INTONATION AND DISCOURSE STRUCTURE IN ENGLISH: PHONOLOGICAL AND PHONETIC MARKERS OF LOCAL AND GLOBAL DISCOURSE STRUCTURE

DISSERTATION

Presented in Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy in the Graduate
School of The Ohio State University

By

Rebecca Herman, B.A.

* * * * *

The Ohio State University
1998

Dissertation Committee:

Professor Mary Beckman, Adviser

Professor Keith Johnson

Professor Craige Roberts

Approved by

_____
Adviser
Department of Linguistics

Copyright by

Rebecca Herman

1998

ABSTRACT

This dissertation examines the relationship between sound structure and discourse structure in English. In English, phrases are marked by edge tones which have discourse functions. A high edge tone indicates that the phrase is to be interpreted with respect to a subsequent phrase, while a low edge tone means that the phrase does not form an interpretive unit with the subsequent phrase. How the tones in a phrase marked by a low edge tone are implemented phonetically conveys how complete a sub-topic is. Thus the choice of edge tone conveys local discourse relationships between pairs of phrases while the phonetic implementation conveys global discourse relationships– the grouping of sets of phrases into topics and sub-topics.

These hypotheses about the discourse functions of edge tone categories and phonetic implementation were tested experimentally using pairs of discourses with at least two sub-topics in which the sub-topics occurred in different orders. The expectation is that the phrase at the end of a sub-topic will have the same choice of edge tones in both orders of occurrence, but that the tones in the phrase will be implemented differently in the two cases. The experiment involved short discourses which were recorded, transcribed tonally, played to listeners to see if they could distinguish the utterances by discourse position, and analyzed acoustically.

The results showed that the same edge tones were used in a large proportion of the utterances, verifying the prediction that edge tones mark local discourse relationships. In

the utterances which were identifiable by listeners as to discourse location and which had the same or similar sequences of tones, the phonetic implementations of the tones varied in systematic ways between the two utterances in a pair, thus verifying the hypothesis that phonetic implementation of tones signals global discourse relationships.

In sum, this dissertation explores how phonological and phonetic aspects of the intonational structure function as linguistic cues to local and global information structure in discourses.

ACKNOWLEDGMENTS

I would like to thank my adviser, Mary Beckman, for setting me on an interesting path and for help along the way. Thank you for always reminding me of the forest (and sometimes even an adjacent grove or two) when I was lost among all the trees. I am also very grateful to Keith Johnson for always helping me to get experiments set up and running properly, for inspiring me to do a perception test and a multi-dimensional scaling analysis, for helping with statistics questions and with interpretation of data in general, and for always having good advice for me. A special thanks to Craige Roberts for helping me to clarify the issues, for the insightful comments, and for always having an encouraging word. I would also like to thank Jignesh Patel for being a wonderful systems administrator. His work is much appreciated. Thanks also to all of the undergraduate subjects from Linguistics 201 and Linguistics 371 who participated in my experiments, and to their instructors who let me solicit subjects in their classes. This dissertation would not have been possible without them. A special thanks to Julie McGory for ToBI transribing all of my materials. I couldn't have asked for a better transcriber. I would also like to thank Laurie Maynell for last-minute recording of example sentences.

I wish to thank my parents for encouraging me along every step by step by step of my thousand mile journey. Thanks also to Gid Herman for always looking out for me. And thanks to Jonathan, Liz, Rachel, and Naomi Herman for always being there for me.

I also want to thank Rachna Prasad, who was there week-by-week, listening, advising, sympathizing, sharing, and always, always, always cheering me on. Also, thanks to Molly Homer for the friendship, the support, and the confidence in me. Thanks for always cheering me up when I am down, and for sharing my happiness when things are better. Thanks also to Svetlana Godjevac for many helpful discussions, for explaining semantic and pragmatic issues no matter how often I asked, and for the words of comfort and encouragement. I also want to thank Allison Blodgett, Paul Davis, Mariapaola D'Imperio, Barbara Gili-Fivela, Stefanie Jannedy, Hélène Loevenbruck, Laurie Maynell, Robert Poletto, Jennifer Venditti, and Pauline Welby for many, many long discussions and for the friendship. I would also like to thank my colleagues who make working in the Linguistics Department at Ohio State such a pleasant, fulfilling experience, including Kim Ainsworth-Darnell, Mary Bradshaw, Matt Makashay, Amanda Miller-Ockhuizen, Ben Munson, Frederick Parkinson, Shu-Hui Peng, Liz Strand, Jenny Vannest, Steve Winters, and Kiyoko Yoneyama. Thanks also to Khalil Iskarous, Abdulkadir Hassan Mohammed, Maria McCormick, Sharon and Ken Housinger, Susannah and Ted Levin, and Margaret and David Lewis for friendship over the years. This dissertation would not have been possible without the friendship and support of all of the people listed here, for which I am truly grateful.

Part of the support for work on this dissertation came from an NSF graduate student fellowship research allowance. Work on this dissertation was completed with support from an Ohio State University Presidential Fellowship.

VITA

July 6, 1970..... Born – Bloomington, Indiana

1992..... B.A. Linguistics, University of Illinois, Urbana.

1992-93, 1994-1996..... NSF Graduate Fellow, The Ohio State University

1993-1994, 1996-1997..... Teaching Associate, The Ohio State University

1997-1998..... Presidential Fellow, The Ohio State University

PUBLICATIONS

1. R. Herman, M. Beckman, and K. Honda. Subglottal Pressure and Final Lowering in English. *Proceedings of the Fourth International Conference on Spoken Language Processing*. (1997)

2. R. Herman. Syntactically Governed Accentuation in Balinese. *OSU Working Papers in Linguistics, Papers from the Linguistics Laboratory*. no. 50. Kim Ainsworth-Darnell and Mariapaola D'Imperio, eds. (1997)

3. R. Herman. Final Lowering in Kipare. *Phonology*. 13:2. (1996)

4. R. Herman. Prosodic Structure in SiSwati. *OSU Working Papers in Linguistics, Papers in Phonology*. no. 48. David Dowty, Rebecca Herman, Elizabeth Hume, and Panayiotis Pappas, eds. (1996)

5. R. Herman. Final Lowering in Kipare. *OSU Working Papers in Linguistics, Papers from the Linguistics Laboratory*. no. 45. Stefanie Jannedy, ed. (1995)

6. R. Herman. The Structure of Affricates in Kuvi (A Constraint-Based Account). *Proceedings of the Eleventh Eastern States Conference on Linguistics*. Janet M. Fuller, Ho Han, and David Parkinson, eds. (1994)

7. R. Herman. 'La Double Vie de W' or the status of [w] in Karuk. *Studies in the Linguistic Sciences*. vol. 24, no. 1/2. (1994)

FIELDS OF STUDY

Major Field: Linguistics

TABLE OF CONTENTS

Page

Abstract ii

Acknowledgments iv

Vita vi

List of Figures xi

Chapters:

1. Introduction 1

1.1 Basic Assumptions About Intonation and Discourse Structure 1
1.2 Phonetic Correlates of Discourse Structure 5
1.3 A Phonological Model of Intonation 10
1.4 An Explicit Model of Discourse Structure 18
1.5 The Meaning of Edge Tones and the Discourse Functions of Phonetic Implementations 47
1.6 Two Hypotheses about Phonological Edge Tones, Phonetic Implementation, and Local and Global Discourse Structures 56
1.6.1 Phonological Edge Tones and Local Discourse Structures 56
1.6.2 Phonetic Implementations and Global Discourse Structures 59
1.7 Testing the Hypotheses– Experimental Design 60
1.8 Summary 67

2. Methods 68

2.1 Recording 68
2.1.1 Materials 68
2.1.2 Subjects 74
2.1.3 Methods 76
2.2 Transcription of Intonational Tones 76
2.2.1 Transcription 76
2.2.2 Mapping Tone Similarity Space 77
2.3 Perception Test 81
2.3.1 Materials 81
2.3.2 Subjects 82
2.3.3 Methods 84
2.4 Acoustic Analysis 88
2.5 Summary 93

3. Results 95

3.1 Results– Tone Transcriptions 95
3.1.1 Edge Tones 95
3.1.2 Tone Similarity Judgements 99
3.2 Results– Perception Test 106
3.2.1 Reliability of Listener Judgements 107
3.2.2 Examination of Utterances with High Edge Tones 111
3.2.3 Testing Selected Tokens for Phonological Tone Similarity 112
3.2.4 Reliable Mis-identification 113
3.2.5 Summary 114
3.3 Acoustic Analysis 114
3.4 Summary 143

4. Discussion and Conclusion 146

4.1 Implications of Experimental Results 146
4.2 Directions for Future Research 153
4.2.1 Discourse Structure 153
4.2.2 The Phonology of Intonation 156
4.2.3 The Phonetics of Intonation 157
4.2.4 The Cross-Linguistic Perspective 163
4.3 Conclusion 163

List of References 169

LIST OF FIGURES

Figure……Page
1. Example of structured information vs. unstructured information.....2
2. Phonological structure in intonation– a final fall.....4
3. Phonological structure in intonation– a final rise ……4
4. Intrinsic vowel length and final lengthening ……7
5. Pragmatic differences cued by intonation alone– a declarative sentence. "Pragmatic differences can be marked by intonation."....11
6. Pragmatic differences cued by intonation alone– an echo/incredulous question "Pragmatic differences can be marked by intonation?!" ……11
7. Example with and without a phrase break……16
8. Example of pitch accents, phrase accents, and boundary tones.....17
9. Example of linguistic structure of discourse ……19
10. Example of linguistic structure and discourse segments……20
11. Example of ambiguous discourse segmentation. ……21
12. Example of discourse segments with hierarchical arrangement……22
13. Example of discourse segment purposes and hierarchical arrangement.....24
14. Example linguistic structure of instruction monologue ……25
15. Example of discourse segmentation of instruction monologue ……26
16. Example of discourse segmentation of instruction monologue ……26
17. Discourse segment purposes of instruction monologue ……27
18. Extended example of hierarchy of dominance relations.....29
19. Example list of "mother-of" relations. ……30
20. Example list of "sister-of" relations. ……30
21. Example of satisfaction-precedence relation.....31
22. Example with one dominant DSP and two subordinate sister DSPs ……34

23. Example focus stack, stage 0. ……35
24. Example focus stack, stage 1. ……35
25. Example focus stack, stage 2. ……37
26. Example focus stack, stage 3. ……37
27. Example focus stack, stage 4. ……38
28. Example focus stack, stage 5. ……39
29. Example focus stack, stage 6. ……40
30. Example focus stack, transitive dominance, stage 1. ……41
31. Example focus stack, transitive dominance, stage 2. ……41
32. Example focus stack, transitive dominance, stage 3. ……41
33. Dominance hierarchy with three daughter DSPs.....43
34. Example focus stack, stage N, complete set of daughter discourse segments? ……43
35. Example focus stack, stage N+1, complete set of daughter discourse segments? ……44
36. Example of contrast between H% and L%. ……48
37. Example focus stack, binding of pronoun given H%.……50
38. Example focus stack, binding of pronoun given H% on previous utterance. .50
39. Example of intentional structure with three daughter DSPs ……53
40. Example of two sister DSPs ……57
41. Example of two sister DSPs in both orders. ……58
42. Example of a DSP's effect in two different discourse positions.……60
43. Example logical structure of intentional hierarchy.……61
44. Example logical structures, subordinate DSPs in both orders. ……62
45. Dialogue 1……69
46. Dialogue 2……70
47. Dialogue 3……70
48. Dialogue 4……71
49. Dialogue 5……71
50 Dialogue 6……72
51. Dialogue 7……72
52. Dialogue 8……73
53. Dialogue 9……73
54. Dialogue 10 ……74
55. Subject information from recording session. ……75

56. Subject information from perception experiment. ……83
57. Sample instruction sheet.……86
58. Sample waveform, RMS amplitude, and F0 of the word “dollars” extracted from the sentence “They raised the rent by forty dollars.” ……91
59. Number of edge tones of each type. ……96
60. Expected number of edge tones of each type, no difference by discourse position ……97
61. Expected number of edge tones by each type, pooling together H%.....97
62. Observed number of edge tones by each type, pooling together H% ……98
63. Deviances: (observed - expected)^2 / expected……92
64. Transcriber’s tone similarity judgements for pairs of pitch accents. ……100
65. Transcriber’s tone similarity judgements for pairs of edge tones.……100
66. One-dimensional scaling solution– pitch accent tone similarity space.....101
67. One-dimensional scaling solution– pitch accent tone similarity space, revised. ……101
68. Two-dimensional scaling solution– pitch accent tone similarity space.……102
69. Two-dimensional scaling solution– pitch accent tone similarity space, revised. ……103
70. One-dimensional scaling solution– edge tone similarity space.……104
71. Two-dimensional scaling solution– edge tone similarity space. ……105
72. Number of correct identifications by listener.....108
73. Example RMS amplitude traces– overlaid.……116
74. Example F0 traces– overlaid ……117
75. Example composite figure (labelled) ……118
76. Composite figure of measurements for Dialogue 3, Sentence A.....119
77. Composite figure of measurements for Dialogue 3, Sentence B.....120
78. Composite figure of measurements for Dialogue 4, Sentence A.....121
79. Composite figure of measurements for Dialogue 4, Sentence B.....122
80. Composite figure of measurements for Dialogue 5, Sentence A.....123
81. Composite figure of measurements for Dialogue 5, Sentence B.....124
82. Composite figure of measurements for Dialogue 6, Sentence A.....125
83. Composite figure of measurements for Dialogue 7, Sentence A.....126
84. Composite figure of measurements for Dialogue 8, Sentence A.....127
85. Composite figure of measurements for Dialogue 8, Sentence B.....129
86. Composite figure of measurements for Dialogue 9, Sentence A.....131

87. Composite figure of measurements for Dialogue 9, Sentence B.....132
88. Histogram of RMS amplitude differences in final pitch accented syllable.....133
89. Histogram of F0 differences in final pitch accented syllable.....134
90. Summary table of differences in acoustic measures.....135
91. Composite figure of mis-identified tokens for speaker S3.....137
92. Composite figure of mis-identified tokens for speaker S5.....138
93. Composite figure for mis-identified tokens.....139
94. Example F0 traces– overlaid.....142
95. More example F0 traces– overlaid.....142
96. More example F0 traces– overlaid.....143

CHAPTER 1

INTRODUCTION

1.1 Basic Assumptions About Intonation and Discourse Structure

A hallmark of human language use is that it shows structure at many different levels, such that at each level, a difference in linguistic form can express a difference in linguistic function. For example, the fine-grained structuring of spectral detail can distinguish words from each other such that a contrast in form expresses a contrast in lexical semantic meaning. This level of structure can be exemplified by the word “tall” being distinguished from the word “call” by a difference in the sound structure, with the difference in sound structure expressing a lexical semantic meaning difference.

The structure of language holds at coarser-grained levels as well, such as the structuring of words into meaningful strings. This level of structure can be exemplified by the string of words “Rachel saw Naomi” being distinguished from the string of words “Naomi saw Rachel” by a difference in word order, with the difference in word order expressing a difference in thematic roles of the participants in this case.

In fact, the structure of language extends even further, beyond just the word order in single sentences to an even coarser-grained level of structure which encompasses several sentences together. The examples in Figure 1 illustrate this aspect of linguistic structure.

a)

Sue won't go out with John unless he shaves off his beard and stops drinking.

He still had his beard this morning.

He hasn't bought a razor in three months.

Also, he was at a bar last night with his old college buddies.

His friends had to drive him home at the end of the evening.

vs.

b)

At the end of the evening, John was driven home by his friends.

In the last three months, John has not bought a razor.

Sue won't go out with John unless he shaves off his beard and stops drinking.

John had his beard this morning.

John was at a bar last night with his friends from college.

Figure 1. Example of structured information vs. unstructured information.

The string of sentences in Figure 1(a) sounds more coherent and natural than the string of sentences in Figure 1(b), even though they contain the same information. From the string of sentences in (1a), the reader can easily answer the question: "Does John stand a chance with Sue?" with a resounding "No." On the other hand, this question is not so readily answered from the string of sentences in (1b). The same set of propositions is expressed in (1a) as in (1b), but the information is harder for the reader to extract because it is not structured. What helps the reader to discover the logical structure in (1a) more readily than in (1b) is the use of linguistic cues such as the use of pronouns, the order in which the information is presented, and the use of the connector word "also."

An example such as the one given in Figure 1 illustrates that there is structure at the linguistic level encompassing more than a single sentence. Furthermore, at this level, the differences in linguistic form encompass a wide array of phenomena, ranging from the use of pronouns to the choice of word order in sentences and even to the choice of the order in which the sentences themselves are expressed. Similarly, at this level, the differences in linguistic function which are expressed by differences in linguistic form are more complex than lexical semantic meaning or thematic role, and seem to consist of a structure on the information. For example, which entities being discussed are most salient in the discourse is one aspect of the information structure, as is how one sentence's topic relates to another sentence's topic. In natural language use, various linguistic cues are used to help the interlocutor uncover the information structure which underlies the discourse.

One of the linguistic cues to information structure in discourses, along with pronominalization, word order, and connectives, is intonation, which is itself structured such that there is a phonological level at which there are categorical distinctions and a phonetic level at which the implementation of the categories can differ. The examples in Figure 2 and Figure 3 illustrate a phonological category difference between a final fall (marked by L-L% at its rightmost edge) and a final rise (marked by L-H% at its rightmost edge). The top panel in each figure shows the acoustic waveform, a record of the acoustic energy in the signal over time, with time on the x-axis and amplitude on the y-axis. The middle panel in the figure shows the fundamental frequency (F0), a record of the frequency of vibration of the vocal folds as calculated from the glottal pulses in the waveform, with time on the x-axis and frequency (in Hz) on the y-axis. The F0 is closely related to pitch, especially in the range of speech. The bottom panel shows the tones in the utterance (to be discussed further in section 1.3) and the words in the utterance, with the vertical line to the right of the typed word aligning with the end of the word.

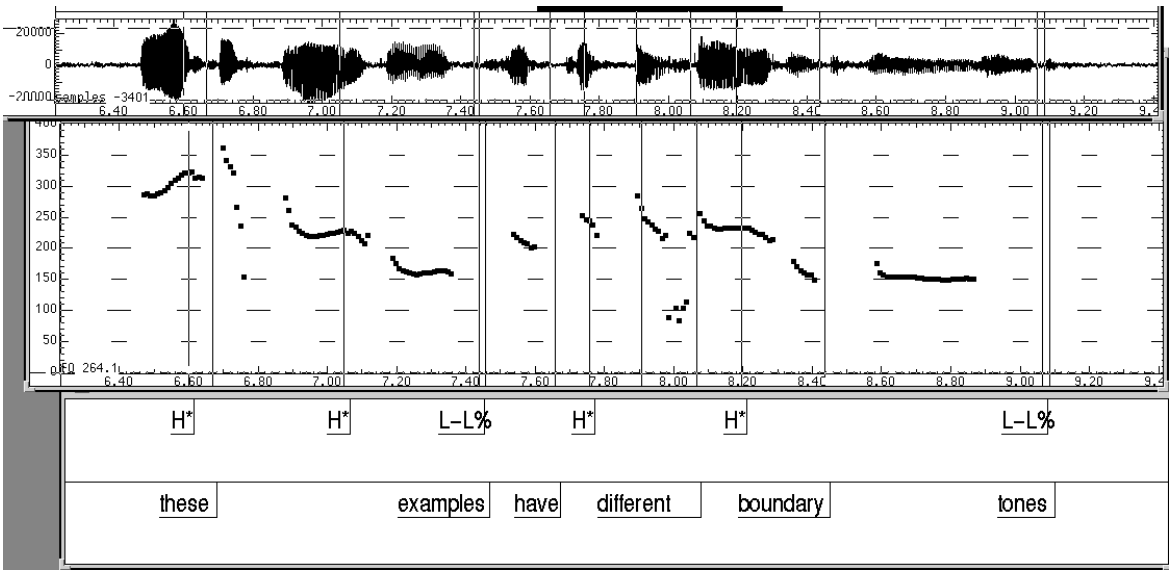


Figure 2. Phonological structure in intonation– a final fall.

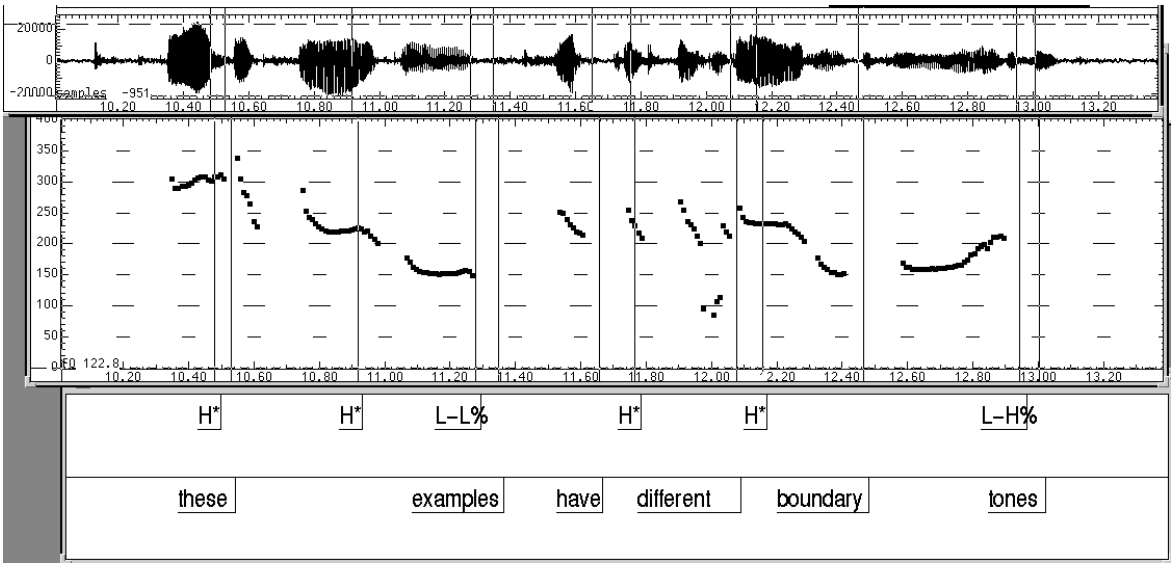


Figure 3. Phonological structure in intonation– a final rise.

Furthermore, exmining the category of final fall, as will be discussed in further detail in Chapter 3 (Results), the phonetic implementations of tones can be different. Thus,

the intonation also has structure, such that there are category differences in addition to differences of implementation within each category.

The question which will be explored in this dissertation is how the different aspects of intonational structure (the phonetics and the phonology) help to convey the information structure in discourses in American English, and furthermore, which aspects of the intonational structure convey which aspects of the information structure.

