classifying the accenting patterns of baseNPs relative to TTS accent predictions for sentences in isolation. When NewTTS processes sentences in isolation, it is reasonable to assume, based on an understanding of the two mentioned subsystems involved, that it generally assigns citation-style accentuation to the sentences.

7.2 Predictors

The higher-level experiments incorporate as predictors all of the features examined in the narrative study, while broadening the range of values considered for each feature. The experiments also relate the features from the narrative study to the lower-level features investigated in the baseline machine-learning experiments, by incorporating the Lex model tested in Chapter 5 into the class of predictors and by extensively testing surface order information for both text and speech units. Finally, certain intermediate-level features, representing syntactic constituency information, are considered in accent prediction experiments for the first time. Below, the features derived from the Lex model, surface position features and syntactic constituency features are described first. Next, the features from the narrative study, grammatical function, form of referring expression and attentional status, are described.

7.2.1 Features from the Lex Model

The experiments reported in Chapter 5 demonstrated the usefulness of POS and word lemma information in word-based accent prediction. The basic model incorporating these features, the Lex model, included broad class part-of-speech tags and word lemma information. The Lex model was adapted for the new, constituent-based machine learning experiments by defining two set-valued features, BROAD CLASS SEQUENCE and LEMMA SEQUENCE. For each baseNP, a broad class sequence or lemma sequence feature value consists of an ordered list of the broad class part-of-speech tags or word lemmas for the words making up the baseNP. For example, the lemma sequence for the NP, the Harvard Square T stop, is {the, Harvard, Square, T, stop}. The corresponding broad class sequence is {det, n, n, n, n}. RIPPER is able to handle these set-valued
features in combination with other single-valued features, so no further transformation of the Lex model features for constituent-based accent prediction is necessary.

7.2.2 Phrasal, Sentence and Segment Position

Most of the positional features examined in the baseline experiments are considered in the higher-level experiments. However, distance in the new experiments is measured in terms of baseNPs. That is, three values, $n$, $m$, and $p$, represent each position feature, where $n$ equals the number of baseNPs from the beginning of a unit to the current baseNP, $m$ equals the number of baseNPs from the end of the unit to the current baseNP and $p$ equals the total number of baseNPs in the given unit.

The units that are considered are: intermediate phrase, intonational phrase, sentence and discourse segment. Discourse segment information was not included in the baseline machine learning experiments, which used task position instead.

7.2.3 Syntactic Constituency

BaseNP type

The BASENP TYPE feature represents local syntactic constituency information. BaseNPs that are not dominated by any NP node are labeled as SIMPLE-BASENPS. BaseNPs that occur in complex NPs (and are thus dominated by at least one NP node) are labeled according to whether the baseNP contains the head word for the dominating NP. Those that are dominated by only one NP node and contain the head word for the dominating NP are HEAD-BASENPS; all other NPs in a complex NP are CHILD-BASENPS. Thus, there is no distinction made between baseNPs that are sisters of a head-baseNP and other baseNPs that are more deeply embedded in the complex NP and may even contain a head word for a NP embedded in the complex NP, as in the example below.

\[
[[\text{the BUS}]_{\text{head-baseNP}} \text{ which is taking } [[\text{US}]_{\text{child-baseNP}} \text{ to } [[\text{the T stop}]_{\text{child-baseNP}} \text{ at } \text{Logan AIRPORT}]]_{\text{complex-NP}} ]_{\text{SBAR}} ]_{\text{complex-NP}}
\]

The baseNPs, \text{us} and \text{the T stop}, which are embedded in a relative clause, are both child-baseNPs even though within the relative clause itself, neither is a sister of a head-baseNP.

Conjoined noun phrases involve additional categories of baseNPs. BaseNPs in conjoined NPs that are dominated by only one NP node are labeled CONJUNCT-BASENPS. Those that are dominated by more than one NP node occur in child-baseNPs and are labeled CONJUNCT-IN-CHILD-BASENPS, as in:

\[
[[\text{the CORNER}]_{\text{head-baseNP}} \text{ of } [[\text{BEACON}]_{\text{conjoint-in-child-baseNP}} \text{ and } \text{MASSACHUSETTS Avenue}]_{\text{conjoint-in-child-baseNP}} ]_{\text{conjoined-NP}} ]_{\text{complex-NP}}
\]

Finally, modifiers of conjunct-baseNPs are labeled CHILD-OF-CONJUNCT-BASENPS, as in:

\[
[[\text{DUCK}]_{\text{conjoint-baseNP}} [[\text{SEAFOOD}]_{\text{conjoint-baseNP}} \text{ and } [[\text{that SORT}]_{\text{conjoint-baseNP}} \text{ of } [[\text{THING}]_{\text{child-of-conjunct-baseNP}} ]_{\text{complex-NP}} ]_{\text{conjoined-NP}}]
\]

The complete classification system for baseNP type is presented in Table 7.2.

Containing clause type

The CLAUSE TYPE feature represents global syntactic constituency information. Four types of clauses are coded: MATRIX, SUBORDINATE, PREDICATE COMPLEMENT and RELATIVE. Each baseNP is assigned the clause type of the lowest level clause or nearest dominating clausal node in the parse tree, which contains the baseNP.
### BaseNP type

<table>
<thead>
<tr>
<th>BaseNP type</th>
<th>Class description</th>
</tr>
</thead>
<tbody>
<tr>
<td>simple-baseNP</td>
<td>baseNP not dominated by any higher NP node</td>
</tr>
<tr>
<td>head-baseNP</td>
<td>baseNP in complex NP, contains head word for complex NP</td>
</tr>
<tr>
<td>child-baseNP</td>
<td>baseNP in complex NP, does not contain head word for complex NP</td>
</tr>
<tr>
<td>conjunct-baseNP</td>
<td>baseNP in conjoined NP containing head word for conjunct, no NP node dominates the conjoined NP</td>
</tr>
<tr>
<td>child-of-conjunct-baseNP</td>
<td>baseNP in conjunct, does not contain head word of conjunct</td>
</tr>
<tr>
<td>conjunct-in-child-baseNP</td>
<td>baseNP in conjoined NP that is a child of a head-baseNP</td>
</tr>
</tbody>
</table>

Table 7.2: Classification system for BASENP TYPE.

#### 7.2.4 Lexico-syntactic and Lexico-semantic Features

**Grammatical function**

The **GRAMMATICAL FUNCTION** feature values, described in Chapter 6, are SUBJECT (SBJ), DIRECT OBJECT (DO), INDIRECT OBJECT (IO), PREDICATE COMPLEMENT (PC) and ADJUNCT (ADJ). The last three categories of grammatical function were omitted from analysis in the narrative study because of low numbers of occurrences.

**Form of expression**

The **FORM OF EXPRESSION** feature values are analyzed as described in Chapter 6. A much wider of variety of forms than were analyzed in the narrative study is included. The fifteen form of expression categories, whose definitions were presented in Table 6.1, are: ADV, CARD, DEF, DEM, DEMPRO, EXPL, GEN, INDEF, NAME, POSS, PRO1, PRO2, PRO3, QUANT, VN. No grouping of forms into explicit forms or reduced forms is defined, since in theory, such groupings, if useful, should be learned by RIPPER.

#### 7.2.5 Global Focusing

The global focusing status of baseNPs is computed using two sets of analyses: the discourse segmentations and coreference coding. Consensus segmentations, as described in Chapter 6, provide a linear segmentation, where each segment is said to begin at a phrase with a consensus segment-initial label. Given a consensus linear segmentation, the global focusing model is used to determine global focusing status. For each baseNP, if it does not occur in a referential chain, and thus is realized only once in the discourse, it is assigned the SINGLE-MENTION focusing status. The remaining statuses apply to baseNPs that do occur in referential chains. If a baseNP in a chain is not previously mentioned in the discourse, it is assigned the FIRST-MENTION status. If its most recent coreferring expression is in the current segment, the baseNP is in IMMEDIATE focus. If its most recent coreferring expression is in the immediately previous segment, the
baseNP is in NEIGHBORING focus.\footnote{The term, neighboring focus space, introduced in Chapter 3, refers to the last focus space which was pushed on the focus stack before the current focus space. For any segment, this corresponds, in linear or hierarchical segmentations, to the focus space for the linearly preceding segment.} If its most recent coreferring expression occurs in the discourse but not in either the current or immediately previous segments, then the baseNP is assigned STACK focus. Because of the linear nature of consensus segmentations, it is not possible to simulate hierarchical embeddings of discourse segments in the focusing model.

7.3 Constituent Accenting Experiments

The machine learning experiments on constituent accentuation were conducted on the H1-spon and H3-spon subcorpora described in Section 4.3. Several sets of experiments are reported on various feature sets, organized roughly from lower-level features to higher-level features. As discussed, the accent classification system and predictors are defined at the constituent level. The results of these experiments are thus not comparable to those of the baseline experiments reported in Chapter 5. The new classification system of accentuation relative to citation-form differs in several ways from the classification system of accented versus unaccented applied in word-based accent prediction, as in Chapter 5 of this thesis and in previous work (Hirschberg, 1993). One difference is that the constituent-based classification system does not treat tokens occurring outside of NPs, or more precisely, outside of baseNPs. There may be dependencies between verb and argument accentuation, or preposition and object accentuation, for example. Researchers have only begun to explore these possibilities in an empirical manner (Altenberg, 1987). A second difference is that if a NP is classified as non-citation form, the exact accenting pattern — which word(s) gain or lose prominence relative to citation-form accentuation — must still be determined by additional processing. This may seem to make the task of constituent-based accent prediction easier than word-based prediction. However, from another perspective, if say four out of five words in a NP are produced with citation-form accentuation, a word-based system that predicts the citation-form accentuation for all five words would achieve 80% correct accentuation for the words in the NP. A constituent-based system that classifies the token as citation-form would receive no partial credit for coming close to the right accenting pattern. In this sense, the task of constituent-based prediction of deviations from citation-form accentuation is harder than word-based prediction. The most critical difference between the two systems, however, is that the citation-form class of NPs includes both regularly deaccented tokens such as pronouns, and frequently accented tokens such as proper names. Computational learning research has shown that learning descriptions of classes described by disjunctive properties is considerably more difficult than learning descriptions of classes described by conjunctive properties, in both theory and practice. The RIPPER system, like C4.5 and CART, does not learn disjunctive rules.

7.3.1 Corpus Measures

As noted, these constituent-based experiments consider a subset of the words in the corpus, namely those occurring in noun phrases. Table 7.3 provides data on the number of baseNPs in the corpus and number of words in the baseNPs. Table 7.3 shows that half of the words in the corpus occur in baseNPs. Table 7.4 gives a breakdown on baseNPs by accent class. For H1-spon, 75.8% of baseNPs exhibit citation form accenting. For H3-spon, the number is 60.2%. The citation form accent percentages serve as the baselines for the higher-level feature experiments. That is, any gain in baseNP accent prediction in accuracy above 75.8% correct for H1-spon and above 60.2% correct for H3-spon represents an improvement over citation form accentuation models.
Table 7.3: Counts for baseNPs and words in baseNPs.