1.2 Phonetic Correlates of Discourse Structure

It is relatively uncontroversial that at least one of the functions of intonation is to mark structure at the discourse level. Research on intonation in discourse has found phonetic correlates of hierarchical discourse structures such as prosodic paragraphs or discourse segments. This work (as applied to English) is exemplified by Lehiste (1975), Lehiste and Wang (1977), Lehiste (1979), Kreiman (1982), Grosz and Hirschberg (1992), and Hirschberg and Nakatani (1996). The types of experiments conducted in this line of research have two components. The first component is a listening experiment to see if there is consistency across listeners when labelling discourse organization. Sometimes the same materials are manipulated to remove segmental information by low-pass filtering to see if listeners can still consistently identify discourse structure using only intonational cues. The second component of such studies is an acoustic analysis to examine what acoustic factors may have been responsible for the perception of a discourse boundary (such as a paragraph break) at a particular location. Such work has converged on a set of phonetic factors which seem to correlate with discourse structures.

One phonetic factor marking discourse structure is a high fundamental frequency on the first peak in the first sentence in a paragraph (Lehiste, 1975). Another phonetic factor is pause length, such that pauses are longer at paragraph boundaries than at sentence boundaries (Lehiste and Wang, 1977). Low terminal fundamental frequency and laryngealization at the end of discourse units are other phonetic factors (Lehiste and Wang,

1977). Yet another phonetic factor is the amount of pre-boundary lengthening (Lehiste, 1979). Another factor is the presence of a non-level intonational contour (Kreiman, 1982). Pitch-range manipulations as well as preceding and subsequent pause have also been found to be factors (Grosz and Hirschberg, 1992). Discourse-segment-initial phrases have been found to be marked by higher maximum and average fundamental frequency and higher maximum and average RMS amplitude, while segment-final phrases have been found to be marked by lower fundamental frequency maximum and average, lower RMS amplitude maximum and average, and faster speaking rate (Hirschberg and Nakatani, 1996). Such studies have been concerned with the connection between phonetic factors and discourse structure. There are two areas in which this line of work is lacking, though.

The first deficiency in some of the studies on phonetic correlates of discourse structure is in the area of intonational phonology. In order to compare two tokens by making phonetic measurements of the two tokens and seeing how they differ, it is necessary for the phonological categories of the two tokens to be the same. This is analogous to examples from studies of segmental phenomena, in which it is clear that within-category comparisons are necessary when making phonetic comparisons. For example, the amount of final lengthening is relative to the intrinsic duration of a vowel. Low vowels have a longer intrinsic duration than high vowels. The following example illustrates the potential problem with across-category comparisons. In this example, in the sentence “She wore a blue and white hat and a lightweight suit,” the vowel [ˇ] in the word “hat” is .20 seconds long, while the vowel [u] in the word “suit” is .14 seconds long.

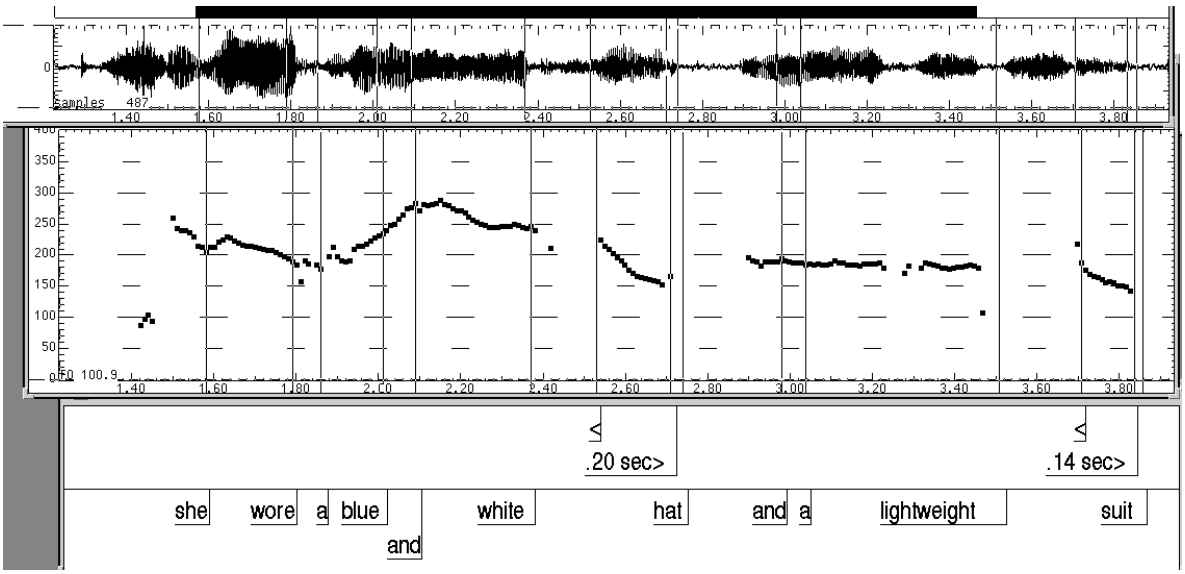


Figure 4. Intrinsic vowel length and final lengthening.

In the example in Figure 4, it would be incorrect procedure to compare the length of the vowel in “hat” and the vowel in “suit” and based on those comparisons, to posit a stronger boundary after the word “hat” than after the word “suit” simply because its low vowel is long, while the word “suit” has a shorter high vowel. Rather, it must be recognized that these two vowels belong to different categories, and that these two categories have different properties. Then, measurements of those properties can be relativized to the category, without confounding the difference in category with the measurement in question. Similar precautions apply in the study of phonetic correlates of discourse structure, in which case it would be erroneous to compare the low point at the beginning of a final rise vs. the final low at the end of a final fall and to posit that “terminal low points are invariant” (see, e.g., Maeda, 1976) without ensuring that the “low points” in question belong to the same phonological category. If such precautions are not taken, then this leads to an “apples and oranges” comparison problem. Such comparison problems have arisen in the studies of the phonetic correlates of discourse structure. See, for

example, the claim by Lehiste and Wang (1977) of low terminal fundamental frequency as a phonetic correlate of a boundary, without a definition of the phonological category of the terminal point. Similarly, Grosz and Hirschberg (1992) and Hirschberg and Nakatani (1996) examined the F0 maximum and the amplitude maximum in a phrase, without controlling for the tone of the syllable whose F0 or amplitude was being measured. As will be shown later in this study, it is important to separate out categorical differences (such as the difference between a final rise and the lack of a final rise) vs. gradient differences (such as the difference between a final fall and a more pronounced fall). As Ladd (1996) points out, a lack of concern with the phonology of intonation in favor of instrumental studies of acoustic correlates of concepts such as "focus" or "contrast" or "giveness" (or, in this case, "finality") is flawed in that this relationship between form and meaning is thus treated differently than the relationship between form and meaning in other aspects of language, in which functional categories mediate between the physical acoustic realization and the meaning. (p. 20-21). Ladd also makes the point that postulating the existence of phonological categories does not imply the limitation of investigation to impressionistic work, but rather provides empirically testable issues (p. 22). For example, postulating the separation of peak tones from edge tones would not limit the study of phonetic correlates of discourse structure to impressions of where peaks are and where edges are, but rather would allow predictions about sound structure to be stated more clearly, and ensure that phonetic measurements would really be comparing tokens which were in the same category, and hence really were comparable. For example, studies which discuss "expanded pitch range" have sometimes taken as a measurement point the highest point in the first peak of the sentence, in which case it is important to recognize that the first peak is a tone which can be categorized as some particular phonological intonational tone. In such a case, it is also important to check whether it is just the first peak which is higher in one case vs. another, or whether all of the peaks are higher in the phrase which has "expanded pitch

range," but this again requires the recognition of the peaks whose height is being measured as phonological tone targets. In this way, the location and type of phonological tone present can be separated out from the definition of the pitch range. Furthermore, such a pursuit would be able to separate out manipulations of the pitch range at the first peaks from compression of the pitch range in the last phrase (known as "final lowering"), again, by recognition of the phonological differences between the tone targets at different locations. Rather than a definition of pitch range blending together both initial raising and final lowering, these two components can be separated out and examined individually. Thus, even studies of the phonetic correlates of discourse structure need to take into account the phonological categories of intonation, if not by controlling for which categories are present then at least by comparing only tokens which are of the same phonological category.

The current study will avoid this problem by having the tones in an utterance transcribed according to an explicit intonational model and then comparing measurements from within the same tone category (at the same time, proposing an original methodology for defining "tone categories"). Thus, this work will avoid the "apples and oranges" comparison problem by explicitly comparing only elements which belong to the same category and seeing how they differ in their implementation.

The second area in which some of the studies of the phonetic correlates of discourse structure are lacking is in the lack of explicit theories of discourse structure. The earlier studies tend to adopt the distinction between "sentences" and "paragraphs" as used in written language, but without defining for speech what the discourse units are and what relationships hold among them. This is not true of the work of Grosz and Hirschberg (1992) or Hirschberg and Nakatani (1996), for example, who use the notion of "discourse segment" from Grosz and Sidner (1986). However, Grosz and Hirschberg (1992) and Hirschberg and Nakatani (1996) use the discourse segmentation only insofar as they

identify utterances as being at the beginning, middle or end of a discourse segment, without elaborating further on the relationships holding between discourse segments.

The current study will avoid this problem by using an explicit model of discourse structure and by controlling for the information structure (including the hierarchical relationships among discourse segments) in the discourses which are used in the experimental corpus. Thus, this work will be able to explicitly define different components of discourse structure when discussing how this structure is marked by linguistic cues.

In order to help resolve these two problems in the present study and to better understand which aspects of discourse structure are conveyed by which aspects of the sound structure, an explicit model of the phonological structure of the intonation of English will be described as well as an explicit model of the components of discourse structure. An understanding of these models will allow for testable hypotheses to be made about the functions of phonetic and phonological markers of discourse structure. Framing such testable hypotheses will allow for a study of the phonetic correlates of discourse structure without the confounds of phonological differences or discourse-structural differences, which were problems in earlier works in this area.

1.3 A Phonological Model of Intonation

First, in discussing the model of intonation it must be stated that it is assumed here that intonation is a component of grammar and is not just a way to convey emotions or feelings. Evidence that intonation is a linguistic phenomenon and hence must be modelled as a part of the grammar comes from the fact that the form-function mapping between intonational tones and their pragmatic functions is arbitrary. Even simple cases in which pragmatic functions can be informally classified according to “sentence type” vary from language to language. For example, it is relatively uncontroversial that echo or incredulous questions in English serve a different pragmatic function than declaratives, and that such utterances can be distinguished by the intonation alone, without using differences in lexical

choice, as seen in Figures 5 and 6. In this example, the F0 of the declarative sentence differs markedly from the F0 of the echo/incredulous question, as is apparent pre-theoretically simply from comparison of the F0 traces.

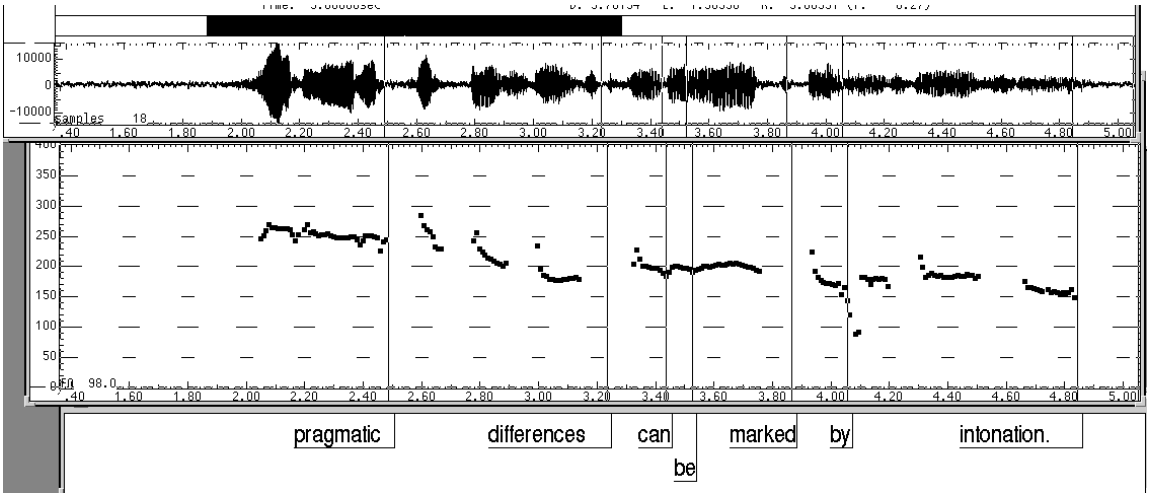


Figure 5. Pragmatic differences cued by intonation alone– a declarative sentence. "Pragmatic differences can be marked by intonation."

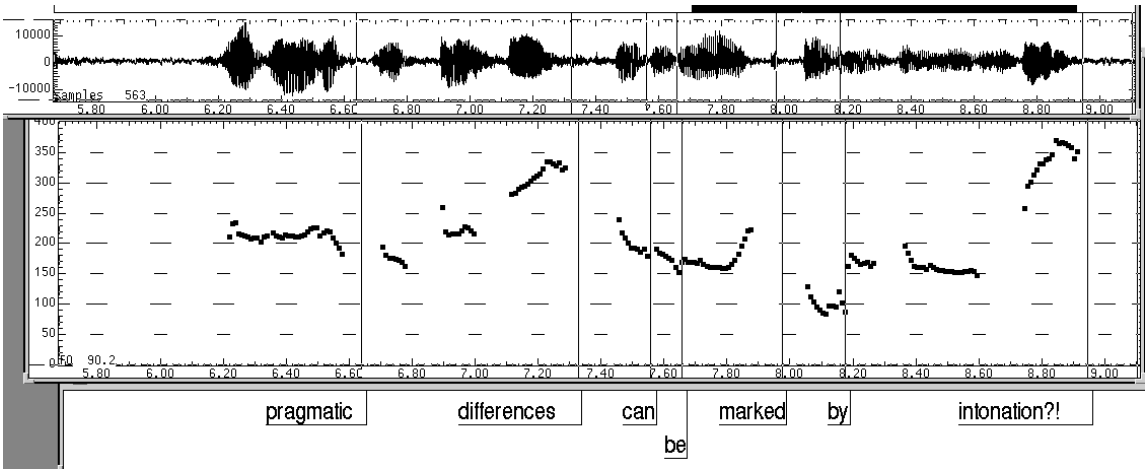


Figure 6. Pragmatic differences cued by intonation alone– an echo/incredulous question. "Pragmatic differences can be marked by intonation?!"

Thus, in incredulous/echo questions there is a clear form difference in the intonation and a clear functional difference in the pragmatics. This arbitrary mapping must be incorporated into the model of the speaker/hearer's knowledge of English. However, even in such clear-cut cases in which it is widely agreed that a functional difference is present and that the functional difference is conveyed by the intonation, there is a lot of cross-linguistic variation. (For a discussion of question formation using intonation in Kipare, Korean, and Mandarin Chinese, see, for example, Herman, 1996.) Thus, in each language, the form-function mapping is conventionalized differently. This means that a model of intonation is part of the model of grammar, and cannot be assumed to fall out from universal principles of any kind (such as for example, running out of breath at the end of a sentence causing the fundamental frequency to lower).

Since the intonation is part of the grammar, a phonemic-style or morphemic-style linguistic analysis of the intonation is necessary to discover the components of the tune and what their functions are. What makes a "phonemic analysis" of intonation more difficult than a phonemic analysis of segmental contrasts is that the functions involved are not lexical semantic meanings but rather discourse functions. Thus, a phonemic analysis of intonation cannot use the traditional field work paradigm of eliciting lists of words with different meanings and observing the sound differences between them. Rather, the same (phonemic) principle must be used, but with a slight modification. The phonemic principle involves matching differences in form to differences in function and similarities in form to similarities in function. This is necessary in a study of the intonation too, but recognizing that the functional differences are not lexical semantic differences but rather differences in information structure. This recognition does not make the phonemic analysis of intonational contours an impossible task, however, since there do exist formal definitions of discourse structure, which will be discussed further in section 1.4. For now, the form side of this form-function mapping will be described, with the understanding that the historical basis

for the discovery of the sound structure as described here has relied on phonemic principles of analysis.

The goal in the linguistic analysis of intonational tunes in English has been to identify intonational tones which bear pragmatic meaning. In some studies, such as Bolinger (1958), Liberman and Sag (1974), and Hirschberg and Ward (1992), entire tunes are examined to see what their corresponding meaning is. This approach has been refined by work such as Pierrehumbert and Hirschberg (1990) to decompose the tune into individual tonal morphemes which are associated with pragmatic meanings. The analogues of segmental "phonemes" in this type of analysis are the elements H and L, which combine in various ways to form tonal "morphemes," (such as H* or L+H*, to be discussed later) which are the minimal meaning-bearing units in the system.

Another potential objection relates to the question of whether the tune as a whole has a pragmatic meaning or whether the individual tones are the units which carry meaning. Certain intonational contours in English seem to be idiomatic and are referred to by names in the intonational literature, such as the "calling contour," the "surprise-redundancy contour," and the "uncertainty contour." The existence of such contours might be seen as calling into question the enterprise of decomposing the tune down into tones (as Pierrehumbert and Hirschberg, 1990, do). One potential explanation which unifies the studies on particular intonational contours with the compositional approach is that these are stylized idiomatic contours, analogous to frozen idiomatic expressions. Another potential explanation is that the pragmatic meaning of the tune is due to just one of the components of the contour and not to the entire contour. (See, e.g., Pierrehumbert and Hirshberg, 1990, p. 295) This issue is testable by minimal manipulations of the stylized contours to see which elements of the pragmatic function remain after the manipulations and hence which elements of the pragmatic function can be attributed to which elements of the sound structure. For now, since studies on the nature of intonation have in fact been successful at

finding functional categories at the level of the tones and since this approach is more in line with linguistic analysis at other levels, it will be assumed that the tune can be further decomposed and that the tones are the level of functional category to be investigated in intonational analysis.

Studies on the nature of intonation have been successful at finding functional categories. For example, Beckman and Pierrehumbert (1986, p. 272, based on the work of Pierrehumbert, 1980) point out that positing a "downstep" phenomenon in English in which the pitch range is lowered and compressed allows for the postulation of only two tones, L and H, in English intonation contours. The many potential intermediate levels seen in acoustic analysis are thus a result of the downstep rule and not of expanding the inventory of tone levels. Thus, the phonology of intonation can be established using phonemic principles and given a rich enough phonetic implementation.

There is a rich history of studies on the phonemic system of English intonation, having roots in the British school (see e.g. Palmer, 1922, Kingdon, 1958, O'Connor and Arnold, 1961, and Halliday, 1970). For a discussion of the history of the British school and comparisons of the British school to other lines of research, see Cruttenden (1986) or Ladd (1996). In this thesis, I will be working with one system for describing the grammar of English intonation developed in the early 1980s which has been very influential in subsequent studies on intonation.

This particular system used for describing the intonation of mainstream American English (as well as for Standard Southern British English and mainstream Australian English) is based on the analysis of the phonology of English intonation by Pierrehumbert (1980). This particular analysis has resulted in a transcription system for transcribing English intonation, the ToBI (Tones and Break Indices) system (Beckman and Ayers, 1994). In the ToBI system, there are three different types of units which are posited, with

each type being associated with one of three different levels in the prosodic hierarchy. The tones at each level can be either high (H) or low (L).

There are tones associated with particular syllables (pitch accents, marked with "*" such as H* or L*). Also posited are bi-tonal pitch accents composed of two tones, only one of which (the "starred" tone, which is symbolized by a tone marked with an asterisk) is associated with a particular syllable and the other of which is the "trailing" or "leading" tone and which follows or precedes the starred tone. An example would be L+H*, in which the H component is associated with a particular syllable and the L is the leading tone, marked by a "+".

A higher level in the prosodic hierarchy for English is the intermediate phrase, and there are also tones aligned with the edges of intermediate phrases (called phrase accents, marked with "-" i.e. H- or L-). The final pitch accent in the intermediate phrase is the "nuclear" accent. One or more intermediate phrases comprise an intonational phrase, and there are also tones aligned with the edges of intonational phrases (boundary tones, marked with "%" i.e. H% or L%). Thus, each intonational phrase contains one or more intermediate phrases. In this thesis, the phrase accents and boundary tones together will be referred to as "edge tones," to set them apart from the pitch accents which are associated with particular syllablea.

Another possibility is for a H tone to be downstepped, and this applies to starred H tones, starred H tones in bitonal pitch accents, H tones which are leading or trailing tones in bitonal pitch accents, and H phrase accents. Example transcriptions would be !H* or L+!H*. Downstepping is the contraction of the pitch range at the occurrence of the tone in question after an earlier H tone has set the pitch range.

This model of intonation includes two aspects. The first aspect is the tune or melody, which includes information such as whether a tone is high or low. The second aspect is the metrical structure and phrasing, which includes information such as whether a

syllable is stressed or reduced and whether it is phrase-initial, phrase-medial, or phrase-final. For example, the two utterances in the following figure are differentiated only by a difference in phrasing, with the example on the left having an intonational phrase break between the word “stop” and the word “state,” and having the interpretation “you must stop because it is the state law” (which is the intended interpretation of this phrase as seen on school buses). The example on the right, on the other hand, does not have an intonational phrase break between the word “stop and the word “state,” and although it has the same sequence of words and the same pitch accents, its meaning would be “state law should be stopped.” Thus, the metrical structure and phrasing aspect of intonation is crucial in interpretation, but in this thesis, I will be more concerned with aspects of the tune, not with aspects of the stress or rhythm.

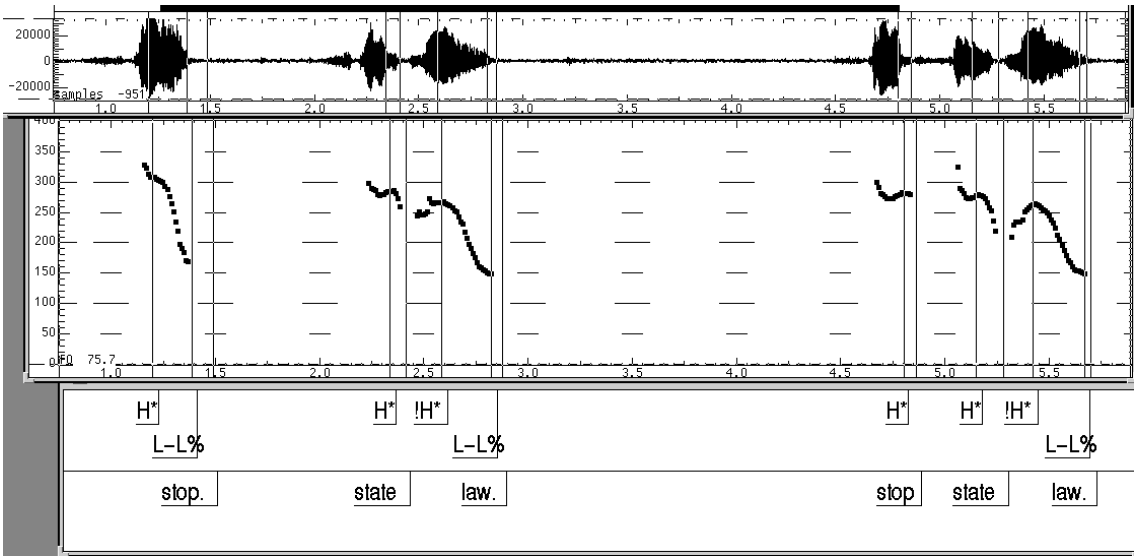


Figure 7. Example with and without a phrase break.