<table>
<thead>
<tr>
<th></th>
<th>H1-spon</th>
<th>H3-spon</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>total words</td>
<td>2359</td>
<td>1616</td>
<td>3975</td>
</tr>
<tr>
<td>baseNPs</td>
<td>621</td>
<td>410</td>
<td>1031</td>
</tr>
<tr>
<td>words in baseNPs</td>
<td>1203</td>
<td>817</td>
<td>2020</td>
</tr>
<tr>
<td>percent of words in baseNPs</td>
<td>51.0%</td>
<td>50.6%</td>
<td>50.8%</td>
</tr>
</tbody>
</table>

Table 7.4: Distribution of tokens in accent classes for all baseNPs.

<table>
<thead>
<tr>
<th>Accent class</th>
<th>H1-spon all baseNPs</th>
<th>H3-spon all baseNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>%</td>
</tr>
<tr>
<td>citation form</td>
<td>471</td>
<td>75.8%</td>
</tr>
<tr>
<td>supra</td>
<td>73</td>
<td>11.8%</td>
</tr>
<tr>
<td>reduced</td>
<td>68</td>
<td>11.9%</td>
</tr>
<tr>
<td>shift</td>
<td>9</td>
<td>1.4%</td>
</tr>
<tr>
<td>TOTAL</td>
<td>621</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 7.5: Distribution of tokens in accent classes for simple-baseNPs.

<table>
<thead>
<tr>
<th>baseNP type</th>
<th>H1-spon</th>
<th>H3-spon</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N</td>
<td>%</td>
</tr>
<tr>
<td>simple</td>
<td>447</td>
<td>72.0%</td>
</tr>
<tr>
<td>head</td>
<td>61</td>
<td>9.8%</td>
</tr>
<tr>
<td>child</td>
<td>74</td>
<td>11.9%</td>
</tr>
<tr>
<td>conjunct</td>
<td>35</td>
<td>5.6%</td>
</tr>
<tr>
<td>child-in-conjunct</td>
<td>4</td>
<td>0.6%</td>
</tr>
<tr>
<td>conjunct-in-child</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>TOTAL</td>
<td>621</td>
<td>100%</td>
</tr>
</tbody>
</table>

In addition to running experiments on data for all baseNPs, a subclass of baseNPs, namely the non-recursive NPs or simple-baseNPs, was tested separately. There are few empirical findings on accenting principles for lengthy, complex NPs, such as those containing relative clauses or participial phrase modifiers. Testing simple-baseNPs separately provides insight as to whether their accenting principles differ from those of baseNPs that occur in complex noun phrases. Table 7.5 shows the distributional data for baseNP types in the corpus, and Table 7.6 shows the accent class distribution for simple-baseNPs. The baseline percent correct figures for simple-baseNPs are 74.7% for H1-spon and 59.6% for H3-spon.

As in the word-based experiments reported in Chapter 5, all constituent-based accent prediction experiments were run with RIPPER using 10-fold cross-validation, with two rule optimization passes in noisy data mode. Results from experiments are reported in terms of the average percent correct on the test sets from all of the cross-validation runs, along with the standard deviation for this average.
### 7.3.2 Lex Model Features

First, the features derived from the Lex model were tested. Table 7.7 shows that word lemma sequence information performs as well alone as the combination of broad class POS sequence and word lemma sequence information. Relative to the baseline percentages for citation form, performance gains amount to about 4% and 2.5% for the best average results for H1-spon and H3-spon respectively. Only the improvement of the word lemma sequence model over the baseline for H1-spon, however, is statistically significant at $p < .05$.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>H1-spon</th>
<th>H3-spon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Broad class sequence</td>
<td>78.58% ± 1.30%</td>
<td>59.51% ± 2.72%</td>
</tr>
<tr>
<td>Word lemma sequence</td>
<td>80.05% ± 1.85%</td>
<td>62.93% ± 2.68%</td>
</tr>
<tr>
<td>Combined</td>
<td>78.75% ± 1.40%</td>
<td>62.44% ± 1.81%</td>
</tr>
</tbody>
</table>

Table 7.7: Classification rates for experiments on lexical features.

### 7.3.3 Surface Position Features

Next, the surface position and length features were examined. In the order tested, they were: length of baseNP in words, intermediate and intonational phrasal position and total length, sentence position and total length, and discourse segment position and total length. Finally, all features were tested together. As indicated by “na” in Table 7.8, segmentation data were not available for H3-spon due to circumstances beyond the author’s control. The combined experiment for H3-spon therefore included baseNP length, supra

<table>
<thead>
<tr>
<th>Experiment</th>
<th>H1-spon</th>
<th>H3-spon</th>
</tr>
</thead>
<tbody>
<tr>
<td>BaseNP length</td>
<td>76.80% ± 1.32%</td>
<td>62.20% ± 2.61%</td>
</tr>
<tr>
<td>Phrasal positions</td>
<td>75.52% ± 2.03%</td>
<td>59.51% ± 2.55%</td>
</tr>
<tr>
<td>Sentence position</td>
<td>75.83% ± 2.75%</td>
<td>59.51% ± 2.31%</td>
</tr>
<tr>
<td>Segment position</td>
<td>75.85% ± 1.80%</td>
<td>na ± na</td>
</tr>
<tr>
<td>Combined</td>
<td>74.55% ± 2.85%</td>
<td>59.27% ± 2.63%</td>
</tr>
</tbody>
</table>

Table 7.8: Classification rates for experiments on surface position features.
prosodic phrase position and sentence position features. The results show that the effect of any one feature on baseNP accent prediction is negligible, since the means are well within one standard deviation from the baseline percent correct figures of 75.8% for H1-spon and 60.2% for H3-spon. Finally, no useful interaction of these variables was learned, as demonstrated by the slightly weaker performance for these features in combination.

### 7.3.4 Syntactic Constituency Features

Syntactic constituency features capture more abstract structural information than do surface position features. However, no rules at all could be induced from these features, as indicated by inspection of the learned rule sets. There is no evidence, therefore, that local NP constituency or global clausal constituency information aids in the automatic prediction of accenting patterns.

However, the syntactic features coded on the directions corpus represent only a limited amount of local and global constituency information. Distributional analysis of baseNPs in various clause types (e.g. relative clauses, subordinate clauses, sentential complements) and in various NP positions (e.g. child-baseNP, conjunct-baseNP) is needed to determine the possible effects of syntactic constituency on accenting patterns, and whether such effects arise not in isolation but in combination with other features such as the information status of the individual NP components.

### 7.3.5 Lexico-syntactic and Lexico-semantic Features

Grammatical function and form of expression were two variables shown to affect accenting patterns of referring NPs in the narrative study. RIPPER results show no rules are learned for these features for either H1-spon or H3-spon. For both speech sets, form of expression and grammatical function together achieve about a 2% gain in correct predictions, although these gains are not significant.

Table 7.10 shows the performance figures.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>H1-spon</th>
<th>H3-spon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammatical function</td>
<td>75.83% ± 1.93%</td>
<td>62.68% ± 2.74%</td>
</tr>
<tr>
<td>Form of expression</td>
<td>78.10% ± 1.54%</td>
<td>61.95% ± 1.89%</td>
</tr>
<tr>
<td>Combined</td>
<td>78.09% ± 1.29%</td>
<td>62.68% ± 2.36%</td>
</tr>
</tbody>
</table>

Table 7.10: Classification rates for lexico-syntactic and lexico-semantic feature experiments.
7.3.6 Global Focusing Features

The global focus feature was tested alone and in combination with various features classes. In Table 7.11, the global focus enhanced model tested the following features: global focus, form of expression, grammatical function, broad class sequence, word lemma sequence, baseNP type and clause type. Global focusing information was more useful in combination with other feature classes, in particular the ones tested in the narrative study — grammatical function and form of expression — as well as the new constituent-based lexical and syntactic features. The improvements over the baseline of 2-3%, however, are not statistically significant.

### Table 7.11: Classification rates for global focus experiments.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>H1-spon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global focus</td>
<td>75.85% ± 2.07%</td>
</tr>
<tr>
<td>Global focus, form of expression, grammatical function</td>
<td>78.11% ± 1.28%</td>
</tr>
<tr>
<td>Global focus enhanced</td>
<td>79.22% ± 1.96%</td>
</tr>
</tbody>
</table>

7.3.7 Combinations of Classes of Features

Three combinations of feature classes, shown in Table 7.12, were tested. The first combination tests lexico-syntactic, lexico-semantic and syntactic constituency features. The second tests lexico-syntactic, lexico-semantic and Lex model features. The final tests lexico-syntactic, lexico-semantic, Lex model and syntactic constituency features. Results are shown in Table 7.13. The average classification success rate of 63.17%

### Table 7.12: Features in combination experiments.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combination 1</td>
<td>grammatical fn, form of expression, baseNP type, clause type</td>
</tr>
<tr>
<td>Combination 2</td>
<td>grammatical fn, form of expression, broadseq, lemmaseq</td>
</tr>
<tr>
<td>Combination 3</td>
<td>all features tested above</td>
</tr>
</tbody>
</table>

### Table 7.13: Classification rates for combination experiments.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>H1-spon</th>
<th>H3-spon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combination 1</td>
<td>77.61% ± 1.39%</td>
<td>60.98% ± 2.60%</td>
</tr>
<tr>
<td>Combination 2</td>
<td>78.74% ± 1.48%</td>
<td>63.17% ± 1.90%</td>
</tr>
<tr>
<td>Combination 3</td>
<td>79.06% ± 1.53%</td>
<td>61.95% ± 2.27%</td>
</tr>
</tbody>
</table>

for H3-spon on Combination 2 is the highest obtained for all the experiments for H3-spon. This shows an improvement in prediction of almost 3%. The highest average classification success rate for H1-spon is 79.06% for Combination 3, showing an improvement of 3.3%. Neither of these improvements, however, is significant.
7.3.8 Experiments on simple-baseNPs

The three sets of experiments that showed the best performance gains are reported for simple-baseNPs. These are: word lemma sequence information alone, lemma sequence and broad class sequence information together, and Combination 2 (grammatical function, form of expression, broad class sequence, word lemma sequence). Classification rates are provided in Table 7.14. For H3-spon, the lemma sequence model delivers the highest classification rate, 65.71%, of all H3-spon models. The improvement over the simple-baseNP baseline of 56.9% is just barely statistically significant at \( p < 0.05 \). Nevertheless, this represents an increase in classification accuracy of 6%, for an error rate reduction of 15%. The average classification success rate of 80.93% for H1-spon on Combination 2 represents a more significant improvement relative to the baseline classification rate for H1-spon simple-baseNPs of 74.7%. Learning on Combination 2 features increases the classification success rate for H1-spon from the baseline by 6.23%. This represents a 25% reduction of the baseline error rate. Using the Student’s curve approximation at \( p < 0.01 \), the lower bound on performance for the Combination 2 model on H1-spon equals 76.5%, which is greater than the baseline rate of 74.7%. Thus, at \( p < 0.01 \), the learned system for Combination 2 on simple-baseNPs for H1-spon performs significantly better than the citation form baseline.