Furthermore, I will be more concerned with deciphering the meanings of the phrase accents and boundary tones (the edge tones), not with the pitch accents. The predictions

will be stated in terms of phrase accents and boundary tones, not pitch accents. For a study of the discourse functions of pitch accents, see, for example, Nakatani (1997).

The tonal system just described, developed on the basis of linguistic analyses of intonational contours (Pierrehumbert, 1980 and Beckman and Pierrehumbert, 1986), can be used to transcribe the intonation of utterances, as seen in the following example, repeated from Figure 2. In this example, the utterance is composed of two intermediate phrases, each of which is also an intonational phrase, thus making two intonational phrases in the utterance. Each intermediate phrase is delimited by a L- phrase accent and each intonational phrase is delimited by a L% boundary tone. There is a H pitch accent associated with two of the syllables in the first intermediate phrase (the word "these" and the second syllable in the word "examples"), and a H pitch accent associated with two of the syllables in the second intermediate phrase as well (the first syllable in the word "different" and the first syllable in the word "boundary").

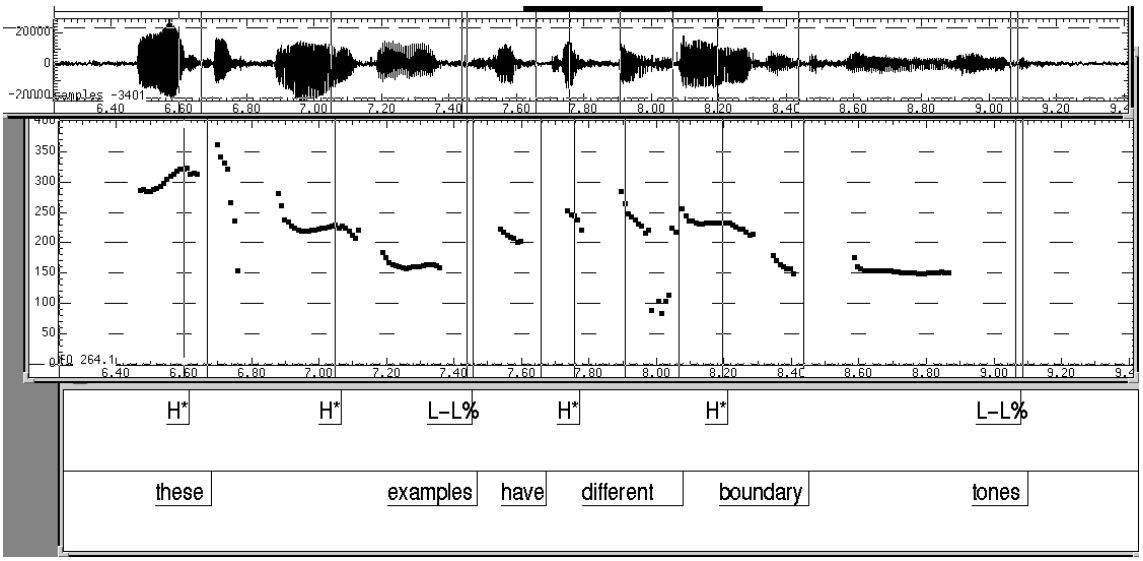


Figure 8. Example of pitch accents, phrase accents, and boundary tones.

This transcription system has been subjected to tests of inter-transcriber agreement (Pitrelli et al., 1994) and consistent behavior in transcribers was found, indicating that it is an adequate tool for the transcription of intonational tunes.

Alternative accounts exist, such as Gussenhoven's (1984) analysis of the edge tones as a complex with the nuclear accent. For the purposes of this work, because of the rigorous testing for intertranscriber consistency to which it has been subjected, the ToBI transcription system will be used. One further benefit of the ToBI system based on Pierrehumbert (1980) is that it does separate out the pitch accents from the phrase accents and boundary tones, which will turn out to be convenient for stating predictions about discourse structure. Thus, for the purposes of this thesis, the ToBI system will be used as a convenient transcription system, providing a simple way of stating the predictions regarding the use of intonation to mark discourse segmentation.

1.4 An Explicit Model of Discourse Structure

In order to discuss discourse structure, there should be some theory of what the components of discourse structure are and how they fit together. Here too, there is a rich tradition of studies and theories. Such work ranges from studies in social psychology and interactions among conversation participants to work in rhetoric and the structure of narratives to research in linguistics and models of pragmatics and information structure. Rather than exploring the wide variety of theories of discourse structure, in this thesis I will be working with one system of modelling discourse structure, which has been used in several studies of intonation and discourse to date.

This is the model described by Grosz and Sidner (1986). Grosz and Sidner's (1986) model of discourse structure has three components, the linguistic structure (the text itself, the actual sequence of utterances or "discourse segments"), the intentional structure (the purposes of each discourse segment and how they are hierarchically related to each

other), and the attentional structure (a model of the conversational participants' focus of attention as the discourse unfolds).

The linguistic structure is simply the text of the discourse as uttered by the conversational participants. As an example, the discourse below is given.

Sue won't go out with John unless he shaves off his beard and stops drinking.
He still had his beard this morning.
He hasn't bought a razor in three months.
Also, he was at a bar last night with his old college buddies.
His friends had to drive him home at the end of the evening.

Figure 9. Example of linguistic structure of discourse.

The discourse in Figure 9 is a coherent narrative, and can be used to answer the question "Does John stand a chance with Sue?" Thus, the discourse as a whole is a unit. Furthermore, it seems that the discourse can be broken down into smaller units, as shown in the following figure.

I. Sue won't go out with John unless he shaves off his beard and stops drinking.

II. He still had his beard this morning.
He hasn't bought a razor in three months.

III. Also, he was at a bar last night with his old college buddies.
His friends had to drive him home at the end of the evening.

Figure 10. Example of linguistic structure and discourse segments.

It is difficult to define exactly how the discourse is divided into discourse segments and what the properties of discourse segments are. Despite the difficulties in defining the properties of discourse segments, there seems to be agreement that discourses can be broken down into such segments. Grosz and Sidner note that the intuition that the discourse can be broken up into discourse segments has been reported across a wide range of discourse types (1986, p. 177). One empirically testable issue here is whether there is widespread agreement among listeners/readers on where the discourse segment boundaries are. Nakatani et al. (1995) provide a guide to discourse annotation usable by discourse-segmenters unfamiliar with any particular theory of discourse structure. This guide can be used by listeners to annotate discourses as to their segmentation (and other properties, to be described shortly). Studies on discourse structure annotation which have tested inter-annotater consistency have found high levels of agreement. For example, Grosz and Hirschberg's (1992) study of professionally read newscasts found that "For segment beginnings and endings, mean percent agreement among our labelers was better than 74% in all conditions." (p. 432) Hirschberg and Nakatani's (1996) study of the Boston

Directions Corpus (consisting of elicited speech using direction-giving tasks) also examined inter-labeler reliability in discourse segmentation. They measured reliability using the "kappa coefficient (k)" as a test statistic and found that the average k scores showed weak inter-labeler reliability for labelers who were labeling from text alone while labelers who were labeling from text and speech had average k scores of .8 or better, indicating a high degree of consistency among transcribers. Thus, there is some experimental evidence that people do agree on discourse segmentation. However, there are some cases in which the discourse segmentation is more controversial, and annotaters make different judgements about where segment boundaries are. For example, Walker, Joshi, and Prince (1998) explain that listeners may disagree about where segment boundaries are "either because they construct different mental representations of the segmentation of a discourse, or because segments are naturally defined at varying levels of granularity." (p. 20) As an example, they give the following short discourse, with segment B being ambiguous as to whether it is in the same discourse segment as A or whether it initiates a new discourse segment.

A. Her sister not being home, she hung up.
B. Her sister came home a short time later,
C. heard her messages,
D. heard her sister calling for help.

Figure 11. Example of ambiguous discourse segmentation.

Thus, although it has been shown that in general there is consensus among annotaters as to how discourses should be segmented, there are also cases in which the discourse segmentation is ambiguous due to differing degrees of granularity of

segmentation or due to different understandings of the discourse. Nonetheless, it is uncontroversial that the text of a discourse, consisting of a series of utterances, can be aggregated into discourse segments, and it is the text with its division into segments which comprises the "linguistic structure" in this model.

Further examining the example discourse, it might also be noted that the discourse segments can be arranged hierarchically, shown typographically with indentation.

I. Sue won't go out with John unless he shaves off his beard and stops drinking.

II. He still had his beard this morning.
He hasn't bought a razor in three months.

III. Also, he was at a bar last night with his old college buddies.
His friends had to drive him home at the end of the evening.

Figure 12. Example of discourse segments with hierarchical arrangement.

In this hierarchical arrangement, the first discourse segment dominates the second and third discourse segments. The second and third discourse segments, although subordinate to the first, do not dominate each other. Argumentation for the segmentation of the discourse into segments as well as for the arrangement of the discourse segments into a hierarchical structure might take as evidence the purposes of the different utterances. That is, an argument for the segmentation of the discourse into three discourse segments might argue that the first discourse segment's purpose is to set up the conditions under which Sue would get involved with John. The second and third discourse segments' purposes are to establish whether each of the two conditions has been met. The second and third sentences

in the discourse (the two sentences in Figure 12 (II)) work together to help establish whether the first condition for Sue to go out with John has been met. Since they are achieving the same purpose and addressing the same goal in the conversation, these two sentences can be justified as grouping together into a discourse segment. Thus, in justifying the discourse segmentation which breaks up the text into groups of utterances, the purposes of the discourse segments in the conversation can be used as argumentation.

This type of argumentation, namely, identifying the purposes or goals of parts of the discourse in order to justify the segmentation, acknowledges the basic, essential purposive nature of discourse. The incorporation into the model of discourse structure of the insight that utterances in a discourse are intended to "achieve some purpose, or goal, which may in turn be part of a larger plan for the discourse," (Roberts, 1998, p. 371) motivates the postulation of the second component of Grosz and Sidner's model of discourse structure, the intentional structure. The intentional structure identifies the purpose for each discourse segment (why it is being uttered) as well as the relationships between the discourse segment purposes (their hierarchical arrangement). An example of intentional structure which has both discourse segment purposes and their hierarchical arrangement is the example dialogue used earlier. This example is further annotated below to include the discourse segment purposes, which follow the "WHY?" label, and the hierarchical arrangement, which is illustrated typographically with indentation.

WHY? Set up the conditions under which Sue would get involved with John

Sue won't go out with John unless he shaves off his beard and stops drinking.

WHY? Establish whether the first condition (that John shave his beard) has been met

He still had his beard this morning.

He hasn't bought a razor in three months.

WHY? Establish whether the second condition (that John stop drinking) has been met

Also, he was at a bar last night with his old college buddies.

His friends had to drive him home at the end of the evening.

Figure 13. Example of discourse segment purposes and hierarchical arrangement.

As Nakatani et al. (1995) say, the discourse segment purposes are arranged in a hierarchical organization, similar to an outline. Roberts (1996) notes that interlocutors in a conversation have goals which they are trying to achieve. As the interlocutors attempt to pursue their conversational goal, they may set up a strategy of inquiry which pursues sub-goals or sub-questions which serve to anwer the meta-goal or super-question, thus imposing a hierarchical structure on the conversational goals. Grosz and Sidner (1986) note that although the list of discourse purposes is open-ended, it seems that there is only a small number of relations between discourse segment purposes.

Grosz and Sidner (1986) list two relationships that can hold between discourse segment purposes, dominance and satisfaction precedence. When one discourse segment purpose (DSP2) is subordinate to another discourse segment purpose (DSP1) and the satisfaction of DSP2 provides part of the satisfaction of DSP1, then DSP1 dominates

DSP2. On the other hand, when one discourse segment purpose (DSP1) must be satisfied before another discourse segment purpose (DSP2) can be satisfied, then DSP1 satisfaction-precedes DSP2. The dominance relation will be exemplified first.

In order to exemplify the relationships among discourse segment purposes, a longer discourse than the one shown above will be used in order to have a more complete example. Nakatani et al.'s (1995) example discourse, an instruction monologue, will be used. The introductory part of the instruction monologue is shown first.

We're going to be making sole, stuffed with shrimp mousse.
In the small bag is the sole and the shrimp.
And there are ten small sole fillets
and there's a half a pound of medium shrimp.
Okay, and you're going to need a blender to make the mousse
So you should get your blender out.

Figure 14. Example linguistic structure of instruction monologue.

Within this larger discourse segment, it can be seen that there are two sub-parts, the introductory remark and the beginning of the instructions. Within the second sub-part containing the beginning of the instructions, there are also two sub-parts, one related to the ingredients and one related to the equipment. Thus, the discourse can be divided into three discourse segments as follows (as proposed by Nakatani et al., 1995).

I. We’re going to be making sole, stuffed with shrimp mousse.

II. In the small bag is the sole and the shrimp.
And there are ten small sole fillets
and there’s a half a pound of medium shrimp.

III. Okay, and you’re going to need a blender to make the mousse
So you should get your blender out.

Figure 15. Example of discourse segmentation of instruction monologue.

Furthermore, it can be seen that the introductory remark dominates the two following sub-parts, which begin the instruction section, but that the two instruction sub-sections do not dominate each other.

I. We’re going to be making sole, stuffed with shrimp mousse.

II. In the small bag is the sole and the shrimp.
And there are ten small sole fillets
and there’s a half a pound of medium shrimp.

III. Okay, and you’re going to need a blender to make the mousse
So you should get your blender out.

Figure 16. Example of discourse segmentation of instruction monologue.

Furthermore, in justifying this particular segmentation of the discourse, the discourse segment purposes can be drawn on as argumentation. The discourse segment purposes are given in the following example, as adopted from Nakatani et al. (1995).

WHY? Teach new cook to make stuffed sole.
We're going to be making sole, stuffed with shrimp mousse.

WHY? Explain steps of initial preparation of ingredients and equipment.
WHY? Identify ingredients.
In the small bag is the sole and the shrimp.
And there are ten small sole fillets
and there's a half a pound of medium shrimp.

WHY? Instruct new cook to get equipment ready.
Okay, and you're going to need a blender to make the mousse
So you should get your blender out.

Figure 17. Discourse segment purposes of instruction monologue.

Note that in the example in Figure 17, the first discourse segment purpose dominates the second (annotated typographically with indentation) because the satisfaction of DSP2 (explaining the steps of the initial preparation of ingredients and equipment) contributes to the satisfaction of DSP1 (the overall goal of teaching the new cook how to make stuffed sole). Note also that the two subordinate discourse segment purposes in Figure 17 group together into a larger discourse segment with its own discourse segment purpose (that of explaining the steps of the initial preparation) even though that dominant

discourse segment purpose does not correspond to some single utterance in the discourse, but rather to a larger segment composed of smaller segments.

Nakatani et al. give an analogy for dominance relationships by saying that they are "similar to the way in which one action (e.g., making an appetizer) is part of the accomplishment of another, encompassing action (e.g., making dinner)." (p. 8)

Thus, the DSPs in a discourse can be organized into a hierarchy of dominance relations. To exemplify the possible relationships in a dominance hierarchy of intentions, a more extended subset of the Nakatani et al. discourse annotation will be used, including the example just given and continuing on to the preparation of the mousse. Again, the discourse segment purposes are listed with their corresponding discourse segments underneath.

DSP1 WHY? Explain steps of initial preparation of ingredients and equipment

DSP2 WHY? Identify ingredients

In the small bag is the sole and the shrimp.

And there are ten small sole fillets

and there's half a pound of medium shrimp.

DSP3 WHY? Instruct new cook to get equipment ready

Okay, and you're going to need a blender to make the mousse

So you should get your blender out.

DSP4 WHY? Explain how to make shrimp mousse

Okay, the first thing you want to do, we should do is we should make the shrimp mousse.

DSP5 WHY? Tell how to prepare shrimp

And, what you want to do is you want to take the shrimp, okay

and you want to peel and devein them.

DSP6 WHY? Describe peeling

Okay, what you do is you peel the outer shell off.

Figure 18. Extended example of hierarchy of dominance relations.

In this set of discourse segment purposes, there are actually two types of dominance relationships between pairs of DSPs, which could also be understood as "dominance" and "non-dominance." The dominance relations can be characterized as "mother-of" relations and the non-dominance relations can be characterized as "sister-of" relations (although Nakatani et al. do not formalize the hierarchy in this way, but rather discuss the structure as a topical outline with subsegments and bigger segments). It is not

clear if this exhausts the types of possible relationships, but it is a way to characterize the relationships between DSPs in a simple discourse. Both of these types of relationships result in lists of ordered pairs.

The list of "mother-of" ordered pairs is as follows, with any DSP1 which dominates another DSP2 (indicated typographically in the discourse by indentation of the dominated DSP2) being in the "mother-of" relation to it.

<<DSP1, DSP2>, <DSP1, DSP3>, <DSP4, DSP5>, <DSP5, DSP6>, <DSP4, DSP6>>

Figure 19. Example list of "mother-of" relations.

This "mother-of" relation or "dominance" relation in the intentional structure listed here as a set of ordered pairs thus has several mathematical properties. It is transitive, as can be seen by the fact that since DSP4 dominates DSP5 and DSP5 dominates DSP6, then DSP4 also dominates DSP6. It is also asymmetric, since if any DSP1 dominates DSP2, then DSP2 does not and in fact cannot dominate DSP1. It is irreflexive, in that no DSP dominates itself. And it is non-connected, in that there are pairs of DSPs which do not have a dominance relation between them, such as DSP1 and DSP4. Thus, since this hierarchy is transitive, asymmetric, irreflexive, and non-connected, it can be classified as a partial order on the disourse segment purposes.

The list of "sister-of" relations is as follows.

<<DSP1, DSP4>, <DSP2, DSP3>>

Figure 20. Example list of "sister-of" relations.

This list is transitive, but would need a set of three sister DSPs in order to demonstrate that. It is transitive because for any DSP1 which is a sister of DSP2, if there were a DSP3 which was also a sister of DSP2, then DSP1 would also be a sister of DSP3. It is also symmetric, as can be seen from the fact that if there is a DSP1 which is a sister of DSP2, then DSP2 is also in the sister-of relation to DSP1. It is also irreflexive, since a DSP1 is not a sister to itself. This relationship is also non-connected in that there are elements in the hierarchy which are not in a sister-of relation to each other.

These two relationships, mother-of and sister-of (or dominance and non-dominance), serve to characterize the dominance relationships in the intentional hierarchy of a simple discourse.

The other type of relationship, besides dominance, which is described by Grosz and Sidner is the relationship of satisfaction-precedence. This relationship requires that DSP1 must be satisfied before DSP2 can be satisfied. Again, discourse segment purposes in instruction monologues with several steps are good examples, because they often contain several instructions which must be completed in a certain order. The following example is from Nakatani et al. (1995).

DSP1 WHY? Tell how to find vein by cutting.

DSP2 WHY? Tell how to remove vein.

Figure 21. Example of satisfaction-precedence relation.

In this case, DSP1, finding the vein, must be satisfied before DSP2, removing the vein, can be satisfied. Thus, DSP1 satisfaction-precedes DSP2. This relation is transitive, because if a DSP1 satisfaction precedes a DSP2, and further a DSP2 satisfaction precedes a DSP3, then DSP1 must also be in a satisfaction precedence relation with DSP3. It is

asymmetric, in that if DSP1 satisfaction precedes DSP2, then DSP2 cannot satisfaction precede DSP1. It is irreflexive, in that a DSP cannot satisfaction precede itself. And it is non-connected, because there are pairs of DSPs which do not have a dominance relation among them. Thus, satisfaction precedence is a partial order on discourse segment purposes.

Thus, the intentional structure is composed of two elements, the discourse segment purposes themselves and the hierarchical relationships among discourse segment purposes, which consist of mother-of and sister-of relations and which impose a partial order on the discourse segment purposes.

In general, the intentional structure can be viewed as a logical structure, as suggested by Roberts (1996). If the discourse segment purposes or intentions are formalized as questions (the "question under discussion"), then the relationships holding between discourse segment purposes (as just described) can be understood formally as entailment relations holding between questions. This will not be explored further here, but the generalization is that the intentional structure can be interpreted as reflecting the logical structure of the discourse in a formal semantic way.

The two components of discourse structure discussed so far are the linguistic structure, which is the text of the discourse and its segmentation, and the intentional structure, which provides purposes for each segment and the hierarchical relationships among the purposes. The third component in this model of discourse structure is the attentional structure, which models changes in conversational participants' attentional state. That is, intuitively, the intentional component models the "why?" of discourse segments while the attentional component models what the discourse is about at that point in the conversation. In other words, it models what is relevant to the discourse at that point in the conversation, what conversational participants are focussing on at that point in the

conversation, and what is salient to the conversational participants at that point in the conversation.

The attentional structure is modelled as a stack, the "focus stack" (which is a non-deterministic push-down automaton). In this stack, each discourse segment has a focus space containing entities made salient in that discourse segment as well as the DSP for that discourse segment (which "reflects the fact that the conversational participants are focused not only on what they are talking about, but also why they are talking about it" (Grosz and Sidner, 1986, p. 181)). Roberts (1998, p. 373) argues that the "entities" in the focus stack are the "entities mentioned in the discourse segment with the properties predicated of them," and she assumes that these entities are discourse referents. The discourse referents are thus pointers or indices to what is being referred to. In sum, the focus spaces in the focus stack are assumed to contain pointers to what is being referred to (discourse referents), the properties predicated of those entities and the relationships holding between them, and the discourse segment purposes. The focus stack containing this information is a dynamic space, which changes as the conversation progresses and as the logical structure of the discourse develops and emerges. Discourse referents of entities in the current discourse segment (and their properties and relations, and the discourse segment purposes of that discourse segment) are pushed onto the focus stack as they become salient. Then, focus spaces pop off the focus stack as they are completed and move out of attention. Single focus spaces can pop off individually as they are completed. Alternatively, if completely addressing the dominated (daughter) DSP2 allows for the completion of the dominating (mother) DSP1 as well, then after DSP2 has been completed then DSP1 is also complete and both focus spaces can be popped off of the focus stack. And in fact, it seems that groups of focus spaces can pop off together, depending on how many DSPs transitively dominate each other. As Grosz and Sidner explain, "the focusing structure is

parasitic upon the intentional structure, in the sense that the relationships among DSPs determine pushes and pops." (p. 180)

The following explanation with accompanying diagrams, based loosely on Grosz and Sidner (1986), illustrates the focus stack over the course of a short discourse.