![Table 7.14: Classification rates for simple-baseNP experiments.](image)

In the rule sets learned by RIPPER for the H1-spon Combination 2 experiments, interactions of the different features in specific rules can be observed. Two rule sets that performed with error rates of 13.6% and 13.7% on the test data from different partitions are presented in Figure 7.1. In the rules themselves, the tilde character is a two-place operator, \( X \sim Y \), signifying that \( Y \) is a member of the set-value for feature \( X \). Inspection of the rule sets reveals that there are few non-lexical rules learned. The exception seems to be the rule that adverbial noun phrases belong to the supra accent class. The rule set for run 4 in Figure 7.2, has the highest error rate for the cross-validation runs at 29.5%, although it is extremely similar to the other two rule sets. However, it fails to make use of form of expression information to learn the noun phrase adverbial rule (i.e. \( \text{supra} \sim\text{form=}\text{adv} \)) and instead makes the generalization that any baseNP containing an adverbial gets supra accent (i.e. \( \text{supra} \sim\text{broadseq} \sim\text{adv} \)). In this case, the less specific rule is less correct, showing that the abstraction of POS class information that the form of expression feature provides can sometimes provide crucial and new information.

Finally, it should be noted that all reported experiments were run with partition sizes of 2, 4, 6, 8 and 10 for cross-validated training and testing. Results for different partition sizes were not significantly different, so the most reliable results from \( k \)-fold cross-validation experiments, which are those for the highest \( k \) of 10, are reported in this thesis.

7.4 Assessment

There are several lessons to be learned from these experiments. First, it is possible to improve on the performance of citation form accenting algorithms, using a number of features that are necessary to achieve reasonable citation form accentuation using word-based training as well as higher-level features identified...
Combination 2 rule set for H1-spon: run 9

reduced : - form=name, broadseq ~ det, lemmaseq ~ Harvard (6/2).
supra : - broadseq ~ adv (20/3).
supra : - gf=adj, lemmaseq ~ this (4/0).
supra : - gf=adj, lemmaseq ~ Cowperwaithe (3/0).
supra : - lemmaseq ~ I (3/1).
default citation (295/66).
Error rate on holdout data is 13.6

Combination 2 rule set for H1-spon: run 10

reduced : - broadseq ~ n, lemmaseq ~ the, lemmaseq ~ Square (6/2).
supra : - form=adv (14/2).
supra : - gf=adj, lemmaseq ~ Cowperwaithe (3/0).
supra : - lemmaseq ~ this (4/0).
supra : - lemmaseq ~ I (3/1).
default citation (293/68).
Error rate on holdout data is 13.7

<table>
<thead>
<tr>
<th>Code</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>broadseq</td>
<td>sequence of broad class POS tags for words in NP (adv=adverb, det=determiner, n=noun)</td>
</tr>
<tr>
<td>form</td>
<td>form of referring expression (adv=adverbial noun, name=proper name)</td>
</tr>
<tr>
<td>gf</td>
<td>grammatical function (adj=adjunct)</td>
</tr>
<tr>
<td>lemmaseq</td>
<td>sequence of word lemmas for words in NP</td>
</tr>
</tbody>
</table>

Figure 7.1: Highest performing learned rule sets for H1-spon, Combination 2.

as relevant to determining the discourse focusing function of accent. Second, it is perhaps counterintuitively not easier to predict deviations from citation form accentuation for speakers who exhibit a great deal of non-citation-style accenting behavior, such as speaker H3. In terms of both absolute percentages and error rate reduction, improvements in H1-spon accent prediction over the baseline exceeded those for H3-spon, although about 15% more of H3-spon’s tokens exhibited non-citation form accentuation. It may be the case that H3-spon exhibits more features associated with unplanned, on-line spontaneous speech productions than does H1-spon. There is no measure during a corpus collection of this nature for determining whether a speaker has fully planned his or her message, or whether s/he has only figured out the gist of what s/he wants to say before speaking. Differences in certain speech characteristics between H1-spon and H3-spon, reported in Chapter 4, reflect the more spontaneous nature of H3-spon. The higher rate of disfluencies (.029 repairs/word for H1-spon versus .043 for H3-spon) and much shorter average length of intermediate phrases (4.2 words for H1-spon versus 2.8 words for H3-spon) and intonational phrases (6.5 words for H1-spon versus 4.4 words for H3-spon) convey the fact that H3’s spontaneous speech was more hesitant and disfluent. In addition, observations made during syntactic analysis revealed a greater number of ungrammaticalities and run-on constructions, such as continuative relative clauses and parentheticals, for H3-spon compared to H1-spon. These textual aspects of H3-spon’s speaking style also reflect the more unplanned nature of
Combination 2 rule set for H1-spon: run 4

reduced :- lemmaseq ∼ the, form=name, lemmaseq ∼ Harvard (6/2).
supra :- broadseq ∼ adv (18/2).
supra :- lemmaseq ∼ this (4/0).
supra :- lemmaseq ∼ I (4/1).
default citation (303/63).
Error rate on holdout data is 29.5.

<table>
<thead>
<tr>
<th>Code</th>
<th>Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>broadseq</td>
<td>sequence of broad class POS tags for words in NP (adv=adverb)</td>
</tr>
<tr>
<td>form</td>
<td>form of referring expression (name=proper name)</td>
</tr>
<tr>
<td>lemmaseq</td>
<td>sequence of word lemmas for words in NP</td>
</tr>
</tbody>
</table>

Figure 7.2: Lowest performing learned rule set for H1-spon, Combination 2.

H3’s spontaneous speech. It may also be the case that underlying differences in verbal facilities and performance skills were responsible for some of these differences in the two speakers’ productions. These differences are familiar aspects of studies on individual speaker differences, yet remain hard to quantify and measure in precise terms. As for the third lesson, it is evident from the nature of the rules learned in the constituent-based experiments that rules that express deviations from citation-form accentuation may often be lexicalized rules of some sort. For example, the lemma sequence model delivered the best experimental result for H3-spon, giving an error rate reduction of 15% on simple-baseNPs. Lexicalized linguistic rules have proven to be useful in training part-of-speech tagging using Brill’s transformation-based tagger (1994). Lexicalized rules arise in other areas of language processing, such as lexicalized syntax grammars and more abstractly in head-driven grammars. The specific lexicalized rules learned for reduced and supra accentuation classes would not have followed from any theoretical or empirical proposals in the literature. As for part-of-speech tagging, it may be that domain dependent training using automatic learning is the appropriate way to develop practical, accurate models of accenting patterns in different corpora. And especially for different speakers in the same domain, automatic learning methods seem to be the only efficient way to capture perhaps idiolectical variation in accenting and other linguistic features. Finally, it should be noted that two of the features demonstrated to be related to accentuation in the narrative study, grammatical function and form of referring expression, do contribute information useful for significantly improving upon baseline citation-form accentuation, as demonstrated by the Combination 2 experiments, which reduced the error rate for H1-spon simple-baseNPs by 25%.
Chapter 8

Conclusion

This thesis explored the role of prominence in language processing from two perspectives. First, the functional prosody perspective narrowed the scope of the thesis to an examination of the discourse focusing nature of prominence, in relation to other functions of prominence as well as to other grammatical aspects of language. Second, the computational perspective led to the adoption of a computational discourse modeling framework in which a new account of prominence interpretation in discourse was developed. Original theoretical proposals concerning the discourse focusing nature of prominence and its principled interactions with lexical, semantic, syntactic and other linguistic factors were investigated in distributional analysis of spontaneous narrative speech. The investigation of the spontaneous narrative monologue revealed that accent information could be tightly integrated with other factors involved in both local and global attentional modeling, in particular form of expression, grammatical function and hierarchical discourse structure. The principles uncovered in the study were integrated into discourse processing algorithms assigning prominence a central role in attentional modeling and reference resolution. Next, to explore the usefulness of these results for speech generation, the interactions of prominence and numerous other linguistic factors known to influence accent were tested in machine learning experiments on a multi-speaker corpus of spontaneous direction-giving monologues. The results of the machine learning experiments begin to address the extent to which the principles and factors influencing prominence can be incorporated directly into accent prediction systems that can be used in current speech synthesis technology. Word-based and constituent-based machine learning experiments on accent prediction demonstrated that these higher-level features could be combined with a wide range of lower-level linguistic features, such as lexical category and word lemma information, to significantly improve on citation form accent assignment using an automatically trainable system.

This thesis also contributes original findings on the prosodic nature of spontaneous speech and preliminary insights into individual speaker differences. Few results on accenting principles for spontaneous speech have been developed in either the theoretical or empirical literature. Word-based accent prediction experiments on the direction-giving monologue speech, both spontaneous and read, therefore provide data on task-oriented monologue speech to compare with the results of Hirschberg (1993) for accent prediction in spontaneous dialogue speech and professionally read news stories. While providing comparable performance on one speaker for the genre of extended, spontaneous monologue speech, the study also revealed in more detail than previous studies the nature of individual speaker differences that arise in accenting patterns in spontaneous speech.

The studies in this thesis provide a broad and deep analysis of the discourse focusing nature of prominence. By breadth is meant not only the consideration of speech and text dimensions, but also the analysis of multiple levels of structure within prosodic structures and other linguistic structures in text. Text analyses ranged from surface position features to the hierarchical segmental structure of discourse. By depth is meant the rigor with which each coding scheme is defined, using the latest computational linguistic, linguistic and
prosodic theories and practices. The annotated corpora studied in this thesis incorporate a considerable number of rigorous, documented and transferable coding systems, many of which are accompanied by useful tools and most of which have been facilitated by individuals working within enormous collective efforts, such as the Penn Treebank initiative, the ToBI group and the Discourse Resource Initiative (Hirschman et al., 1996), which have focused on developing standard annotation or transcription schemes for text, speech and discourse features respectively. However, it has not been the case, to date, that researchers from these separate communities have come together to develop corpora that are fully and rigorously labelled in all three spheres – lexico-syntactic, acoustic-prosodic and discourse. The enterprise described in this thesis can be viewed as a case study relating one speech feature to structure in each of these spheres. As an initial attempt to see where the connections between these linguistic spheres may lie, it contributes to a better understanding of how to process prominence in a speech understanding architecture integrating speech and text processing. In particular, this thesis contributes a potential theoretical basis not only for integrating prominence into traditionally text-based NLP, but also for linking the interpretation of prominence to models for its generation. While the principles of prominence interpretation may be directly used in message-to-speech synthesis, in which discourse structure and coreference relations are directly encoded, the machine learning experiments show that they may also lead to significant improvements in text-to-speech synthesis systems for restricted domains. In further exploring the prosody-discourse interface, new foundations for integrating this important speech cue into NLP systems will be uncovered.
Appendix A

Boston Directions Corpus Collection Protocols

A.1 Spontaneous Speech Recording

INSTRUCTIONS TO SUBJECTS

In this experiment, you will be asked to describe to a partner how to get to several places in the Boston area, by walking and/or by taking public transportation. You will be given a map of the T and a collection of street maps in case you need to refresh your memory. You may use the felt-tipped pens provided to trace the routes on the map and to jot down notes about your planned route. Feel free to refer to your maps at any time, but please do not rustle them while you are speaking.