The organization of the discourse segments and discourse segment purposes in this example is such that DSP1 dominates DSP2, DSP1 dominates DSP3, and DSP2 and DSP3 do not dominate each other in any way. The example given above, which has this structure, is repeated here with both the discourse segment purposes and the discourse segments themselves.

DSP1 WHY? Set up the conditions under which Sue would get involved with John

Sue won't go out with John unless he shaves off his beard and stops drinking.

DSP2 WHY? Establish whether the first condition (that John shave his beard) has been met

He still had his beard this morning.

He hasn't bought a razor in three months.

DSP3 WHY? Establish whether the second condition (that John stop drinking) has been met

Also, he was at a bar last night with his old college buddies.

His friends had to drive him home at the end of the evening.

Figure 22. Example with one dominant DSP and two subordinate sister DSPs

At the beginning of the discourse, the focus stack is empty, since the conversation is not about anything yet.

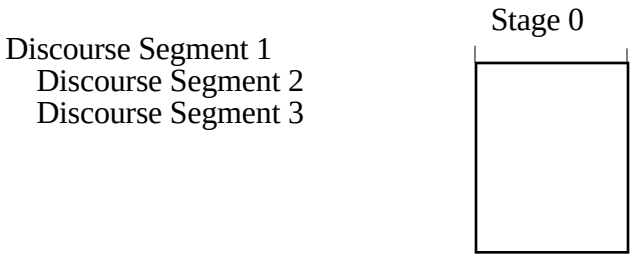


Figure 23. Example focus stack, stage 0.

Then, the discourse referents (and its properties and relations, and the discourse segment purposes) corresponding to the entities in the first discourse segment are pushed onto the focus stack, so it is the only entity in the focus stack at stage 1. For simplicity, the sum of the information in each focus space will be abbreviated as DR1, DR2, and so on, which will be understood to stand for the discourse referents referred to in the corresponding portion of the linguistic structure in the first discourse segment, along with its properties and the discourse segment purpose. At this stage, since the discourse referents for the first discourse segment are the only elements in the focus stack so far, there is just the one focus space.

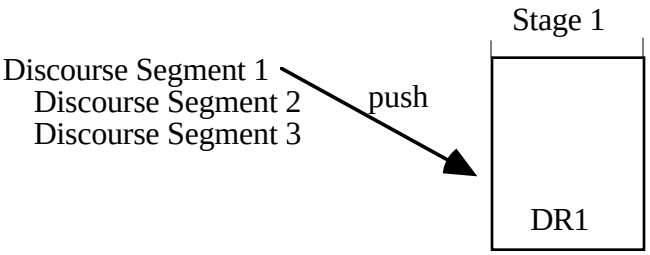


Figure 24. Example focus stack, stage 1.

The discourse referents for the first discourse segment, DR1, remain in the focus stack, since that information is still salient in the attentional state in the discourse and so there is no reason to pop it out of the focus stack yet. But now discourse segment 2 is reached and so a focus space containing all of the information in DR2 is also pushed into the focus stack, on top of the first focus space.

The actual course of events is not entirely clear, since according to Grosz and Sidner, the attentional state is a property of the discourse itself, not of the discourse participants. Thus, it is not clear at which point the logical structure of the discourse becomes a property of the discourse and is recognized as such by the discourse participants. It could be assumed that the focus spaces are added incrementally, as each utterance is reached. This assumption is in fact problematic for cases in which there seems to be a discourse segment purpose for a larger discourse segment and not for a single utterance, as in Figure 17. In that example, it cannot be the case that the discourse referents and discourse segment purposes for single utterances are pushed onto the stack as the utterance is encountered. Also, it may be that the segmentation of utterances into discourse segments relies on information from several utterances together. This is a subject for further study, and the following discussion will treat the focus stack as if the discourse referents and discourse segment purposes within a focus space are simply added to the focus stack incrementally as each discourse segment is reached, for the sake of ease of exposition.

The information in focus spaces which are lower down in the stack is accessible to the attentional state higher up in the stack. Only entities which are currently in the focus stack can be used in interpretation. Intuitively, that means that things which have already been said but which have not yet been completely addressed and so are not a closed topic yet can still be referred to and used in the conversation by conversational participants. Since the discourse referents of earlier discourse segments are lower in the stack (because,

schematically, focus spaces are pushed in at the top), then that means that their discourse segments have been uttered earlier and thus would still be accessible when the next discourse segment is uttered.

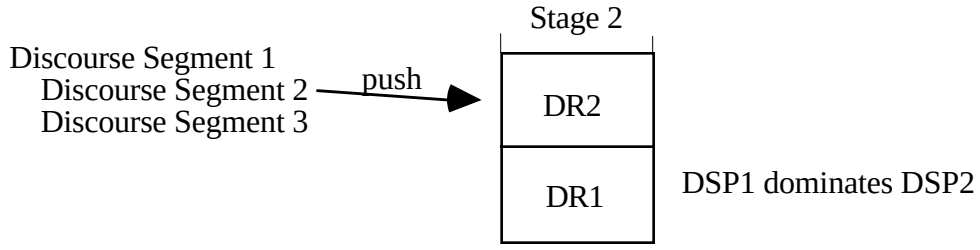


Figure 25. Example focus stack, stage 2.

Upon completion of the second discourse segment, its corresponding focus space can be popped off of the focus stack. It is thus no longer relevant to the discussion. (This is an example of what Nakatani (1996) refers to as a “push-pop”.) In other words, that sub-topic is closed for discussion and the entities in DR2 cannot be used for interpretation any more (in the ideal, clean and neat model of discourse structure in which focus spaces pop off of the focus stack completely and in one go).

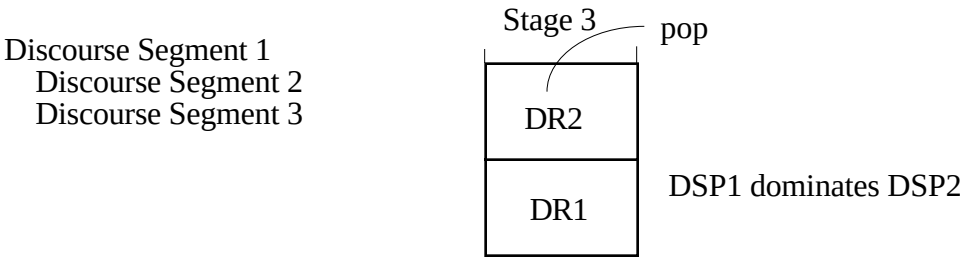


Figure 26. Example focus stack, stage 3.

Now, since the first discourse segment is still in the attentional state in the discourse because it is still the super-topic of the conversation and has not been completely addressed yet, the focus space containing the information abbreviated here as “DR1” remains on the focus stack. Only one of the first discourse segment’s two sub-topics has been addressed (and it has not been decided that any of the topics cannot be addressed– see Roberts, 1996, p. 100, on the issue of practically unanswerable questions or issues), so the topic introduced in the first discourse segment is still “out there” and is still “on the table,” modelled by the focus space containing DR1 still being in the focus stack. At this stage in the logical development of the discourse, discourse segment 3 is reached and the discourse referents associated with it (along with their properties and relations, and the corresponding discourse segment purpose) are pushed onto the focus stack as well. Again, the information in the focus space containing DR1 is still accessible, since it is still relevant and still in the focus stack.

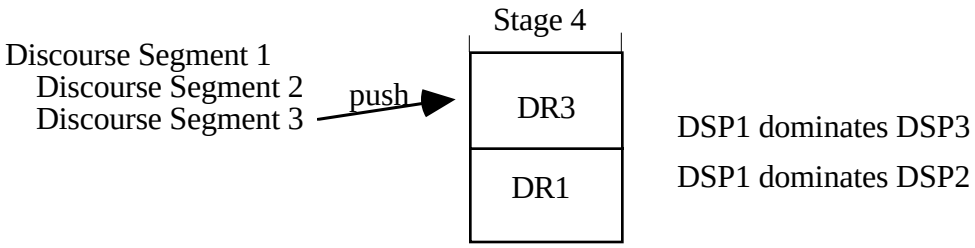


Figure 27. Example focus stack, stage 4.

Again, upon completion of the third discourse segment, its corresponding focus space can be popped off the focus stack, and it is no longer relevant in the conversation. (Again, this would be an example of what Nakatani (1996) refers to as a “push-pop.”) So the sub-topic related to the third discourse segment is closed for discussion, and the information in the focus space containing DR3 cannot be used for interpretation any more.

At this stage in the logical development of the discourse, the topics raised by both daughters of the first discourse segment have been addressed. Thus, the two sub-parts of the dominant discourse segment have been closed, meaning that the dominant discourse segment itself has been completely addressed and need no longer be salient in the discourse. Thus, the focus space containing DR1 can also be popped off the focus stack, since the issue raised by the first discourse segment has been completely addressed and is no longer relevant to the conversational participants. This is schematized with the focus spaces for both DR3 and DR1 being popped from the focus stack together, but again, the issue of whether there is an order to popping off these two focus spaces or whether the recognition of the logical structure of the discourse allows both focus spaces to be popped from the focus stack together is an open question, whose answer will depend on which view of the attentional state is taken. (Is the focus stack purely a reflection of logical structure or does it involve some component of the psychological state of the conversational participants, and would these different views have different effects in the model?)

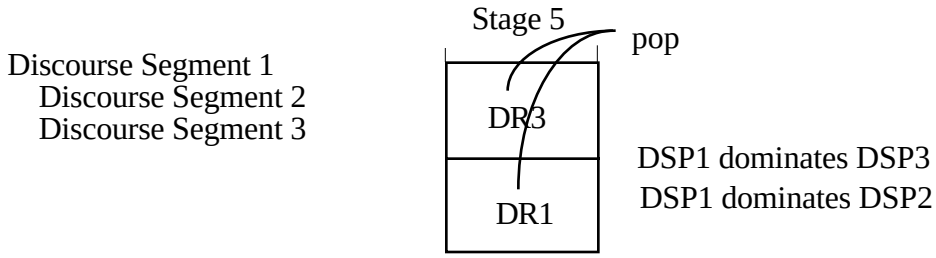


Figure 28. Example focus stack, stage 5.

At the end of the conversation, the focus stack is empty since the conversation is no longer happening, but the intentional structure has been completely established.

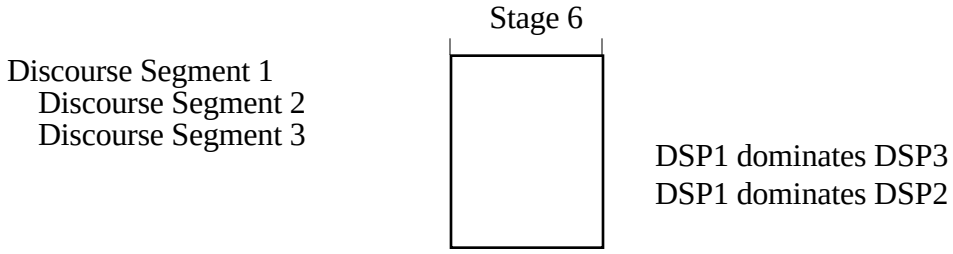


Figure 29. Example focus stack, stage 6.

It can be noted that each dominance relation (which is part of the intentional structure) is at some point expressed in the focus stack (as one focus space being next to, or "local" to another focus space). For example, the fact that DSP1 dominates DSP2 is expressed in the focus stack at stage 2 and the fact that DSP1 dominates DSP3 is expressed in the focus stack at stage 4.

The expression of transitive dominance relations in the focus stack holds as well. If DSP1 dominates DSP2, and DSP2 dominates DSP3, then by transitivity DSP1 dominates DSP3. This holds in the focus stack as well, as shown in the following example, in which the transitive dominance of DSP3 by DSP1 is expressed in both the intentional structure and the focus stack (by the focus space containing DR3 being above the focus space containing DR1 in the focus stack at stage 3).

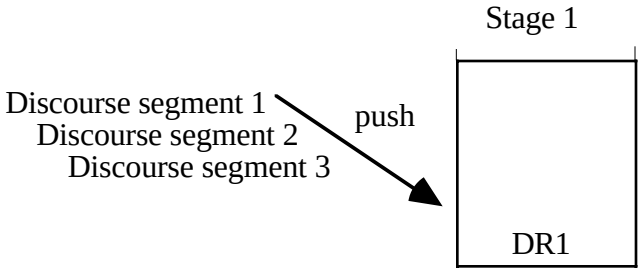


Figure 30. Example focus stack, transitive dominance, stage 1.

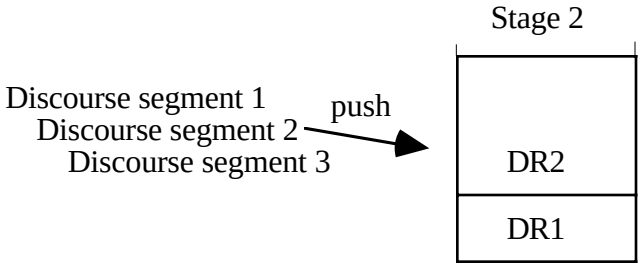


Figure 31. Example focus stack, transitive dominance, stage 2.

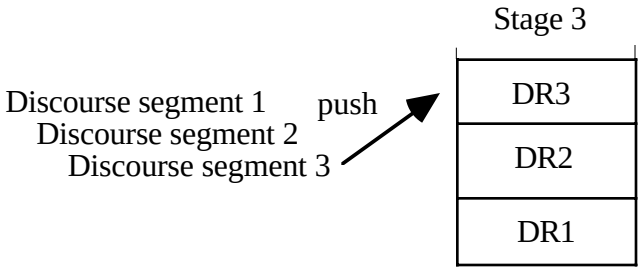


Figure 32. Example focus stack, transitive dominance, stage 3.

Thus, the focus stack is a push-down automaton which reflects the dominance relations in the intentional structure as the logical structure of the discourse is developed. However, there are some parts of the dominance hierarchy which cannot be expressed in the focus stack, namely, the more "global" aspects of the intentional structure. For one, it is impossible to express sisterhood of discourse segments in the focus stack. (Whether this really holds is an open question, because it is not clear whether focus spaces are popped completely and cleanly from the focus space when that sub-topic is complete or whether it is more a case of fading activation of that focus space. This is another area in which further study and experimentation is necessary.) For example, note that when two discourse segments are sisters (as with the second and third discourse segments in the example in Figure 22), at no one stage in the course of changes to the focus stack is this sister-of relationship expressed in the focus stack itself.

As another example of a property of the dominance hierarchy which is not expressed in the focus stack, since "the focus structure does not include the intentional structure as a whole" (Grosz and Sidner, 1986, p. 180), then that means that the question of whether a set of daughter discourse segments exhausts the set of daughters or not cannot be answered just by examining the attentional structure. That is, it cannot be known from the stages in the development of the focus stack whether the daughters which have already been pushed onto and popped off of the focus stack form the complete set of the daughters of the dominating discourse segment or not. Figure 33 illustrates this. The figure shows that the dominance hierarchy contains three daughters of DSP1. Thus, from the dominance hierarchy, it is clear than when the focus spaces containing DSP2, DSP3, and DSP4 have all been popped from the focus stack, then the focus space containing DSP1 can be popped from the focus stack as well.

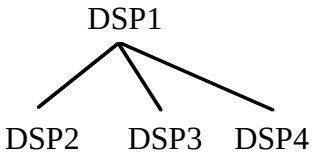


Figure 33. Dominance hierarchy with 3 daughter DSPs.

However, from the focus structure, it cannot be known whether the fourth discourse segment DSP4 is the last of the complete set of daughters of DSP1 or whether there is some other potential DSP5 which is also a member of the set of daughters of DSP1.

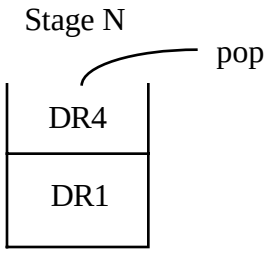


Figure 34. Example focus stack, stage N, complete set of daughter discourse segments?

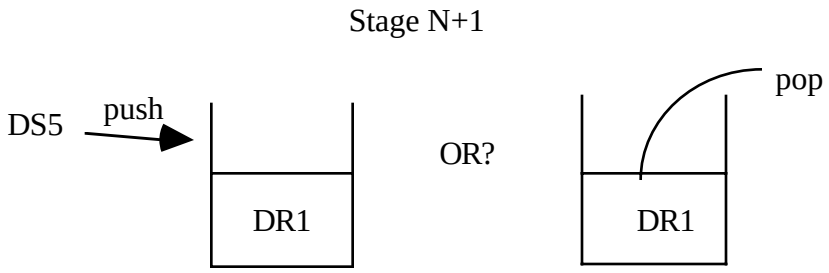


Figure 35. Example focus stack, stage N+1, complete set of daughter discourse segments?

The idea that the question of whether the complete set of daughters has been addressed cannot be answered in the focus stack makes intuitive sense as well, since it is not clear from what is "out there" or from what is relevant in the conversation at some stage in the discourse (that is, what is in the focus stack) whether the larger topic has been completely addressed or whether there are still some sub-topics which need to be addressed. The complete list of sub-topics comes from the mind of the conversational participants and from what their purposes and their agendas are, which is exactly what the intentional structure is expressing. So it makes intuitive sense that although dominance relations, including transitive ones, are expressible in the focus stack, there are elements of the intentional structures, such as whether the complete set of daughter DSPs has been expressed, which are not expressible in the attentional structure.

In sum, the attentional structure expresses all of the dominance relations in the dominance hierarchy that are active at some point during the conversation. That is, all of the ordered pairs in the list of "mother-of" relations are expressed in the focus stack at some point during the course of the discourse. However, the attentional structure cannot express certain aspects of the intentional structure (what I am calling here more "global"

relationships) such as “sister-of” relations or whether the complete set of daughters to some mother have been satisfied yet.

The focus stack, in modelling the attentional structure, has proven useful in a variety of linguistic problems. For example, Grosz and Sidner (1986) suggest that the search for possible referents of definite noun phrases and pronouns is constrained to entities which are within the focus stack at that stage in the discourse (p. 191). Roberts (1998) develops this idea and shows that pronouns’ antecedents come from within a logically accessible set of referents and within that set, from within the attentionally accessible space, interpreted in terms of Grosz and Sidner’s focus stack. (p. 390) That means that pronouns can be understood to be referring only to entities which are still current in the conversation . (This understanding does not capture cataphora, in which the pronoun precedes its “antecedent.” Such cases may involve linguistic markers pointing to the need to hold off on binding the pronoun until further entities are introduced into the attentional structure, as in the example in Figure 36, below). The focus stack may also prove useful in defining consistently thorny issues in pragmatics such as “relative salience” (see e.g. Roberts, 1998, p. 390), “topic” (see e.g. Grosz and Sidner, 1986, p. 191), interruptions and cohesiveness in discourses (see e.g. Grosz and Sidner, 1986, p. 192), or “focus” (see e.g. Roberts, 1996 for a related discussion). Thus, extending the ideas about pronouns’ antecedents coming from attentionally accessible discourse referents, it would seem that in a very general (and intuitive) sense, interpretation can occur only with respect to entities which are in the focus stack together, and hence which are accessible to the interlocutors in the conversation at that particular stage in the logical development of the discourse. This general summary does not, of course, address issues such as deixis accompanying use of demonstrative pronouns such as “this” or “that”, as discussed by Roberts (1998, p. 367), but such issues would go too far afield of the general point here, which is the claim that the attentional structure must be one of the components of the

discourse structure, and that interpretation (including, but not limited to, the search for antecedents of pronouns) can occur relative to the discourse referents of entities which are within the focus stack.

In summary, the theory of discourse structure proposed by Grosz and Sidner (1986) has three components– the linguistic structure (the text of the discourse and the discourse segmentation), the intentional structure (the discourse segment purposes and the hierarchical relationships among them), and the attentional structure (which is modelled as a focus stack realizing the dominance relations over the course of the conversation). In this thesis, I will be using this model of discourse structure, but will not be proposing modifications to it, since I am more concerned with the "form" side of the form–meaning relationship. However, the use of this model of discourse structure in this thesis will allow for advances over earlier studies of phonetic correlates of discourse structure by allowing for explicit definitions of what the discourse structure is. The use of this model will allow the work discussed here to progress beyond the ideas proposed in earlier studies of differences in "sentence" vs. "paragraph" boundaries being expressed intonationally (see e.g. Lehiste, 1975) by explicitly recognizing the hierarchical relationships between subordinate and dominant discourse segments, and also by allowing for the possibility of more than two levels. It will also and allow for progress beyond the idea of utterances having different positions within a discourse segment (see e.g. Grosz and Hirschberg, 1992) by recognizing hierarchical dominance and sister-of relations holding between discourse segments.

Recognizing the three components of discourse structure proposed by Grosz and Sidner (1986), it can be seen that what is essential for a conversation to proceed felicitously is that the intentional and attentional structures be cued by the linguistic structure. The exact ways in which aspects of intonation are used to express these various facets of the discourse structure will be further discussed in the next section.

1.5 The Meaning of Edge Tones and the Discourse Functions of Phonetic Implementations

Turning back to the idea that there are phonetic correlates of discourse structure, the question is exactly what aspect of discourse structure is being conveyed by the phonetic correlates. That is, which component of the Grosz and Sidner (1986) model of discourse theory do the phonetic correlates (such as decreases in F0, presence of laryngealization, etc.) mark? In order to address this question, first the intonational phonology needs to be understood, in order not to confound the meanings of the different edge tones with the discourse functions of the phonetics.

Exactly what meanings are to be assigned to intonational edge tones is addressed by Pierrehumbert and Hirschberg's (1990) account of the meaning of intonational contours. The discussion here will build on their approach. Phrase accents are seen as conveying the degree of relatedness of one intermediate phrase to a subsequent intermediate phrase. (p. 287) Boundary tones indicate "whether the current phrase is to be interpreted with particular respect to a succeeding [intonational] phrase or not." (p. 287)

According to Pierrehumbert and Hirschberg, intermediate phrases marked with H- form a unit with a subsequent intermediate phrase. On the other hand, a L- indicates separation of the intermediate phrase marked by the L- from the subsequent intermediate phrase. (p. 302) Similarly, if there is a H%, then the intonational phrase marked with a H% should be interpreted with respect to a subsequent intonational phrase (p. 305). On the other hand, a L% does not convey directionality in interpretation.

(There are seeming counterexamples in which the H% is the final tone in the utterance and the utterance is the final one in the discourse. For example, there is a final H% in the answer/background "Anna came with Manny" example in Pierrehumbert (1980). The explanation for this type of case may be revealed through further investigation of the discourse structure of cases in which the answer is presented first, followed by an almost

afterthought of a background, in which more clarifying material is presented even after the complete answer has already been given. This remains a topic for further research, and if the answer/background case turns out to have an explanation in the discourse structure, then the proposed meanings of the phrase accents and boundary tones may be accepted as stated, whereas if there is not some explanation from the discourse structure, then the meaning of the phrase accents and boundary tones may need some refinement to account for such cases.)