Each task that you will be asked to perform is described on a separate page. Carefully read the instructions on the page. You will then have a few minutes to think about the route you want to describe. When you have planned your route, inform your partner that you are ready, and s/he will nod to signal that you may begin speaking.

As you provide directions, your partner will trace the route you describe on a large composite map of the Boston area. You will be able to see your partner, but not the composite map. Your partner will not be allowed to speak to you while you are giving directions; if you are speaking too fast, your partner will raise a hand to tell you to slow down, and will nod when you have slowed down enough.

You should assume that your partner is from out of town and has no knowledge of the Boston area. For each task, you should also assume that your partner has no knowledge of what you have said in the previous tasks.

TASK 1
You are at the Harvard Square T stop. How would you get to the Kendall Square station on the T? (Remember that you may use the T map provided to figure out the route you want to describe.)
TASK 2
You are at the Harvard Square T stop. How would you get to Copley station on the T?

TASK 3
You are at your place of residence in the Boston area. What is your favorite coffee or ice cream shop? How do you get from your residence to this shop?

TASK 4
What is your favorite bookstore or music store? What route would you take to get from this store back to your home?

TASK 5
What route would you take to get from the steps of the entrance to the MIT “Dome” building at 77 Massachusetts Avenue to the New England Aquarium in Boston?

TASK 6
A prospective graduate student is coming to visit Harvard from California. This student is a native Californian who has never been to Boston before. You must tell her how to get from Terminal C at the airport to the statue of John Harvard in front of University Hall in Harvard Yard, where she will be met. Note that she cannot afford taxis, so she will need to take public transportation (buses or the T) and walk; she will have very little baggage.

TASK 7
The prospective graduate student wants to take her host out to dinner during her visit. Describe three of your favorite Harvard Square eating establishments for her to choose from. You may give her some basis for choosing among your suggestions, by talking about price, location, food quality, or ambiance, for example.

TASK 8
The student must visit Widener Library in the morning and the office of your own academic department in the afternoon. She will eat lunch at a place you recommend in the Square. How can she accomplish all these activities, starting and ending at the front of the Science Center?

TASK 9
The student has a free day in Boston and would like to see some sights. Choose two from the list of sights below (the two you are most familiar with will be fine) and give the student directions on how to get to them. Describe a single excursion that starts and ends at the Harvard Square T stop. The sights can be visited in any order you wish. For each sight, if at all possible, please provide the student with directions on how to get to the sight entrance, not just the general vicinity. Also, suggest one or two activities to do at each sight, or describe a few interesting features that your visiting student should take note of.

- Boston Public Garden
- Children’s Museum/Computer Museum
• Faneuil Hall/Quincy Market
• Fenway Park
• Filene’s Basement/Downtown Crossing
• Museum of Fine Arts (MFA)
• Old North Church
• State House
• Symphony Hall

Now you have a few minutes to plan your directions; you may use the felt-tipped pens to trace the route on the maps.
A.2 Read Speech Recording

INSTRUCTIONS TO SUBJECTS

In this experiment, you will be asked to read aloud the directions that you provided for getting around the Boston area. You will be given texts of your directions, which have no punctuation. You may annotate these texts as you like. Please read them through until you comprehend them well. When you are ready to record, let the experimenter know and read the directions in a natural voice.
Appendix B

Expert Segmentation Instructions

Instructions for Annotating Discourses for Global Segmentation and Parentheticals in Directions Discourses

There are two levels of analysis: global and local; below are instructions for labeling on the global level only.

A. General annotation instructions: global discourse structure

1. MINIMAL DISCOURSE SEGMENT UNITS

You have been given a set of files of phrases. These represent minimal segmentation units. Please do NOT break any lines apart, or join any lines together. Of course, you may include as many units as you like in a given segment using the notation described below.

2. DISCOURSE SEGMENTATION

You are asked to notate the files to reflect the linguistic structure of the discourse, where each segment in the linguistic structure achieves a discourse purpose. Because there is no formal definition of discourse purpose this part is somewhat tricky — you will have to use various parts of your knowledge/intelligence to figure out where there is a change in purpose.

Indicate discourse segment boundaries and the relationship between segments as follows:

a. DS/DSP HEADING
Precede each new segment with its segment number, beginning with 0; on the same line, indicate a short description of the discourse segment purpose for this segment, e.g.:
ds0: give recipe for making apple pie
phrase0
phrase1
.
.

b. GROUPING NOTATION

The initiation of a new discourse segment can be notated by placing a new DS/DSP label at the appropriate position in the file. Note that the numbering for DS/DSP labels should follow their linear order of occurrence. (If, however, you need to “squeeze” in a segment, you may use alphabetic tags: e.g. ds2a, ds2b).

Use indentation (indicated by tabbing) to distinguish between sister segments and embedded segments. Place all discourse units in a given segment at the same indentation level as the DS/DSP label for that segment, e.g.:

ds0: give recipe for making apple pie
phrase0
phrase1
    ------ ds1: describe selection and preparation of apples
    ------- phrase2
    ------- phrase3

    ------- ds2: describe preparation of crust
    ------- phrase4
    ------- phrase5
    ------- phrase6
.
.