Pierrehumbert and Hirschberg (1990, p. 305) give the following pair of examples for the directionality of interpretation, and for the idea that a H% signals that the current intonational phrase is to be interpreted with respect to a subsequent intonational phrase.

I. L-L%
A. My new car manual is almost unreadable.
L-H%
B. It's quite annoying.
L-L%
C. I spent two hours figuring out how to use the jack.

vs.

II. L-H%
A. My new car manual is almost unreadable.
L-L%
B. It's quite annoying.
L-L%
C. I spent two hours figuring out how to use the jack.

Figure 36. Example of contrast between H% and L%.

Pierrehumbert and Hirschberg claim that in the first case, the referent of "it" is likely to be interpreted as "my spending two hours figuring out how to use the jack" whereas in the second case, the referent of "it" is likely to be interpreted as "my new car manual." They suggest that a H% contributes "to the interpretation of intentional structure by signalling the existence of hierarchical relationships..." (p. 306) Recall that in Grosz and

Sidner's (1986) theory of discourse structure and in Roberts' (1998) extension of this theory regarding anaphora, interpretation of pronouns can only occur within the focus stack (which makes intuitive sense, because interpretion will occur relative to what is relevant and salient in the discourse at the time). Thus, I interpret Hirschberg and Pierrehumbert's "meaning" of the H% as marking that the discourse referent for another discourse segment needs to be pushed onto the focus stack before anything can be popped from the focus stack. In other words, the H% marks cataphora, in which the referent for the pronoun occurs after the pronoun itself in the linear order of the conversation. Since the referent comes after the pronoun, that means that more information needs to be added to the focus stack before interpretation and binding of pronouns (which can select only entities within the focus stack and hence attentionally salient) can occur. In the example in Figure 37, the discourse referents associated with the corresponding segment of the linguistic structure in the B sentences (abbreviated here with the relevant pronoun "it" for convenience) in the first case would still need to be in the focus stack when the pronoun "it" was interpreted as referring to "my spending two hours figuring out how to use the jack," because the antecedent of the pronoun can only be a discourse referent which is within the focus stack. Therefore, the binding of the pronoun "it" must be to some element in the focus stack (intuitively, something that is still "out there" or "on the table" in the conversation). (It would require further study to discover whether the presence of a H% requires cataphora or whether it simply permits it. It does seem, however, as will be discussed below, that the L% does not make cataphora available as an option in the interpretation of the pronoun.)

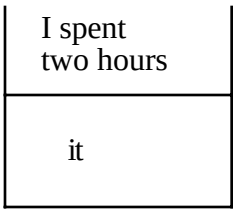


Figure 37. Example focus stack, binding of pronoun given H%.

The discourse segment in B in the second case would be pushed onto the stack containing the discourse referent for the linguistic material in discourse segment A, so “it” in B could be interpreted as referring to “my new car manual” in A, and then both could be popped off of the stack before C was pushed onto the stack with the pronouns having been satisfactorily bound.

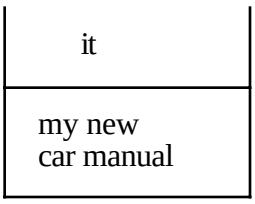


Figure 38. Example focus stack, binding of pronoun given H% on previous utterance.

If, on the other hand, there is a L%, then according to Pierrehumbert and Hirschberg (1990), the current utterance “may be interpreted without respect to subsequent utterances.” (p. 307) In the example above, in the second discourse, that can be interpreted as meaning that the utterance in B does not need to wait for any further information, but

may be interpreted with respect to what is already on the focus stack. This is illustrated in Figure 38, in which "it" can refer to "my new car manual," whose discourse referent is already in the focus stack. That is, in terms of Grosz and Sidner's model of discourse structure, it seems that L% can signal that the focus space of that discourse segment can be popped off the focus stack.

In general, then, extending the notion of "interpretation" beyond just binding of pronouns, it would seem that a H% signals that more material is to be pushed onto the focus stack.

Thus, it would seem that the choice of edge tone is tied to discourse dominance relations which are expressed in the discourse dominance hierarchy (part of the intentional structure). I will be referring to these as "local" relationships in the intentional structure, and they are realized also as locality of focus spaces within the focus stack at different points in time. This interpretation relates the phonological category of an edge tone to the discourse function of expression of the dominance hierarchy.

Note that this is a slightly different use of "local" compared to Nakatani's (1997) use of "local." Nakatani (1997) uses "global" to mean concerning relationships among discourse segments while she uses "local" to mean concerning relationships among utterances within a single segment. In my usage, "global" also means relationships among discourse segments, but specifically the types of relationships which cannot be expressed in the focus stack, such as whether the entire set of daughter DSPs has been expressed. For me, "local" means a dominance relationship between two utterances, but not necessarily in the same discourse segment. Thus, for me, the relationship between "Sue won't go out with John unless he shaves off his beard and stops drinking" and "He still had his beard this morning" would be a local relationship, because it is expressing a dominance relationship between two utterances, but for Nakatani this would be a global relationship because the two utterances are in different discourse segments.

Now that the discourse meanings of the phonological categories of edge tones have been explored, the phonetic correlates of discourse structure can be examined without confounding the meanings of tonal categories with discourse functions of the different phonetic realizations.

Phonetic correlates have been postulated to mark discourse structures such as "sentences" or "paragraphs." Lehiste (1975), one of the first researchers to study this topic, makes use of the idea of sentences and paragraphs (or "prosodic paragraphs"), and more recent work such as Hirschberg and Nakatani (1996) makes use of the idea of "discourse segments" as found in the linguistic structure component of the Grosz and Sidner (1986) model of discourse theory. Hirschberg and Pierrehumbert (1986) suggest that within the category of the low edge tones, the degree of final lowering, how compressed the pitch range is at the end of the utterance, indicates an utterance's position in the "discourse hierarchy" (which seems to be interpretable as the dominance hierarchy in the intentional structure). They further suggest that varying degrees of final lowering correspond to the level of the intentional hierarchy which the utterance "completes." (p. 141)

I interpret this to mean that the phonetic implementation of tones in a phrase ending with L% marks the more "global" level of intentional structure by indicating how many focus spaces can be popped off of the focus stack together. That is, since a L% indicates that the utterance does not need to be interpreted with respect to a subsequent utterance, then that utterance can be popped off of the focus stack. Furthermore, recall that any number of focus spaces can be popped off of the focus stack at the same time (depending on the dominance relations and whether all of the daughters of the mother DSP have been comletely addressed). If all of the dominated (daughter) DSPs of a dominating (mother) DSP have been addressed, then the dominating (mother) DSP can be popped off of the stack as well. This comes into play when there is more than one daughter. For example, in the figure below, when DSP2, DSP3, and DSP4 have all been addressed, then DSP1 can

be be considered to have been addressed as well, but until all three of the subordinate DSPs have been addressed, the dominant DSP must remain on the focus stack (as the over-arching topic of discussion, as it were, in intuitive terms).

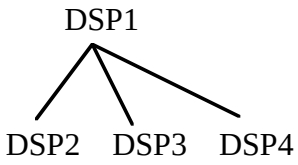


Figure 39. Example of intentional structure with three daughter DSPs.

Thus, if there is a mother DSP with three daughter DSPs, it is only in the intentional structure that the question of whether DSP2, DSP3, and DSP4 exhaust the set of the daughters of DSP1 can be answered, and not in the attentional structure. The fact that DSP1, the dominating DSP, has been addressed, is of course, eventually indirectly expressed in the focus stack as the popping of DSP1, but the knowledge that DSP1 can in fact be popped from the focus stack depends on the recognition of more of the intentional structure than can ever simultaneously be present in the focus stack, because sisters will not be in the focus stack simultaneously. (This is in the cleanest theoretical model, although whether it is true that an earlier sub-topic, though completely finished, cannot have some entity within it used in interpretation of a later discourse segment is an empirical question. For example, would it be possible to have an entity in an earlier discourse segment serve as an antecedent to a pronoun in a later discourse segment? Also, would such cases be subject to other constraints such as recency effects and processing capabilities? These remain questions for future research.) The fact that it is only in the intentional structure that the question of whether the full set of sub-topics has been addressed can be answered makes intuitive sense as well, since it is only in the speakers' plans for the conversation that the

full range of sub-topics is present, and the knowledge of what the full set of sub-topics is going to be is not "out there" available to all of the conversational participants. Thus, conversational participants must indicate somehow in the linguistic structure when the list of sub-topics (or dominated DSPs) has been exhausted, and so complete that portion of the conversation (with its dominant DSP). And this, I suggest, is exactly where the phonetic correlates of discourse structure come into play. Since the negotiation of topics in discourse and the recognition that one's interlocutor is finished with some sub-topic is of vital importance for a discourse to proceed smoothly, it is crucial to indicate the global intentional structures in conversation. The amount of final lowering, more global pitch range manipulations, and other phonetic factors having to do with pausing and other tempo modulation, volume control, or voice quality are used in speech as cues as to how many focus spaces can be popped from the focus stack at one time.

Thus, a first pass at understanding the meanings of the different categories suggests that a H edge tone means that more information should be pushed onto the focus stack while a L edge tone means that the utterance may be popped from the focus stack. Within the L category, the phonetic implementation of the tones in a phrase marked by a L% (within the range allowed by the boundaries of the particular tone categories) is tied to the more global relations in the discourse, such as whether the daughters of the dominant DSP have all been satisfied, and hence how many discourse segments can be popped from the focus stack (how deep into the focus stack conversational participants can go to empty out focus spaces).

A common misunderstanding by beginning ToBI transcribers may help explain the difference between the meaning of tone choice and the discourse function of the phonetic implementation. The two meanings are: "to be interpreted with respect to a subsequent utterance," for H%, versus "there are more sisters to this DSP," when there is no final lowering in the phonetic implementation. The two meanings can be seen to be similar in

some sense because both mark a type of incompleteness. And in fact, beginning transcribers sometimes think that they are hearing a H%, and cite as supporting evidence the sense of "incompleteness" and "more to come," when in fact a closer examination of the F0 trace and the meanings reveals that what they were hearing was a L% without final lowering, not an actual H%, and the sense of "incompleteness" was the sense that "more sister DSPs are to come," not that "this intonational phrase should be interpreted with respect to subsequent phrases."

This understanding of the edge tones' meanings as conveying local dominance relationships while the phonetic implementations convey global relations (in the case of L edge tones) might explain the asymmetry between the existence of "final lowering" (the extra lowering of a final pitch accent when followed by a L edge tone) and the apparent non-existence of "final raising." Presumably, "final raising" would be analogous to "final lowering" in that it would involve the extra raising of a final pitch accent when followed by a H edge tone. This hypothetical phenomenon of "final raising" is not to be confused with the actually existing observed phenomenon of the upstep of a boundary tone following a H-phrase accent, which is discussed by Pierrehumbert (1980). Rather, "final raising" would presumably affect the final pitch accent, not the final boundary tone. The reason for the non-existence of "final raising" may be that in the case of H%, the discourse meaning is that this utterance is to be interpreted with respect to subsequent utterances. However, utterances can only be pushed on to the focus stack one at a time, as they are uttered one by one. Thus, a higher final pitch accent could not serve to indicate, for example, that more utterances were to be pushed onto the focus stack simultaneously, because utterances are intrinsically introduced into the attentional structure one-by-one. On the other hand, in the case of L%, which allows focus spaces to be popped off of the focus stack (dismissed from the conversational participants' attention), there can be any number of focus spaces popped off of the focus stack at a time, as whole topics are wrapped up, so the exact

phonetic implementation of the tones in the phrase marked by the L boundary tone can serve as a cue to the depth of the focus stack that is to be emptied. However, given the uncertainties about the exact modelling of the focus stack and the questions about when and how discourse segments are recognized as segments are pushed into or popped out of the focus stack (and how much of the logical structure must be recognized for the pushing and popping to occur, and whether this involves the logical structure alone or some psychological element of the conversational participants), it may be plausible to postulate a linguistic marker that helps to indicate that several discourse segments are to be pushed onto the focus stack. In English, rather than a gradient "final raising" analogous to the final lowering which indicates the depth of focus spaces which are to be popped, it seems that the linguistic marker which might function to indicate that several focus spaces are to be added to the stack before popping is initial raising, in which the first peak in the first utterance at the beginning of a new topic is higher. If understanding of the discourse function of initial raising is correct, it might be predicted that a greater degree of initial raising would indicate a greater amount of focus spaces still to come, while less initial raising would indicate less focus spaces.

1.6 Two Hypotheses about Phonological Edge Tones, Phonetic Implementation, and Local and Global Discourse Structures

1.6.1 Phonological Edge Tones and Local Discourse Structures

The understanding of the meaning of the edge tones as expressing dominance relations between two discourse segments in the intentional and attentional structures leads to a testable hypothesis– the "same edge tones hypothesis." It would seem that no matter whether a particular dominance relation is expressed earlier or later in the dominance hierarchy, the choice of edge tones should be the same. That is, consider a situation in which DSP1 dominates DSP2, and DSP1 dominates DSP3, but DSP2 and DSP3 are not in

dominance relations with each other and neither DSP2 nor DSP3 dominates another DSP, as illustrated.

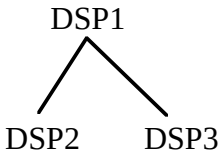


Figure 40. Example of two sister DSPs

As far as these dominance relations go, DSP2 and DSP3 can be expressed in either order in the linguistic structure (the text itself) as long as one does not satisfaction-precede the other. Again, the example from Nakatani et al. (1995, p. 5) will be used in which there is a mother DSP with two daughter DSPs. The two daughter DSPs are sisters and are not in a dominance relation with each other nor in a dominance relation with some other subordinate DSP. Thus, the two daughters could potentially be expressed in either order.

I.

DSP1 WHY? Explain steps of initial preparation of ingredients and equipment.

DSP2 WHY? Identify ingredients.

DSP3 WHY? Instruct new cook to get equipment ready.

vs.

II.

DSP1 WHY? Explain steps of initial preparation of ingredients and equipment.

DSP3 WHY? Instruct new cook to get equipment ready.

DSP2 WHY? Identify ingredients.

Figure 41. Example of two sister DSPs in both orders.

The prediction is that if the choice of edge tones is tied to the dominance relations, then no matter in which order the daughters occur, they will have the same edge tones (and in fact, those tones should be L edge tones, since one discourse segment does not need to be interpreted with respect to the other). So the discourse segment related to DSP2 should have the same edge tones in I as it has in II. (It is not clear yet exactly what the prediction is regarding the pitch-accents, and whether they would be expected to be the same in both I and II or not. Nakatani's work suggests that centering relations are also important in choice of pitch-accent, and the issue of locality in centering is still not clear.)

If the hypothesis that the same dominance relations are expressed by the same tone (no matter what the linear order of utterances) is verified, then that would imply that "locality" in discourse structure is not simply a matter of utterance adjacency, but rather a matter of dominance relations which are modelled in the intentional structure by the dominance hierarchy and which are modelled in the attentional structure as position in the

focus stack at different points in the discourse. So “locality” in both versions of the example above is between discourse segment 1 and discourse segment 2 because DSP1 dominates DSP2, and it should not matter whether discourse segment 3 intervenes between them in the linguistic structure or not.

1.6.2 Phonetic Implementation and Global Discourse Structures

This understanding of the phonetic implementation of tones as expressing more global intentional properties such as “the set of daughter DSPs has been completely addressed” also leads to a testable hypothesis– the “phonetic differences hypothesis.” It would seem that there could be two discourses which are quite similar to each other. In one, a particular dominance relation is expressed earlier in the dominance hierarchy, before other dominance relations, and so only that particular utterance can be popped off of the stack (a “push-pop” in Nakatani’s (1996) terms). In the other discourse, that same dominance relation is expressed later in the dominance hierarchy, in such a position that it is the last of a set of daughter DSPs and so allows the dominating DSP to be completely addressed, thus closing the super-topic as well and allowing for more discourse segments to be popped from the stack. It would seem that although the choice of edge tones should be the same for those two cases (at least, if the earlier prediction can be verified), then exactly how the tones in the phrase are implemented phonetically should be different. That is, consider again a situation in which DSP1 dominates DSP2, and DSP1 dominates DSP3, but DSP2 and DSP3 are not in dominance relations with each other, as in Figure 40 above. As far as these dominance relations go, DSP2 and DSP3 can be expressed in either order in the linguistic structure (the text itself) as long as one does not satisfaction-precede the other. Concentrating on DSP2, the prediction is that in the first order, the utterance expressing DSP2 would have a different phonetic implementation of tones than in the second order (even if the same tones are being expressed). In the first order, DSP2 is the only DSP that can be popped off of the stack after DSP2 has been addressed (because DSP3 is still

required to completely address the topic in DSP1). In the second order DSP2 and DSP1 can both be popped off of the stack after DSP2 has been addressed (because DSP2 exhausts the set of dominated DSPs of DSP1).

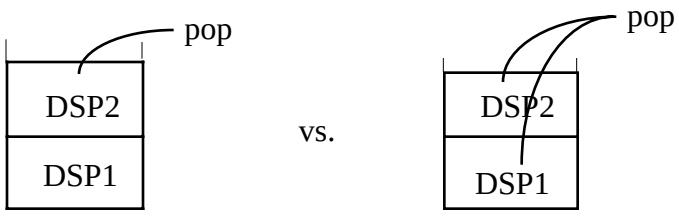


Figure 42. Example of a DSP's effect in two different discourse positions

Thus, the interpretation of the relationship between the phonological tones, the phonetic implementation, and aspects of discourse structure leads to two testable hypotheses, one regarding the relationship between the choice of edge tone and locality within the dominance hierarchy and the other regarding the relationship between phonetic implementation and more global properties of the intentional structure.

1.7 Testing the Hypotheses– Experimental Design

The first prediction (the "same edge tones hypothesis") is that if the choice of edge tones is tied to the dominance relations, then no matter which order the daughter DSPs occur in, they should have the same edge tones (and since they are sisters and not in a dominance relation to some subordinate DSP themselves, then they would be predicted to have L edge tones). The second prediction (the "phonetic differences hypothesis") is that when a particular dominance relation is expressed earlier in the dominance hierarchy (before other dominance relations, and so allowing only that particular DSP to be popped off of the stack), then the tones in that phrase would be implemented differently than in the case when that same dominance relation is expressed later in the dominance hierarchy (in

such a position that it is the last of a set of dominance relations and so allows the dominating DSP to be satisfied as well), allowing for more DSPs to be popped from the stack.

These predictions can be tested experimentally. The experimental corpus necessary is one in which there is a discourse with a dominance hierarchy in which the set of dominance relations is the same in two different versions of the discourse but the dominance relations are expressed in different orders. For this to be true, there needs to be one dominant DSP and at least two daughter DSPs, with neither daughter DSP dominating the other, no "satisfaction-precedence" relationship between them, and no subordinate DSPs to the daughter DSPs. In order to achieve these controlled relationships, a logical structure such as the following can be translated into a small natural language discourse.

if x and y, then z

x

y

Figure 43. Example logical structure of intentional hierarchy.

For example, the conditional (or rather, a negative version of it) could be something like "Sue won't go out with John unless he shaves off his beard and stops drinking." The first sentence sets up the logical structure for the participants, establishing the DSP of *z* as the dominant DSP and establishing that it has two daughters which need to be satisfied in order for the mother DSP to be satisfied. So in the example, the *z* would be "find out if Sue will go out with John," the *x* would be "find out if John has shaved off his beard," and the *y* would be "find out if John has stopped drinking." After setting up the logical structure for the participants, *x* is established, being pushed onto and popping off of the focus stack,

which enables the DSP of *z* to be partly answered, but not completely, so *z* cannot yet be popped off the focus stack itself. Then *y* is established, again being pushed onto and popping off of the focus stack, again directly above *z*, but allowing for the same local relationships within the focus stack at the two different times. But then after *y* is answered, it is logically concluded that *z*, allowing the DSP of *z* to be popped off of the focus stack and the conversation to end. Now, having controlled for the dominance relationships of the DSPs of the *x* and *y* sentences to the dominating DSP discourse segment purpose, reversing the order of *x* and *y* allows the DSP of *x* to be realized in two locations in the dominance hierarchy (either medially or finally), as well as allowing the DSP of *y* to be realized in two locations in the dominance hierarchy (also, either medially or finally). The following two examples show the logical structures that would be true of the two discourses.

if x and y, then z	or:	if y and x, then z
x		y
y		x

Figure 44. Example logical structures, subordinate DSPs in both orders.

Then, if natural language versions of discourses along the lines of these models are recorded being read by native speakers, then the *x* and *y* sentences from both orders can be extracted from the recording. There will be the pair of x-from-medial-discourse-position and x-from-final-discourse-position, which should both have the same dominance relationships to the mother DSP and so would be predicted to have the same (low) edge tones. (One way to check whether the intended intentional structure was achieved in the natural language translation of the logical structure would be to have the discourses

annotated according to the principles set forth in Nakatani et al. (1995). This will not be done in this experiment, but might be useful as a check in future work.) However, the difference between the two will be that x-from-discourse-medial-position will not yet allow *z* to be popped off the focus stack, but will just be popped off the focus stack itself. However, x-from-discourse-final-position will allow both itself and *z* to be popped off the focus stack. Thus, x-from-discourse-medial position should have a different phonetic implementation of tones than x-from-discourse-final position. Given the logical structure described above for the discourses, there will also be the pair of y-from-discourse-medial-position and y-from-discourse-final-position, again both having the same dominance relations to the mother DSP but allowing for different numbers of segments to be popped off the focus stack after the utterance of *y* in each case. This will give two pairs of sentences for intra-pair comparison from every discourse that is read (with each discourse being read twice, with its daughter DSPs in the two orders). This constitutes the design of the experiment.

Thus, this type of design tests whether leaving the dominance relations the same but changing their order results in the same edge tone choices in the two different cases. However, one component which is lacking in this set-up is the converse, which would be to minimally change the dominance relations and to see whether this resulted in different tones, and if so, what the differences in tone choice are. That is, Pierrehumbert and Hirshberg (1990) observed (albeit, anecdotally and not experimentally) what changes in tone resulted from slightly different pragmatic contexts and how changes in tone choice changed the discourse meanings. An extension of their work into an experimental paradigm such as the one proposed above, but changing the dominance relations (rather than leaving them the same) would most likely be enlightening as well, and is another area for future research.