Note in this example that ds1 and ds2 are daughter segments of ds0; ds1 and ds2 are sister segments. In effect, the segments form a tree structure, with DS0 as the root of the tree.

Note the following segmentation conventions:

(1) The multiple embedding of DSPs is allowed, e.g.:

ds0: give recipe for making apple pie
    ------ ds1: motivate making of apple pie
    ------- phrase0
    ------- phrase1
(2) The closing of segments in most cases is not explicitly notated, but is implicit in the indentation levels. That is, if there is a discourse segment closing and resumption of the immediately embedding discourse segment, there will be no preceding line with a purpose label. Compare the segmentations (A) and (B):

(A) segment start after pop
ds1: purpose 1
phrase1
phrase2
~-------~ds2: purpose 2
~-------~phrase3
~--------~phrase4
ds3: purpose 3
phrase5
phrase6

(B) pop returns to embedding space
ds1: purpose 1
phrase1
phrase2
~-------~ds2: purpose 2
~--------~phrase3
~-------~phrase4
phrase5
phrase6

c. ATTENTIONAL & INTENTIONAL RELATIONS

In the default cases modelled above, the intention or purpose of an embedded segment DS1 is related to that of its embedding segment DS0 by a domination relationship. That is, the achievement of DSP1 partially satisfies DSP0. For example, describing the preparation of the apples partially satisfies the goal of giving a recipe for making an apple pie.

However, in some cases, such as interruptions, digressions, flashbacks, and information presented "out of order", a segment DS4 may be embedded attentionally with respect to DS3, but not intentionally. In the case of flashbacks, the purpose of one segment may be (intentionally) dominated by a
segment purpose other than the segment that immediately dominates it in the tree. In the case of interruptions and digressions, the purpose of the segment is not intentionally dominated by the purpose of any (preceding) segment in the tree.

For example, below, DS4 represents a digression, whose purpose is not dominated by that of any other segment in the discourse; this is indicated notationally by placing ’[‘] at the end of its DSP description. Alternatively, the purpose of DS6, a flashback, is dominated by that of DS2; this relationship is indicated by placing ‘]ds2[‘ at the end of the description of DS6’s purpose, e.g.:

ds0: give recipe for making apple pie
ds1: describe selection and preparation of apples
nds2: describe preparation of crust
nds3: describe baking of pie
ds4: describe early medieval pie baking []
ds5: describe tests for doneness
nds6: describe optional crust decorations [ds2]

B. PARENTHETICALS:

Often subordinating constructions of various sorts (e.g. relative clauses, prepositional phrases) are used to supply background information or additional details. These constructions appear to be intended to work something like mini-subsegments. They seem to serve two somewhat different roles:

(a) to provide additional information about an individual
   [e.g. ‘the chairman, Jack Jones of ABC Corp,’] or
(b) to provide background information about an event
   [e.g. ‘Jack won a gold medal in what some consider the first real marathon ever run. His time was...’]

Here the winning of the medal is the main event; the bit about ‘some consider...’ is a subsegment.

The (a) cases are termed ”sentential” parentheticals; the (b) cases, ”discourse” parentheticals. In the (b) cases, the material is often evaluative or provides substantiating evidence or appeals to an authoritative source.

In your annotation files, please mark these two types of parentheticals, sentential and discourse.
1. enclose sentential parentheticals in single parentheses; main info (parenthetical) continuation.

   e.g. ...the chairman (Jack Jones of ABC Corp.)...

2. enclose discourse parentheticals in double parentheses.

   e.g. [in a segment about exercise and cholesterol]:
   ...since Jill ((in what some of her friends considered a move completely out of character)) started running 10 miles at 6 AM every day her cholesterol has been far below normal...

Note that a parenthetical may consist of only part of a phrase, or multiple phrases.

C. Guidelines for some difficult cases:

1. Self-corrections are cases where the speaker interrupts the fluent flow of speech, and "restarts" or "repairs" the utterance, e.g. "How many--how much does a one-way ticket cost?" When you identify a self-correction, first delimit the 'incorrect' or misspoken material by enclosing it in curly braces. Include this material in whatever segment the repairing or correcting material appears in.

   e.g.

ds1: inquire about denver-boston flights
i'd like information on flights from denver to boston
--how many--how much does a one-way ticket cost?

2. Boundary phrases may be difficult to assign to a segment. Sometimes these phrases can be either the last phrase of an embedding segment or the initial phrase of an embedded segment. In these cases, when the meaning allows either interpretation, make the phrase the first phrase of the embedded segment. For example, compare possible segmentations in (A) and (B). The one in (B) would be preferred for this ambiguous case.

   (A)

dS3: describe baking of pie
.
.
bake the pie until it is done
for about one hour

or until the crust is a golden color

pierce the pie with a knife
to test the firmness of the apples

(B)
d3: describe baking of pie

..

d4: describe tests for doneness

bake the pie until it is done
for about one hour
or until the crust is a golden color
pierce the pie with a knife
to test the firmness of the apples
References


Avesani, Cinzia. 1996. Personal communication.


Beckman, Mary E. and Julia Hirschberg. 1994. *The ToBI Annotation Conventions*. Ohio State University, Columbus, February.


Ostendorf, Mari. 1996. Linking speech recognition and language processing through prosody. Electrical and Computer Engineering Department, Boston University.


Venditti, Jennifer. 1996. Personal communication.