Also, since this type of set-up calls for the reading of dialogues, the tone choice cannot be controlled as it can in an experiment such as, for example, Liberman and Pierrehumbert's (1984) use of lists of berry names which were modelled by the experimenters before being recorded by the subjects. That is, Liberman and Pierrehumbert's experiment involved the recording of particular tunes, and examined the implementation of those tunes. The experimenters demonstrated for the subjects the tunes that they wanted them to produce, thus controlling which tones were present in the recorded material. In the current experiment, subjects were asked to speak naturally, and the experimenter did not instruct the speakers as to which tune to use. And in fact, one of the questions under examination regards the tone choice itself. That is, it is of interest which edge tones are chosen by the speakers. Thus, the choice of which tone is used in production cannot be controlled in any event in this type of an experiment. However, this is not to imply that the problems encountered by the phonetic correlates of discourse structure approach are necessarily to be encountered in this approach. Even though the tone choice is not controlled in the materials, given the explicit model of the intonational structure, the choice of tones can be examined, and the acoustic phonetic analysis can be restricted to phrases with identical sequences of phonological intonational tones, in order to avoid comparing apples and oranges, as it were, and thus progressing beyond the earlier studies which compared acoustic measurements without ensuring that the measurements were for elements belonging to the same category.

Now, given a design in which the set of dominance relations is the same but the order in which they are expressed in the hierarchy is varied, the first step in the analysis is to test whether the choice of edge tones of the two utterances in the pair, differing only in position in the dominance hierarchy, but not themselves dominating any subordinate DSPs, are the same (and whether they are both L-L%, as predicted) by having the tunes transcribed according to the ToBI transcription system (Beckman and Ayers 1994). This

will allow an experimental test of whether the choice of edge tone conveys the dominance relations, because given the way the corpus was designed, the dominance relationships between pairs of sentences should be the same for each sentence. The only difference between the two sentences in a pair which have been extracted from the two different orders of reading is which location they were from in the dominance hierarchy. Thus, the prediction as stated above is that the choice of edge tones should be the same for both utterances in each pair. Testing this will allow for establishment of the "same edge tones" hypothesis.

Furthermore, in order to avoid the confound of potential differences in phonological tone category when testing for phonetic differences between utterances and when looking further into the sentence than just the part at the very end which is controlled by the edge tones, the ToBI transcription will again be useful. The transcriptions can be compared to see whether the pitch accents are similar enough betwen the paired tokens to allow for comparison and allow for the conclusion that the difference is a phonetic difference, not a phonological one. This is necessary since the tone choice was not controlled for, as it has been in studies such as Liberman and Pierrehumbert (1984) by modelling the pitch contour for production by the subjects. Actually, testing for similarity between pairs of transcriptions is an interesting problem in its own right, especially when trying to go beyond just a statement such as "these two tonal transcriptions are identical" to statements such as "these two tonal transcriptions are similar enough" by recognition of the fact that the tones may have been transcribed slightly differently due to confusability between two categories of tones, but that the tones should perhaps be viewed as containing the same tone targets. This issue will be discussed further in the Methods chapter. Thus, one of the first steps in testing the "phonetic differences hypothesis" is ensuring that the differences are not phonological differences.

The next step is to test whether x-from-discourse-medial position vs. x-from-discourse-final position are distinguishable perceptually, by playing such pairs to listeners for identification of which sentence was "paragraph final." (Since these are naive listeners, calling this "discourse-segment final" would not be meaningful.) This will determine whether discourse position is perceptible as such to listeners. Thus, in testing the "phonetic differences hypothesis," it is important to test whether these phonetic differences are indeed perceptible. If discourse positions are not perceptible to listeners, then that fact may indicate something about the variable nature of speech and the plethora of other cues which are available in natural conversational settings besides just slight differences in intonation. That is, in natural speech listeners do not just use differences in intonation to judge sentences, but rather rely on a variety of cues such as pronoun use, word order, and connector words. Furthermore, if the sentences are not identifiable as to their discourse location, then this may also help to explain interruptions in naturally occurring discourses. That is, it may be that there are cues which tend to guide listeners in discourse segmentation, but that those cues are more tendencies than absolutes. Thus, it may be that listeners do in fact mis-interpret prosodic cues from time to time, and perhaps think that their interlocutor is done with a sub-topic when in fact the interlocutor has not finished yet. If so, then this may explain why listeners sometimes accidentally interrupt their interlocutor during a conversation before the interlocutor is ready to yield the floor, or why the listener might attempt to change the topic before the speaker is ready to. If the utterances are identifiable as to their discourse locations (or if at least some of them are), then the next step is an acoustic analysis to identify what it was that listeners were using as cues.

If it does turn out that the choice of edge tones is the same when the dominance relations are the same but in different orders, and that the pitch accents chosen are similar enough to allow comparison without the confound of comparing across different categories of phonological tones, and furthermore that listeners can distinguish utterances from

different dominance hierarchy locations, (all of which will be shown to greater or lesser degrees) then it comes down to a study of the phonetic correlates of discourse structure, similar to earlier such studies, but with the understanding here that the "phonetic correlates" are the phonetic implementation of the phonological intonational tones in the phrase within the limits imposed by their categories (and not a hybrid of both phonetic correlates and tone choice) and that the "discourse structure" is the "global" component of intentional structure which cannot be expressed in the focus stack. Thus, this experimental design will allow for examination of intonational cues to discourse structure while avoiding the pitfalls of earlier studies which did not control for both phonological tone and discourse structural relationships.

1.8 Summary

In summary, what is being proposed here is a study of the phonetic correlates of discourse structure, both in terms of what the phonetic correlates are without the confound of differences in phonological intonational tones and also in terms of exactly which aspects of discourse structure are being conveyed. These issues will be explored in an experiment which is set up in such a way that two discourses with the same dominance relations but expressed in different orders result in pairs of utterances which have the same dominance relation to the mother DSP but allow different amounts of DSPs to be popped from the focus stack. The expectation is that in a pair of two corresponding utterances extracted from the two different discourses, the utterances will have the same choice of edge tones (which should be low) but that the phonological tones in the phrase will be implemented differently in the two cases.

CHAPTER 2

METHODS

2.1 Recording

2.1.1 Materials

The study employed read speech rather than spontaneous speech for the more careful controls allowed and for the comparisons possible, since there were specific requirements as to the attentional and intentional structure in the dialogues. Although spontaneous speech may be better because it is more naturalistic, it would be difficult to elicit spontaneous speech which allows for the same type of controls. That is, with spontaneous speech it would be difficult to control dominance relationships in the intentional structure so as to get matched dialogues which had the same dominance relationships in the dominance hierarchy, but in different orders in two different cases, which is what is required to test the predictions.

The corpus consisted of dialogues constructed to be read by native English speakers. The corpus was constructed such that a logical structure is imposed on the dialogue. There is a dominant DSP, and then at least two daughter DSPs. As one example, it can be established in the first sentence that "If x and y, then z." Then the dialogue can continue by implicating "x" and then by implicating "y," thus allowing the dominant DSP to be satisfied and popped off of the focus stack. The daughter DSPs (the x and y), should

not be in a dominance relation with each other, nor should one satisfaction-precede the other. Then the utterances associated with the two daughter DSPs can be read in both the "x y" order and the "y x" order, creating a situation allowing for the comparison of "x" from two different positions in the discourse dominance hierarchy and also allowing for the comparison of "y" from two different positions in the discourse dominance hierarchy. Ten such dialogues were created, so with the two different orders in which the sub-segments were read, that gives 20 dialogues in total to be recorded. The dialogues are listed below.

Are you going to need to change advisors?

I was talking to my advisor the other day.
She said she's going to be doing field research in Kenya.
(B) It's o.k. to have an advisor who's away, but only for one quarter.
(A) She'll be gone for the whole year.

Are you going to need to change advisors?

I was talking to my advisor the other day.
She said she's going to be doing field research in Kenya.
(A) She'll be gone for the whole year.
(B) It's o.k. to have an advisor who's away, but only for one quarter.

Figure 45. Dialogue 1

I know that Sue wanted to donate blood. Did she qualify as a donor?

If you're over 110 pounds, between ages 17 and 65, and haven't taken any medication for two weeks then you can donate blood.
I think she weighs about 130 pounds.
(B) She turned 32 last month.
(A) She hasn't taken any medication since that sinus infection last February.

I know that Sue wanted to donate blood. Did she qualify as a donor?

If you're over 110 pounds, haven't taken any medication for two weeks, and are between ages 17 and 65 then you can donate blood.
I think she weighs about 130 pounds.
(A) She hasn't taken any medication since that sinus infection last February.
(B) She turned 32 last month.

Figure 46. Dialogue 2

Do you think that John stands a chance with Sue?

She won't go out with him unless he shaves off his beard and stops drinking.
He still had his beard this morning.
(B) He hasn't bought a razor in 3 months.
Also, he was at a bar last night with his old college buddies.
(A) His friends had to drive him home at the end of the evening.

Do you think that John stands a chance with Sue?

She won't go out with him unless he stops drinking and shaves off his beard.
He was at a bar last night with his old college buddies.
(A) His friends had to drive him home at the end of the evening.
Also, he still had his beard this morning.
(B) He hasn't bought a razor in 3 months.

Figure 47. Dialogue 3

Did Bill pass the exam?

Only the people who went to the TA's office hours, went to the study session, or stayed up all night reading passed the exam.
He has class during office hours.
(B) I didn't see his name on the sign-in list from the study session.
(A) When I called him at 10 last night, his roommate said he was sleeping.

Did Bill pass the exam?

Only the people who went to the TA's office hours, stayed up all night reading, or went to the study session passed the exam.
He has class during office hours.
(A) When I called him at 10 last night, his roommate said he was sleeping.
(B) I didn't see his name on the sign-in list from the study session.

Figure 48. Dialogue 4

Have you satisfied all of your general education requirements yet?

You need one class in each of three areas– history, culture, and social science.
I took Western Civ last spring.
(B) Next fall I'm going to take Intro Anthropology.
(A) Linguistics 201, which I'm taking right now, counts as my social science.

Have you satisfied all of your general education requirements yet?

You need one class in each of three areas– history, social science, and culture.
I took Western Civ last spring.
(A) Linguistics 201, which I'm taking right now, counts as my social science.
(B) Next fall I'm going to take Intro Anthropology.

Figure 49. Dialogue 5

Do you think Ellen's going to get into graduate schools next year?

In order to get into grad school, you need a good GPA, lots of extracurricular activities, and good GRE scores.
She has a 3.7 cumulative average.
(A) Also, she's involved in lots of clubs and activities.
(B) And she took one of those training classes and so she did really well on her GRE.

Do you think Ellen's going to get into graduate schools next year?

In order to get into grad school, you need a good GPA, good GRE scores, and lots of extracurricular activities.
She has a 3.7 cumulative average.
(B) And she took one of those training classes and so she did really well on her GRE.
(A) Also, she's involved in lots of clubs and activities.

Figure 50. Dialogue 6

Are you going to be living in the same place next year?

I decided that if they didn't fix the leak in the roof or if they raised the rent then I'm going to move.
They patched up the roof again last month.
(B) The workmen told me that they don't think that will fix the problem.
Also, my renewal notice came last week.
(A) They raised the rent by $40.

Are you going to be living in the same place next year?

I decided that if they raised the rent or if they didn't fix the leak in the roof then I'm going to move.
My renewal notice came last week.
(A) They raised the rent by $40.
Also, they patched up the roof again last month.
(B) The workmen told me that they don't think that will fix the problem.

Figure 51. Dialogue 7

Can you come to the movies with us after dinner?

If I finish reading my history assignment and solve this math problem I can go.
I still have two chapters left to read.
(A) They take me about an hour each.
I haven't even started on the math yet.
(B) That usually takes me several hours.

Can you come to the movies with us after dinner?

If I solve this math problem and finish reading my history assignment I can go.
I haven't even started on the math yet.
(B) That usually takes me several hours.
I still have two chapters left to read.
(A) They take me about an hour each.

Figure 52. Dialogue 8

Are you going to be able to make it to the party tonight?

I don't have a car, so I aways need either a ride home or else someone to walk me home.
Pete always gives me a ride.
(A) But he's out of town this weekend.
Sue lives next door to me so we often walk home together.
(B) But she has a big exam she's studying for.

Are you going to be able to make it to the party tonight?

I don't have a car, so I aways need either someone to walk me home or else a ride home.
Sue lives next door to me so we often walk home together.
(B) But she has a big exam she's studying for.
Pete always gives me a ride.
(A) But he's out of town this weekend.

Figure 53. Dialogue 9

Can I get away with wearing these socks with the holes in the toes to the party tonight?

You know the party's going to be at Yuko's house.
(B) She still follows Japanese customs at home even though she's been in the U.S. for four years.
(A) One Japanese custom is that all guests remove their shoes on entering the house.

Can I get away with wearing these socks with the holes in the toes to the party tonight?

You know the party's going to be at Yuko's house.
(A) One Japanese custom is that all guests remove their shoes on entering the house.
(B) She still follows Japanese customs at home even though she's been in the U.S. for four years.

Figure 54. Dialogue 10

Thus, these 10 dialogues, each with two target sentences, give 20 pairs of sentences (with one member of the pair from the AB order and one member of the pair from the BA order) to allow for the testing of the predictions.

2.1.2 Subjects

The subjects in the recording session were all undergraduates in Linguistics 201, an undergraduate introduction to linguistics at the Ohio State University. They were paid $10 for their participation in the experiment. The recording took about half an hour. The date of birth, sex, and places lived of all of the eighteen speakers is listed in the table.

All of these speakers speak a "mainstream" variety of American English. Subject S18 is an African American woman and may have a slightly different variety (including a different intonational system) than most of the other speakers. None of these speakers are fluent bilinguals except Subject S16, who is bilingual in Spanish and English, having spoken Spanish since childhood.

Speaker #	Date of Birth	Sex	Places Lived
S1	11/17/78	M	0-18 South Jersey 18-19 Columbus, OH
S2	11/17/75	F	0-5 Detroit, MI 5-18 Cincinnati, OH 18-21 Columbus, OH
S3	1/12/79	F	0-7 Waterford, MI 7-18 Cleveland, OH 18-19 Columbus, OH
S4	10/6/77	M	0-2 Columbus, OH 3-7 Chillicothe, OH 7-21 Columbus, OH
S5	8/15/78	M	0-19 Columbus, OH
S6	1/28/79	F	0-5 Parma, OH 5-19 Strongsville, OH
S7	12/20/75	F	0-15 Cortland, OH 15-16 Mansfield, OH 16-19 Cortland, OH 19-22 Columbus, OH
S8	10/14/79	F	0-17 Buffalo, NY 17-18 Columbus, OH
S9	4/4/77	F	0-9 mo. Napoleon, OH 9mo.-12 Syracuse, NY 12-18 Solon, OH 18-21 Columbus, OH
S10	2/26/79	F	0-19 Columbus, OH
S11	2/26/77	F	0-20 Norwalk, CT 20-21 Columbus, OH
S12	4/3/79	F	0-19 Columbus, OH
S13	3/3/78	M	0-18 New Jersey 19-21 Columbus, OH
S14	10/4/77	F	0-20 Columbus, OH
S15	1/10/78	F	0-3 Kettering, OH 3-8 Huber Heights, OH 8-15 New Carlisle, OH 15-18 Tipp City, OH 18-20 Columbus, Oh
S16	7/7/78	M	0-20 Masillon , Ohio
S17	9/29/79	F	0-18 Lyndhurst, OH 18 Columbus, OH
S18	6/24/77	M	0-18 Cleveland, OH 18-20 Columbus, OH

Figure 55. Subject information from recording session.

2.1.3 Methods

The recording session took about half an hour to complete. Recordings were made in a double-walled sound-proof booth. The subjects used a head-mounted microphone recording into a TEAC V427-C tape recorder. The head-mounted microphone, which stays a constant distance from the subject's mouth even if the subject moves her or his head, allowed for comparison of amplitude across tokens. The dialogues were printed on cards. Each dialogue had two variants, with the daughter DSs in different orders. The variants were split up so that they occurred in different blocks. The cards were read one time each, taking a short break between the two blocks. The experimenter read the part shown in italics and the subject read the rest. The subjects were asked to speak as naturally as possible, "as if they were actors auditioning for a part and reading a script." The recordings were then digitized using ESPS Waves™ on a Sun Sparc 10 computer at 16 bits, 16 kHz sampling rate, and the sentences under consideration were extracted from the dialogues into individual files. The F0 traces were generated using the Waves™ get-f0 function, which is an auto-correlation method of F0 estimation which uses dynamic programming to adjust to local intensity variations and F0 trends.

2.2 Transcription of Intonational Tones

2.2.1 Transcription

The ToBI transcription was done by an expert labeller, naive to the purposes of the experiment (obviously, not the experimenter). She was paid $15 per hour for her transcription work. There were 720 sentences to be transcribed (18 speakers * 2 sentences per dialogue * 20 dialogues). Each sentence was transcribed with pitch accents, phrase accents, and boundary tones according to ToBI conventions. The inventory of pitch accents was H*, L+H*, L*, L*+H, H+!H*, !H*, L+!H*, and L*+!H. The inventory of phrase accent / boundary tone combinations was L-L%, L-H%, H-H%, H-L%, !H-L%, !H-H%.

2.2.2 Mapping Tone Similarity Space

In order to test for comparability of tone transcriptions, I devised a method of mapping out the transcriber's tone space and using the mapping to infer tone similarity (and hence potential transcription confusability), which then allowed for an acoustic analysis which only compares tokens which have the same or similar classes of tones. (This method of using multi-dimensional scaling was suggested by Keith Johnson.) This was done separately for pitch accents vs. phrase accents and boundary tones.

Other studies have also classed accents together into groups according to similarity along various dimensions. For example, Ayers (1996) found evidence in a psycholinguistic study for a primary perceptual split between nuclear accents versus prenuclear accents and unaccented syllables. Furthermore, within the category of nuclear accents, she found evidence for perceptual differences between expanded pitch range nuclear accents, regular nuclear accents, and downstepped nuclear accents, which nonetheless all behaved similarly to each other in the class of nuclear accents. Thus, she found groups of accents which were similar along perceptual dimensions. Along the same lines, Nakatani (1997) counted as prominent accents with H* in them or bitonal pitch accents, while she counted as non-prominent accents with L* and unaccented tokens. Similarly, in Jannedy's (1997) analysis, no accent, !H*, and L* are categorized as "non-prominent" while H*, L+H*, L*+H and L+!H* are categorized as "prominent." Thus, there is some precedent in identifying classes of pitch accents, although researchers have used various criteria for identifying the classes. In this experiment, the classes are identified on the basis of transcriber tone similarity judgements, and the transcriber's "tone space" is calculated using a multi-dimensional scaling analysis.

The ToBI transcriber who participated in the current experiment was asked to rate the similarity of all possible pairs of tone category labels on a 7-point scale. Pairs included categories such as H* vs. !H*, L+H* vs. H*, etc. "Unaccented" was included as a

category. Pitch-accents were rated separately from phrase accents and boundary tones, so there were pairs such as H* vs. !H* presented for similarity judgement, pairs such as L- vs. !H- presented for similarity judgement, pairs such as L- vs. L-L% presented for judgement, but no pairs such as H* vs. L- presented for judgement. Each pair was presented 3 times (in random order) to the transcriber for similarity judgements. The mean of the three similarity judgements was used. The decision to exclude pairs which involve combinations of pitch accents and phrase accents excludes from the analysis pairs such as for example L-H* vs. L+H*, in which it is not clear whether the L (and usual accompanying low part of the F0 trace) is due to a phrase break being present or to a low leading tone. It remains a subject for future research to see if certain combinations of tones are confusable with other combinations of tones or with individual tones.

The similarity judgement values were given a multi-dimensional scaling solution (using Systat) in order to map out the transcriber's intuitive tone transcription space. To give a non-linguistic (Ohio-centric) example of multi-dimensional scaling, if Cincinnati and Columbus are given a closeness rating of .5, and Columbus and Cleveland are given a closeness rating of .5, but Cincinnati and Cleveland are given a closeness rating of 1, then a multi-dimensional scaling solution will give Cartesian coordinates in which Columbus is plotted half-way between Cincinnati and Cleveland. In this example, it would give coordinates in just one dimension (a line running southwest-to-northeast on the map), but the example can be extended to multiple dimensions. For example, if Dayton were included, then two-dimensional coordinates would be required to place it west of Columbus but not along the diagonal (southwest-to-northeast) line formed by Cincinnati, Columbus, and Cleveland. The plot with the points at particular X-Y coordinates is the type of mapping that will be the result of the multi-dimensional scaling. The solution gives each category X-coordinates (for a one-dimensional solution), X- and Y-coordinates (for a two-dimensional solution), X-, Y-, and Z-coordinates (for a three-dimensional solution), and so

on into multiple dimensions. In this case, the transcriber's tone similarity judgements are used as input, with the pitch accent types as the categories in the first case and phrase accents / boundary tones being the categories in the second case. One and two dimensional maps will be explored, to see whether the one- or two-dimensional map accounts for most of the variance while at the same time having dimensions which are interpretable and in accord with transcribers' intuitions. Thus, the results of the multi-dimensional scaling will be Cartesian coordinates for each category, mapping out how similar to and different from each other the categories are according to how close to or far away from each other they lie on the line or in the plane. If it turns out that there are clusters of tones on the tone map, then the tones in that cluster can be interpreted as being similar, and hence easily confusable in the transcriber's tone space. Given the confusability of these tones with each other, it may be inferred that they have the same tone targets or that they are in the same class of tones.

(Note that this might be a useful tool in future research for measuring inter-transcriber agreement, as a way to map out different transcribers' tone spaces and to see how closely different transcribers' intuitions about similarities and differences match up with each other. This could then be used to search for patterned differences between transcribers, and may be a more sensitive measure of inter-transcriber differences and similarities than simply counting numbers of identical transcriptions. This method could also prove important in the investigation of intonational categories, by providing first clues as to tone groups which are confusable by transcribers, and hence which might need investigation as to whether they are confusable by listeners in natural conversational settings as well. Further investigation of such classes and the similarity or difference in their meanings as well may prove interesting, if a reliable metric of similarity in the semantic or pragmatic domain can be found. For one such investigation, see Gussenhoven and Rietveld (1991).)

This method allows for a finer degree of similarity judgement between pairs of strings of tones, rather than just "identical" vs. "different." For example, in Pitrelli et al's (1994) analysis of inter-transcriber consistency, the authors compared the tone transcriptions of different transcribers by marking either "agreement" if two transcribers transcribed the same tone for the same word or else "disagreement" if that pair of transcribers transcribed different tones for the same word. Then, they counted up numbers of agreements and disagreements. However, it seems possible that transcribers might have consensus judgements as to the similarities and differences among tone categories, and how easily confusable they are with each other. (This is of course testable, using just such a method as was described above which would allow for comparison between transcribers' intuitions of similarity and difference.) This type of impression was translated into one of Pitrelli et al.'s methods in measuring agreement among transcribers, namely to relax the distinction between H* and !H* (to allow H* and !H* to be counted as "the same" category), since the largest disagreement "is ambiguity in application of the downstep diacritic." (p. 125) This means that if one transcriber marked H* for some word and a second transcriber marked !H* for that same word, they would be counted as an "agreement." This is exactly the type of judgement which is captured in the multi-dimensional scaling approach described above, quantifying this type of impression of tone similarity using transcriber's elicited tone similarity judgements, and multi-dimensional scaling to map out the transcriber's tone space.

There is also the potential for substituting the Cartesian coordinate values in for each transcribed tone and then calculating a Euclidean distance between strings of tones (with the Euclidean distance between two points being the distance apart from each other on the plane). This method of calculating the distances between strings of tone transcriptions would be more useful for an experimental design in which there were both cases in which the tunes were expected to be similar to each other (and so the Euclidean distance would be

expected to be quite small, because the coordinate values would be quite similar to each other and hence close to each other in the Cartesian plane) and cases in which the tunes were expected to be different from each other (and so the Euclidean distance would be expected to be larger, because the coordinate values would be far away from each other, and so have a greater distance between them). This method could potentially allow for quantification of the similarity between pairs of strings of transcribed tones, but will not be used here.

Thus, the method described above tests for tone similarity and allows for comparisons from within tonal categories while avoiding comparisons across different classes of tone categories by empirically defining "similar" tones. This method includes into the analysis the transcriber's intuitive tone space and experience of confusability of categories. Thus, the question of whether the "same" tones have been transcribed for the two utterances in a pair of utterances can be answered.

2.3 Perception Test

A perception test was run to test whether subjects could distinguish between sentences which were paragraph-final vs. sentences which were not paragraph-final. The stimuli in the perception test were sentences extracted from the recording.

2.3.1 Materials

The sentences which were used were extracted from two different positions in the dialogues, namely, final position in the dialogue vs. medial position in the dialogue. The entire dialogue had already been digitized and the target sentences had been extracted from the existing sound files and made into individual files. The target sentences were drawn from each of the 18 speakers, each with 2 sentences from each of 10 dialogues. This gives 360 sentences, each of which was paired up with the same sentence but read in a different discourse location, leading to 360 pairs of sentences.

Each pair was presented to listeners twice, once with the paragraph-final sentence first and once with the paragraph-medial sentence first, to control for effects of order, giving 720 pairs total. This was too much for any one listener to listen to and give judgements on. Therefore, these 720 sentences were split into 5 groups, each containing 144 sentences, which was manageable from the listeners' point of view both in terms of time needed for listening and in terms of attention span and ability to concentrate. The 144 sentences that each listener heard consisted of 2 sentences from one dialogue (the A and B sentences from dialogue 1, for example) and 2 sentences from another dialogue (the A and B sentences from dialogue 2, for example), each with its matched pair-mate from the other discourse location, giving 4 sentence pairs. These 4 sentence pairs had been read by 18 different speakers, giving 72 pairs. These 72 pairs were presented to the listeners in both orders, giving 144 sentence pairs. Thus, the within-subjects factor was the discourse location, while the between-subjects factor was the list of pairs used.

2.3.2 Subjects

The subjects in the perception test were Ohio State University undergraduates enrolled in either Linguistics 201, the undergraduate introduction to linguistics, or Linguistics 371, Language and the Mind. They were paid $10 for their participation in the experiment. The experiment took about 30 minutes. The listeners' date of birth, sex, and places lived are listed in the table.

Subject #	Date of Birth	Sex	Places Lived
L1	10/30/76	F	0-18 Defiance, OH 18-21 Columbus, OH
L2	6/22/78	M	0-18 Northbrook, IL 18-20 Columbus, OH
L3	3/27/78	F	0-18 Van Wert, OH 18-20 Columbus, OH
L4	1/21/79	F	0-3 Bath, England 3-19 Pittsburgh, PA
L5	6/15/79	F	0-1 Findley, OH 1-18 Lancaster, OH 18 Columbus, OH
L6	1/28/76	F	0-19 Lexington, OH 19-22 Columbus, OH
L7	11/3/78	M	0-18 Akron, OH 18-19 Columbus, OH
L8	1/12/56	F	0-19 Cincinnati, OH 19-42 Columbus, OH
L9	1/9/75	M	0-18 Chicago, IL 18-23 Columbus, OH
L10	5/31/77	F	0-2 Cincinnati, OH 2-4 Rockford, IL 4-6 Fort Wayne, IN 6-10 York, PA 10-20 Canfield, OH
L11	7/26/74	F	0-5 Lakewood, OH 5-17 Steubenville, OH 17-24 Columbus, OH
L12	10/25/74	M	0-5 Steubenville, OH 5-6 Tiffin, OH 6-11 Lima, OH 11-18 Canton, OH 18-20 Toledo, OH 20-23 Columbus, OH
L13	8/28/78	M	0-17 Toledo, OH 17-19 Columbus, OH
L14	3/31/75	F	0-18 Youngstown, OH 18-23 Columbus, OH
L15	3/15/77	F	0-3 Jackson, MI 3-11 Knoxville, TN 11-12 Chattanooga, TN 12-21 Columbus, OH
L16	3/9/74	F	0-4 Oberlin, OH 5-21 South Bend, IN

(to be continued)

Figure 56. Subject information from perception experiment.

Figure 56, continued.

Subject #	Date of Birth	Sex	Places Lived
L17	10/8/73	F	0-20 Akron, OH 20-24 Columbus, OH
L18	5/1/76	F	0-19 Painesville, OH 19-22 Columbus, OH
L19	8/1/73	M	0-17 Columbus,OH 18 Orlando, FL 19-21 Cincinnati, OH 22-25 Columbus, OH
L20	8/16/74	F	0-22 Lorain, OH 22-24 Columbus, OH
L21	2/26/77	F	0-9 Sidney, OH 9-20 Maplewood, OH 20-21 Columbus, OH
L22	11/13/75	M	0-19 N. Canton, OH 19-22 Columbus, OH
L23	3/8/76	M	0-22 Ohio
L24	5/1/75	F	0-1 Arizona 1-21 Brunswick, OH 21-23 Columbus, OH
L25	4/22/70	M	0-18 Madison, OH 18-20 Columbus, OH
L26	12/24/76	F	0-19 Ottawa, OH 19-20 Dayton, OH 20-21 Columbus, OH

2.3.3 Methods

The pairs of sentences were played in randomized order to subjects over headphones. The pairs were randomized across speakers, meaning that the listener could have heard a pair of utterances spoken by speaker 1 followed by a pair of utterances spoken by speaker 2, or else a pair of utterances spoken by speaker 2 followed by a pair of utterances spoken by speaker 1, or in fact any order of speakers (but with the two utterances within a pair being spoken by the same speaker). The subjects were seated in a double-walled sound-proof booth, were presented with the stimuli over headphones, and indicated their answer on the computer by clicking with the mouse on one of the boxes which appeared on the screen. The question which appeared on the screen was “Which of

these sentences was paragraph-final?" The possible answers in the boxes were "1st," "2nd," or "listen again." Thus, this was a forced-choice test, with the option to listen to each pair again any number of times after the first presentation. The pace was self-selected. On the instruction sheet, they were shown the two possible contexts from which the sentence had been extracted. The instructions given to the listeners were as follows, with each instruction sheet containing two of the dialogues. (The one shown here contains dialogues one and two.)

Instructions:

You will be hearing 144 pairs of sentences. The sentences which you will be hearing have been extracted from one of the following four contexts:

context 1:
Are you going to need to change advisors?

I was talking to my advisor the other day.
She said she's going to be doing field research in Kenya.
It's o.k. to have an advisor who's away, but only for one quarter.
She'll be gone for the whole year.

context 2:
Are you going to need to change advisors?

I was talking to my advisor the other day.
She said she's going to be doing field research in Kenya.
She'll be gone for the whole year.
It's o.k. to have an advisor who's away, but only for one quarter.

context 3:
I know that Sue wanted to donate blood. Did she qualify as a donor?

If you're over 110 pounds, between ages 17 and 65, and haven't taken any medication for two weeks then you can donate blood.
I think she weighs about 130 pounds.
She turned 32 last month.
She hasn't taken any medication since that sinus infection last February.

context 4:
I know that Sue wanted to donate blood. Did she qualify as a donor?

If you're over 110 pounds, haven't taken any medication for two weeks, and are between ages 17 and 65 then you can donate blood.
I think she weighs about 130 pounds.
She hasn't taken any medication since that sinus infection last February.
She turned 32 last month.

(to be continued)

Figure 57. Sample instruction sheet.

Figure 57, continued

In half of the sentences, the extracted sentence you will be hearing was **not** read as the final sentence in the paragraph.
In the other half, the extracted sentence **was** read as the final sentence in the paragraph.

Your task is to listen to each pair of sentences over the headphones and then **decide which sentence** (the first one you heard or the second one you heard) **is the sentence which was paragraph-final**. Use the mouse to click on the appropriate box on the screen, and the next pair will begin immediately after that.

The trial number will appear at the top of the screen.

If you wish to listen to the pair of sentences again, click on the box labelled "listen again." However, remember that your first impression is often the most accurate one.

If you have any questions, please ask.

Thank you!

The reading of the instruction sheet was followed by a quick verbal check to make sure that the subjects understood the task.

After data collection, the listeners' answers were marked as either correct or incorrect, and the number of correct identifications was tallied.

Each pair of discourse-medial vs. discourse-final sentences was heard by 5 listeners. Each listener heard each pair twice, to counter-balance for order. This results in 10 judgements per matched pair of sentences. The 95% confidence interval around 5 is at 3.42 and 6.58. Thus, 7 or more correct identifications give 95% confidence that the listeners were correctly identifying the discourse position and were not just guessing, and 3 or less correct identifications implies a 95% confidence that they were reliably getting the reverse of the actual identifications.

The pairs had been presented in both orders, making it possible to test for bias of presentation order. The proportions of correct responses were calculated by subjects and

(in a separate calculation) by items for paragraph-final-sentence-presented-second vs. paragraph-final-sentence-presented-first. These proportions were transformed using the arcsine transformation. A paired t-test was performed for the pairs of transformed proportions. If significant differences are found, then that means that there was a bias by order of presentation.

2.4 Acoustic Analysis

The acoustic factors under consideration as potentially providing differences between pairs of utterances were RMS (root mean squared) amplitude, fundamental frequency (F0), and duration. F0 is standardly used as an acoustic factor in intonational studies. RMS amplitude was included here as a potential acoustic factor because of the implications of a suggestion by Herman et al. (1996) that "final lowering" involves not just a compression of the range of F0 but also a "final fade"– a decrease in volume governed by a decrease in subglottal pressure. This finding implies that what is being controlled in the use of the phonetic implementations of tones to signal discourse functions is not just an F0 target or even an F0 range (if those are defined purely by the F0 level), but rather an overall "vocal effort." Changes in "vocal effort" might involve a general control strategy which includes a change in subglottal pressure, a change in the tension or relaxation of the muscles in the larynx, and a general change in the effort involved in the articulation. This general control strategy of changing the amount of "vocal effort" has several acoustic results, including changes in energy in the signal (reflected in changes in RMS amplitude), changes in F0, changes in duration, and changes in voice quality (for a discussion of phrase-final lengthening as resulting from a change in the stiffness of gestures, see Beckman, Edwards, and Fletcher, 1992). Thus, although what is being examined here is the implementation of tones, this is not to imply that F0 alone is serving as a cue. Rather, since the RMS amplitude, the F0, and the duration are all hypothesized to be aspects of the acoustic signal which result from a general control strategy in the articulation (namely, a

change in the “vocal effort”), that means that all three of these cues will be examined here in the investigation of the phonetic implementation of tones.

F0 and RMS amplitude are both calculated from the waveform, which is the record of the acoustic energy in the signal. A sample waveform is shown in the bottom panel of Figure 58, with time on the x-axis and amplitude on the y-axis. The RMS amplitude can be calculated by taking the amplitude over a small section, squaring it, taking the mean, and taking the square root, resulting in all positive values which indicate when the energy is greater and when the energy is less. The RMS amplitude curve corresponding to the waveform is shown in the middle panel. The RMS amplitude curves give a very rough estimation of loudness (rough because psycho-acoustic measures indicate that amplitudes at certain frequencies contribute more to loudness than amplitudes at other frequencies).

The fundamental frequency reflects the rate of vibration of the vocal folds. Each glottal pulse has a period, whose inverse is the frequency. The period of the glottal pulses is found by an autocorrelation method, meaning that a small section of the waveform is taken and correlated with a small section of the waveform further along in time. This is done for several incremental steps forward in time, and when the correlation is at its highest, then that means that the two adjacent glottal pulses match in shape, so a complete glottal pulse has been found, and the space of time between the two pulses is the period (whose inverse is the F0). The F0 is roughly equivalent to pitch, especially in the frequency region in which the F0 of the voice for most speech occurs. The RMS amplitude curves and F0 traces were calculated using Waves™ get-f0 program, which is an auto-correlation method of F0 estimation which uses dynamic programming to adjust to local intensity variations and F0 trends. In some cases, at the end of the utterance particularly, changes in voice quality cause problems for the F0 tracking because the F0 at such points is not well-defined. In such cases, there can be “pitch-halving,” in which the calculated F0 value is half of what it should be. The second syllable in the example in Figure 58 is

actually not such an example, as can be calculated by expanding the waveform to make individal glottal pulses identifiable, finding the glottal pulses in the waveform by eye, measuring the period between glottal pulses, and calculating the frequency (F0) by 1/period. The problems of changes in voice quality will be discussed further in 4.2.4

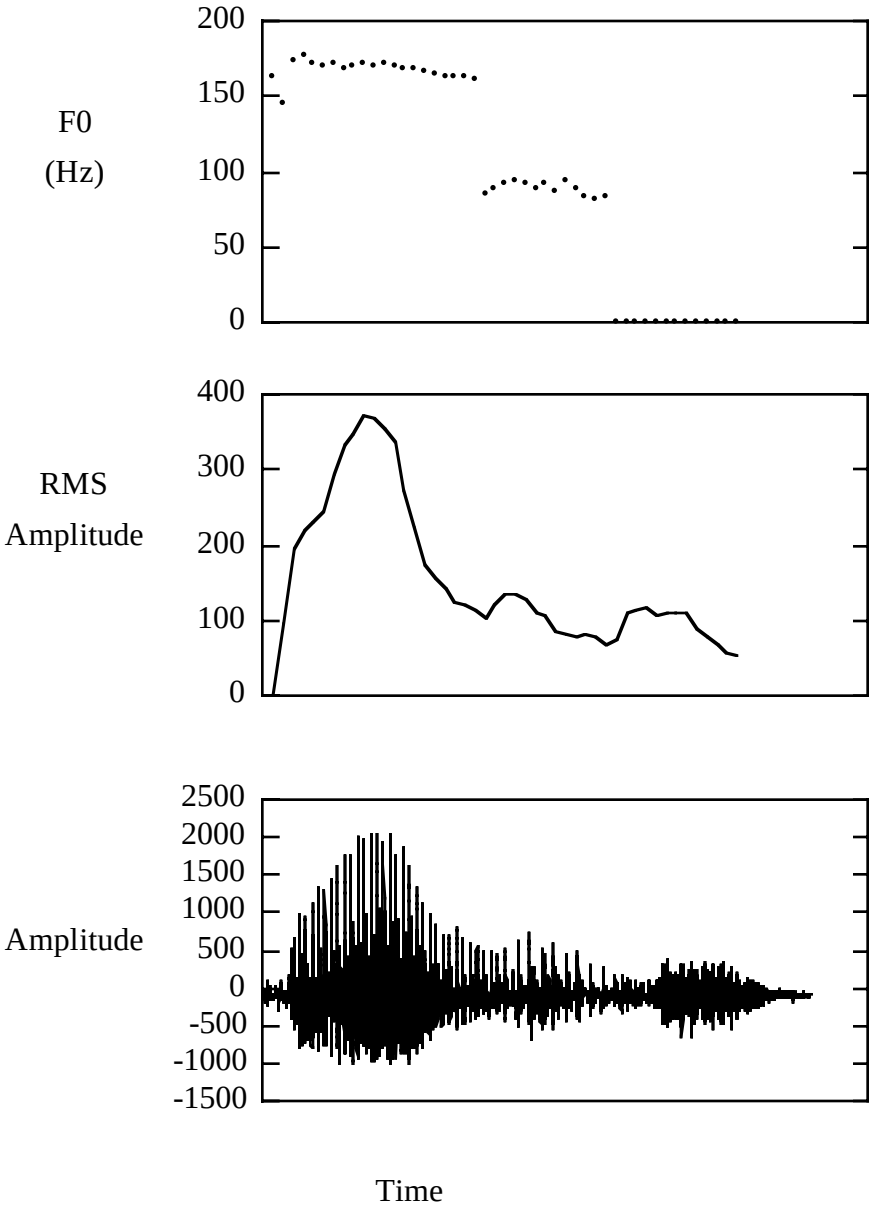


Figure 58. Sample waveform, RMS amplitude, and F0 of the word “dollars” extracted from the sentence “They raised the rent by forty dollars.”

A comparison between tokens was done by measuring the RMS amplitude values at the syllable peak of pitch accented syllables from both discourse-medial and discourse-final

tokens, and comparing the two. As mentioned earlier, the use of a head-mounted microphone keeping the microphone a constant distance away from the speaker's mouth allowed for comparison of RMS amplitude between utterances. The RMS amplitude values were measured at the peak RMS amplitude point in the target (pitch accented) syllable. overlaying. Similarly, a comparison of the F0 values was done by measuring the F0 at the syllable peak of pitch accented syllables from discourse-medial and discourse-final utterances. The F0 values were measured at the the syllable peak (the highest point in the RMS amplitude) for the target syllable rather than at the F0 peak for the target syllable because peaks in the F0 were not always apparent, especially after the first pitch accented syllable in the sentence. Although this did not allow for good comparison of the first pitch accented syllable in the utterance, since there the F0 peak is pronounced in the F0 trace and sometimes does not match up with the syllable peak as defined by the RMS amplitude maximum, it did allow for a constant point of measurement in many of the later syllables, in which the F0 formed more of a plateau over the syllable than a peak. (The defintion of the point at which to measure F0 is another area requiring further research, both from production and perception.) The durations were compared by measuring the duration from the onset of the last pitch accented syllable (at the beginning of the first consonant in the syllable) until the end of the utterance, and comparing the measurements from the discourse-final with the measurements from discourse-medial tokens. Although it is not clear exactly where the final lengthening occurs, if there is final lengthening present, it should occur within that span of syallables. All of these measurements were plotted in composite graphs showing RMS amplitude peaks for pitch accented syllables, F0 at syllable peaks for pitch accented syllables, and the durations of the final portion of the utterance. The differences between the measurements from discourse-medial and discourse-final tokens were also calculated and plotted. These differences were found by subtracting the value for the discourse-final token from the value for the discourse-medial token. From

these composite graphs, the contributions of each of the three factors to the perception of discourse-position of the utterance can be seen.

If the hypothesis that all three of these acoustic factors are the results of a general control strategy of changing the amount of "vocal effort" is true, then that would predict that all three of these factors should co-vary. The exact contribution of each of these factors as a perceptual cue to discourse position would require further investigation, perhaps by synthetic manipulation of each factor.

In sum, the acoustic measurements shown for each pair of tokens are the RMS amplitude peaks from pitch accented syllables, the F0 at the syllable peaks from pitch accented syllables, and the duration from the onset of the last pitch accented syllable until the end of the utterance.

2.5 Summary

There were several steps in the experimental investigation. First, the corpus was designed so as to control for dominance relationships. Then, the materials were recorded by 18 native speakers. After that, the target sentences were extracted from the recording and digitized and F0 traces and RMS contours were calculated. The tones in each of the target sentences were ToBI transcribed by an experienced ToBI labeler. The transcription allowed for the testing of the "same edge tones" hypothesis. Furthermore, the labeler was asked to provide tone similarity judgements, which were used to calculate multi-dimensional scaling solutions and hence to refine the notion of "same category with potential confusability with another category" vs. "different phonological categories with little potential for confusability." This ensured that any phonetic differences found in testing the "phonetic differences hypothesis" were not the result of phonological category differences. Then, perception tests were run using as stimuli pairs of sentences extracted from different discourse positions, which allowed for testing of whether any phonetic differences which were found in support of the "phonetic differences hypothesis" were

perceptible to listeners. And finally, an acoustic analysis using the results of the perception test and the tone similarity analysis was done by examining pairs of sentences which were reliably and correctly identified by subjects as to their discourse position (as well as for pairs of sentences which were reliably mis-identified), which ended in L-L%, and which had similar strings of pitch accents, in order to make sure that in testing the "phonetic differences hypothesis" the tokens under consideration were perceptibly different but phonologically similar. (The reason that it was important to compare pitch accents, but not so crucial here to compare edge tones from earlier phrases in the utterance, was that the RMS amplitude and F0 measurements were taken at syllable peaks. Thus, it was important for the pitch accents– which are associated with particular syllables, in this case, the target syllables– to be the same or from similarity classes). Thus, the only difference between the tokens in a pair was acoustic factors, which would provide evidence for the "phonetic differences hypothesis."

CHAPTER 3

RESULTS

3.1 Results– Tone Transcriptions

3.1.1 Edge Tones

The first step in analyzing the phonological tones is to have them transcribed. The ToBI transcriptions for the entire corpus were examined in order to see what the edge tones were. Given the hypothesis that high edge tones mark that more focus spaces are to be pushed onto the focus stack before popping the current focus space, there were predicted to be only low edge tones at the end of the utterances under consideration, since they were sisters in the intentional hierarchy and did not dominate some other DSP.

The corpus consisted of 720 tokens, 360 from discourse-final position and 360 from discourse-medial position. Of those 720 tokens, there were 4 (0.56%) which had H-H% at the end of the utterance. Two of these were from discourse-medial position and two were from discourse-final position. Three of these were produced by the same subject.

Of the 720 tokens, there were 22 (3.06%) which had L-H% at the end of the utterance. Sixteen of these were from discourse-medial position and six of these were from discourse-final position. Five of these were from Dialogue 7. Of these five, four were sentence B. Six of them were from Dialogue 9. Of these six, four were sentence A and two were sentence B. Thus, several of the tokens with L-H% transcribed were the same

sentence type or from the same dialogue, perhaps indicating that the discourse structure in that dialogue as perceived by participants in the experiment did not match the intended discourse structure of the dialogue as planned by the experimenter.

Of the 720 tokens, there were 25 (3.47%) which had H-L% at the end of the utterance. Eighteen of these were from discourse-medial position and seven were from discourse-final position. Five of these were from Dialogue 2. Of these five, three were sentence A and two were sentence B. Eight of these were from Dialogue 8. Of these eight, 7 were sentence B. Again, the clustering of the majority of H-L% transcriptions within two particular dialogues may indicate that the perceived discourse structure for those dialogues did not match the intended discourse structure.

Thus, there were relatively few high edge tones compared to the number of low edge tones, as predicted. A summary of the results is presented in the table.

	discourse-medial		discourse-final	
H-H%	2	.6 %	2	.6%
L-H%	16	4%	6	1.7%
H-L%	18	5%	7	1.9%
L-L%	324	90%	345	96%

Figure 59. Number of edge tones of each type.

There were more high edge tones in discourse-medial position than in discourse-final position. The predicted frequency, assuming no difference between the discourse-medial and the discourse-final positions, can be calculated as the mean of the observed numbers found for discourse-medial and discourse-final for each category (thus leaving the same proportion of tokens in each category– discourse-medial and discourse-final).

	discourse-medial	discourse-final
H-H%	2	2
L-H%	11	11
H-L%	12	12
L-L%	334	334

Figure 60. Expected number of edge tones of each type, no difference by discourse position.

Since some of these cells contain less than 5 tokens, making a chi-squared test unreliable, the cells containing H% will be pooled together.

	discourse-medial	discourse-final
H-H% and L-H%	13	13
H-L%	12	12
L-L%	334.5	334.5

Figure 61. Expected number of edge tones by each type, pooling together H%

	discourse-medial	discourse-final
H-H% and L-H%	18	8
H-L%	18	7
L-L%	324	345

Figure 62. Observed number of edge tones by each type, pooling together H%.

The deviance can be computed as (observed - expected)^2 / expected. The table of deviances is given in Figure 63.

	discourse-medial	discourse-final
H-H% and L-H%	1.92	1.92
L-H%	3	3
L-L%	.33	.33

total deviance = 10.5

df = (number of columns - 1) * (number of rows - 1) = 2

Figure 63. Deviances: (observed - expected)^2 / expected

A deviance of 10.5 with 2 degrees of freedom is significant at the .01 level for the chi-squared distribution. Thus, it can be concluded that there was a significant difference in the distribution of H edge tones according to discourse position such that there were more H edge tones discourse-medially than discourse-finally.

There are predicted to be no high edge tones discourse-finally, but their existence may be a discourse marker for the interlocutor to give approval that the sentences were read correctly or were understood, thus still falling within the meaning of requiring interpretation with respect to a subsequent utterance (in this case, with respect to the speaker's interlocutor's back-channeling). Ayers (1994, p. 29) discusses the use of high edge tones in spontaneous speech as "grounding or checking to see if the listener understood." She further notes the absence of high edge tones in the same positions in read versions of the same narrative. The use of a high edge tone to seek confirmation from the interlocutor might thus still be seen as a reflection of the intentional structure of the discourse, in this case, the speaker's intentions to shape the discourse by turn-taking.

Discourse-medially, the presence of high edge tones would indicate that the utterance is to be interpreted with respect to a subsequent utterance. Alternatively, the use of a high edge tone might imply a different intentional hierarchy than was intended in the logical structure. That is, if the speaker interpreted two DSPs not as sisters but rather as

being in a dominance or satisfaction-precedence relation with each other, then that would explain the presence of a high edge tone at the end of the discourse medial utterance. This issue could be explored further by annotation of the discourse segments and discourse segment purposes, as described by Nakatani et al. (1995), in order to see how the speakers were interpreting the intentional hierarchy, and whether the interpreted intentional hierarchy matches the intended intentional hierarchy or not.

In sum, the “same edge tones” hypothesis was tested, and it did turn out that the majority of cases had a low edge tone and that very few cases had a high edge tone, as predicted.

3.1.2 Tone Similarity Judgements

In order to measure the similarity of two tone transcriptions, the most stringent criterion is of course to determine whether they have identical sequences of tones. But this stringent criterion can be relaxed so as to admit tone similarity and confusability between categories. In order to do this, there are several steps involved. First, the transcriber’s tone similarity judgements are elicited. Second, a multi-dimensional scaling solution which maps out the transcriber’s tone similarity space is calculated. Then, the map of the tone space is interpreted, and groups of tones close to one another in the tone space are identified and understood to be easily confusable by the transcriber, and can be counted as the same class of tone targets.

Tone similarity judgements of the transcriber were elicited for every pair of pitch accents in the inventory and for every pair of phrase accents or phrase accents and boundary tones in the inventory. For example, the transcriber judged the similarity of H* to L+H*, H* to L*, H* to !H*, and so on for every possible pair in this group. Also, in a separate task, the transcriber judged the similarity of L- to H-, L- to !H-, L- to L-L%, L- to L-H%, and so on for every possible pair in this group. The transcriber was asked to judge similarity on a scale of 1 to 7, with 1 being most similar to each other, and 7 being most

different from each other. The transcriber saw each pair three times in random order, and responses were averaged over the 3 responses for each pair. For example, the pair H* vs. !H* received similarity judgements of 2, 3, and 3, resulting in a mean similarity judgement of 2.67. The results are as follows, separated out for pitch accents vs. phrase accents and boundary tones.

	unacc	H*	L+H*	L*	L*+H	H+!H*	!H*	L+!H*
H*	5.00							
L+H*	7.00	2						
L*	4.67	4.67	6.67					
L*+H	4.67	5.33	6.67	2.00				
H+!H*	3.67	3.00	4.00	3.33	3.33			
!H*	3.67	2.67	5.00	3.33	3.33	1.33		
L+!H*	3.67	2.67	3.67	3.33	3.66	1.33	2.00	
L*+!H	3.33	3.00	5.00	3.33	3.00	2.33	2.66	2.33

Figure 64. Transciber's tone similarity judgements for pairs of pitch accents.

	no break	L-	H-	!H-	L-L%	L-H%	H-H%	H-L%	!H-L%
L-	4.67								
H-	4.67	5							
!H-	4.00	2.33	3.33						
L-L%	7.00	3.67	5.00	3.67					
L-H%	7.00	4.67	4.33	3.67	2.67				
H-H%	7.00	5.33	2.67	3.33	7.00	4.00			
H-L%	7.00	4.33	4.00	4.67	4.33	4.67	3.67		
!H-L%	6.67	4.00	3.67	3.67	4.67	4.67	4.33	3.00	
!H-H%	6.67	5.00	2.67	2.33	5.33	4.00	3.33	3.67	2

Figure 65. Transcriber's tone similarity judgements for pairs of edge tones.

Next, the transcriber's tone similarity space was mapped out using these tone similarity judgements and the statistical technique of multi-dimensional scaling. The multi-

dimensional scaling solution for the pitch accents gives a one-dimensional solution with .88769 of the variance accounted for by the mapping. The map is shown below.

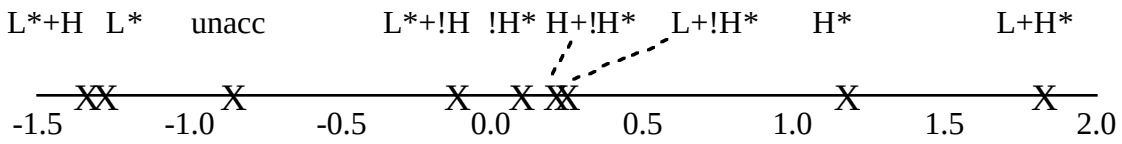


Figure 66. One-dimensional scaling solution– pitch accent tone similarity space.

Since there were no occurrences of L*+!H in the corpus and that tone seems to be transcribed rarely by this transcriber, perhaps indicating less certainty in her judgements of its similarity to other tones, the map can be re-drawn without that category, for the sake of simplifying the space.

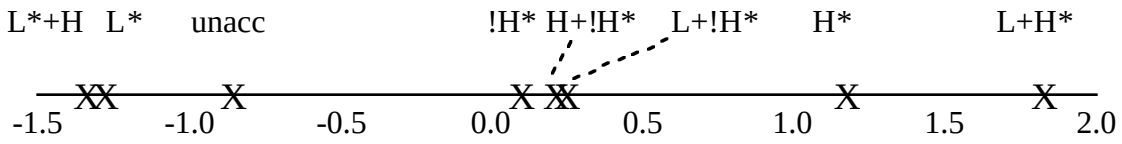


Figure 67. One-dimensional scaling solution- pitch accent tone similarity space, revised.

There are distinct groups of tones, when looking at the starred tones, which appear to be present in this one-dimensional map. For example, tones with L* are in a pair at one end of the map. “Unaccented” is alone as a category. The next group consists of tones with !H*. And the final group, or pair of single member groups, includes H* and L+H*.

Adding a second dimension in the multi-dimensional scaling analysis raises the amount of variance accounted for to .99500. The map is shown in Figure 68.

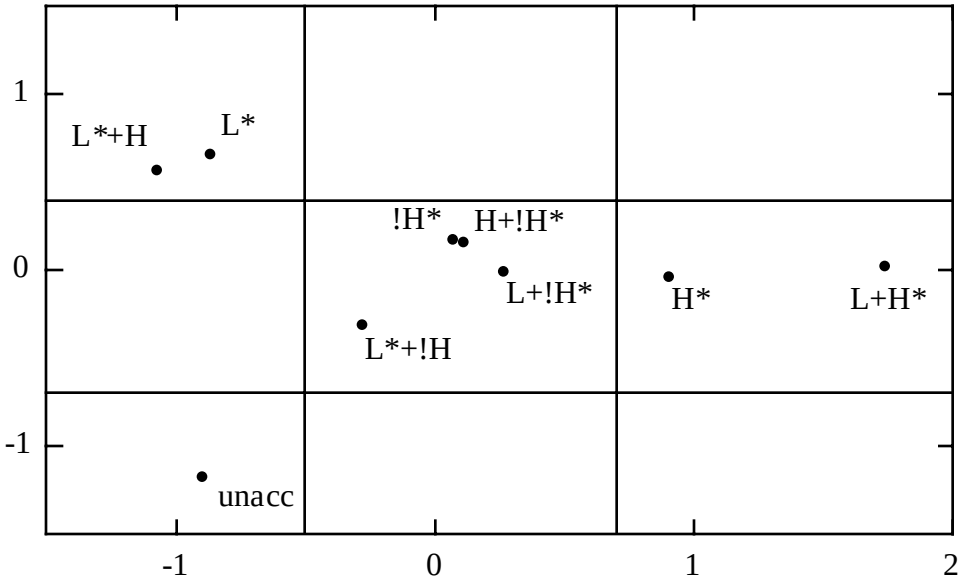


Figure 68. Two-dimensional scaling solution– pitch accent tone similarity space.

If the L*+!H is disregarded, since it did not occur in the corpus at all (and may not be used very often by this transcriber, indicating that she is less confident as to its identity with respect to other tones), as in the next figure, then the dimension marked by the "accented-unaccented" diagonal line corresponds to the height of the starred tone in the pitch accent, and the dimension marked by the "low-high" diagonal line mainly distinguishes unaccented syllables from syllables with a pitch accent of some sort. The diagonal lines in the figure (drawn in by approximation, not by a mathematical method) show these dimensions.

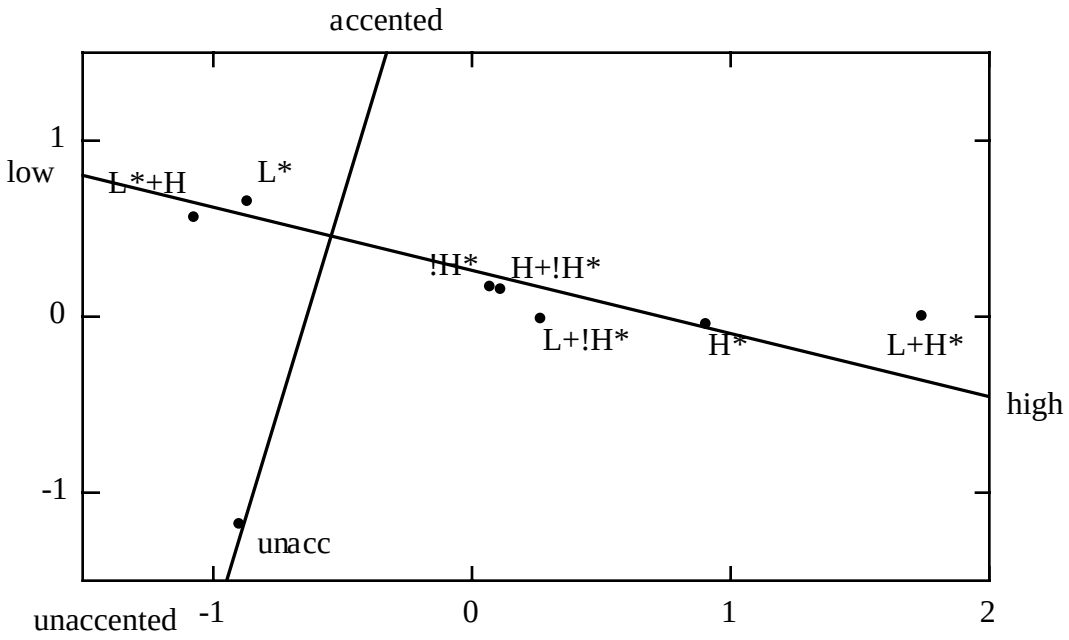


Figure 69. Two-dimensional scaling solution– pitch accent tone similarity space, revised.

Comparing the one-dimensional solution in Figure 67 and the two-dimensional solution in Figure 69, it seems that the one-dimensional solution in this case is adequate (it does account for about 89% of the variance), since it does express a scale which seems to correspond with the height of the starred tones and since the two-dimensional solution mainly pulls out the unaccented syllables from the pitch accented syllables. Thus, the second dimension might correspond to non-tonal cues to accentedness. Thus, the groups found in the one-dimensional solution will be the groups used as tone target categories, broken into groups of pitch accents containing L*, pitch accents containing !H*, pitch accents containing H*, and unaccented syllables. (Pooling together H* and L+H* is a somewhat arbitrary choice, since they are farther apart than L* and unaccented, which are put into two separate categories, but this grouping was done for the sake of simplicity, for the sake of having fewer classes, and because about a third of the cases which have similar pitch accents involve a H* in one token with a L+H* in its matched token. Also, this

grouping seems to capture the F0 level of the starred tone. On the other hand, putting unaccented in with the class of L* and L*+H would have completely eradicated the accented/unaccented distinction, hence those were not pooled together into one class.)

The phrase-accents and boundary tones have .69649 of the variance accounted for with a one-dimensional scaling solution. This single dimension has groups corresponding to degree of break, with no break at one end, then the three phrase accents, and then the pairs of phrase accents and boundary tones. Since phrase accents alone mark a lower level of intonational disjuncture (namely, an intermediate phrase break) while the boundary tones mark a higher level of intonational disjuncture (namely, an intonational phrase break), this map corresponds well with the degrees of intonational disjuncture marked by the different tone categories.

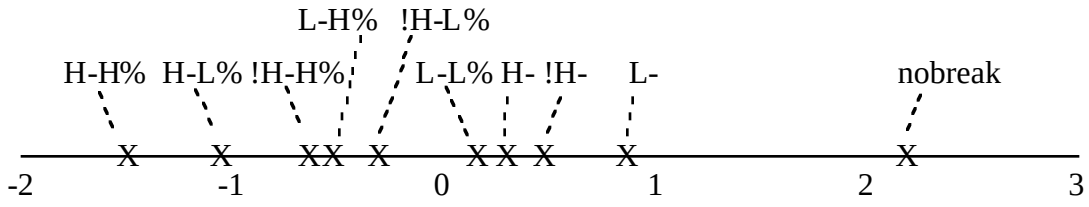


Figure 70. One-dimensional scaling solution– edge tone similarity space.

Adding a second dimension to the mapping increases the amount of variance which is accounted for to .92595. This solution is plotted below. In this case, the x-dimension again seems to correspond to degree of break, with no break at one end, phrase accents in one vertical cluster in the middle, and then the phrase accents and boundary tones at the other end. In this case, though, the y-dimension seems to correspond to L vs. H phrase accent, with L- clustering at the bottom of the plot, H- clustering at the top, and no break in the middle. These groups are demonstrated with vertical and horizontal lines partitioning

the space into areas, with the vertical lines partitioning no break, weak break, and strong break and the horizontal lines partitioning low phrase accent and high phrase accent.

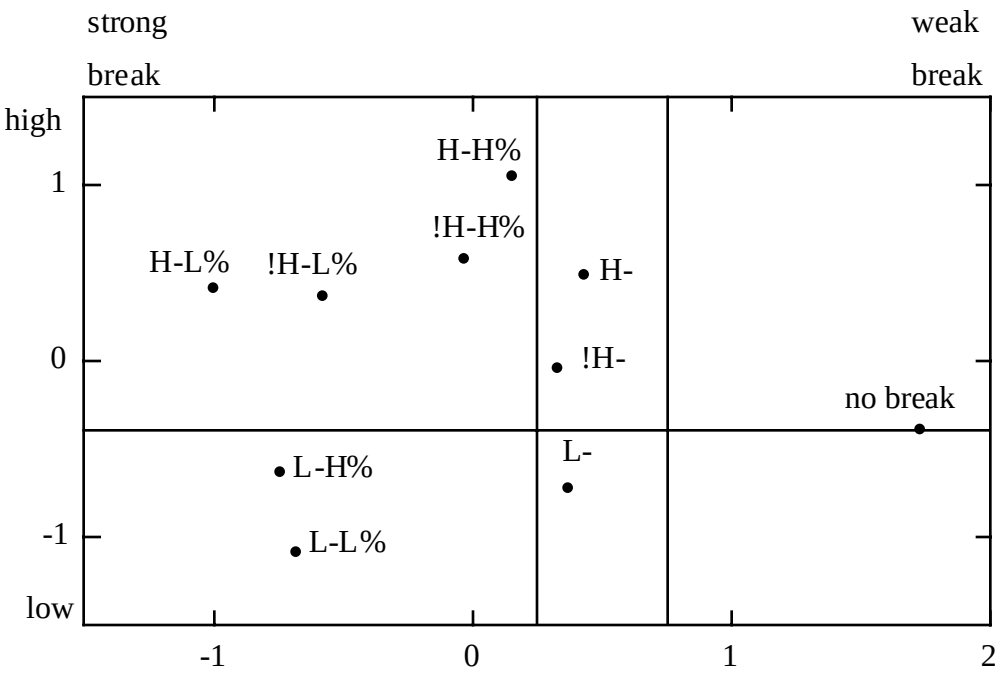


Figure 71. Two-dimensional scaling solution– edge tone similarity space.

Choosing between the one-dimensional and the two-dimensional map requires several criteria. One criterion is simply the amount of variance accounted for. The one dimensional map accounts for about 70% of the variance and the two-dimensional map accounts for about 93% of the variance. Another criterion is whether the dimensions are interpretable in a straightforward fashion. The two dimensions are interpretable as degree of break and height. Thus, since the two-dimensional solution accounts for much of the variance and can be broken into interpretable groups, the two-dimensional solution can be chosen as the best one (although it will not actually be used in the analysis).

The groupings of tones into similar tone targets will be used in the acoustic analysis. In the acoustic analysis, all of the phrases under consideration (at the ends of discourse segments) are full intonational phrases, so the x-dimension in the map above is not relevant. Furthermore, as discussed in 3.1.1, the predictions about the discourse functions of high edge tones need to be tested with respect to the presence of either a H- or a H%, or both. Thus, looking only at cases with L-L% (rather than some class containing L-, as suggested by the multi-dimensional scaling map) still leaves over 300 pairs suitable for further study. Thus, the multi-dimensional scaling solution for the edge tones will not be used in the acoustic analysis. Rather, the main purpose of the multi-dimensional scaling analysis is to avoid the confound of comparing different classes of pitch accents within the final and pre-final intermediate phrases since measurements are taken at syllable peaks in pitch accented syllables, defined as the peak in RMS amplitude (while avoiding the overly restrictive method of comparing only identical tone transcriptions). Thus, the one-dimensional scaling solution for the pitch-accents is the one map which will be used in comparing tones for similarity and confusability when selecting tokens for acoustic comparisons.

In sum, a method for defining tone similarity classes in order to avoid comparisons of phonologically different tones which are not confusable with each other was implemented, leading to the postulation of four groups of tone similarity classes for the pitch accents. This will be used further in testing the "phonetic differences hypothesis," by ensuring that the differences are phonetic and not phonological.

3.2 Results– Perception Test

In order to describe the results of the perception test, first the listeners' performance in general is presented by showing the proportion of tokens correctly identified by each listener. Then, the number of correct judgements per token necessary for statistical significance is presented. Then, the issue of bias in identifications according to order of

presentation of the stimuli is discussed, and based on these results, a more stringent criterion for selecting tokens for consideration is presented. Utterances which had a high edge tone are examined to see whether they were identified correctly and are then weeded out of the group which will be analyzed acoustically. The testing of the remaining selected tokens for tone similarity as an extension of the method in Section 3.1.2 is then discussed, which leaves a group of tokens to be submitted to acoustic analysis which were correctly identified by listeners and which had similar phonological tones.

3.2.1 Reliability of Listener Judgements

The numbers of correct judgements for each listener are given below. L23 was excluded from further investigation because this subject answered “first” for about the last 40 tokens, and expressed frustration at the end of the experiment, so was clearly not performing the task.

Listener #	# of Pairs Judged Correctly (out of 144 possible)	% correct
L1	69	.479
L2	92	.639
L3	84	.583
L4	92	.639
L5	79	.549
L6	89	.618
L7	91	.632
L8	83	.576
L9	104	.722
L10	95	.660
L11	87	.604
L12	98	.681
L13	85	.590
L14	97	.674
L15	79	.549
L16	118	.819
L17	107	.743
L18	65	.451
L19	85	.590
L20	71	.493
L21	71	.493
L22	81	.563
L23	71	.493
L24	102	.708
L25	84	.583
L26	94	.653
total:	2273	2273 / 3744 = .607

mean = total / n = 2273/26 = 87.42

Figure 72. Number of correct identifications by listener.

Thus, listeners identified about 61% of the tokens correctly. Since this mean score is not very high, it might be better to investigate further token-by-token to see which tokens were identified at better than chance and which were not, or if there was a bias toward identifying the utterance heard second as paragraph-final, and so how many tokens were actually identified correctly both times that the listener heard them.

Recall from the methods chapter that there needed to be 7 or more correct identifications for a token in order for the identifications to be at better than chance. However, it is also necessary to test for bias in order of presentation. That is, were listeners more likely to identify a sentence as "paragraph-final" when they heard it second in a pair of sentences? In order to test for bias, listeners' responses in the two orders of presentation were examined. Listeners heard each pair of sentences in both orders. Listeners' answers of "first" or "second" in response to "Which sentence was paragraph-final?" were marked as correct (1) vs. incorrect (0). The proportion of correct responses was calculated by subject and (in a separate calculation) by item, for cases in which the paragraph-final sentence was heard first vs. cases in which the paragraph-final sentence was heard second. These proportions were transformed using the arcsine transformation ($\sin^{-1}$). The transformed proportions were submitted to a paired t-test in pairs of final-medial order vs. medial-final order. In the by-subject condition, the means were different in the paired test at $p < .05$ (with $t=-2.0408$ and $df=24$). (The more conservative Wilcoxon sign test had $p < .07$.) Thus, the difference between the two orders of presentation was marginally significant by subject. In the by-item condition, the means were different in the paired test at $p < .008$ (with $t=-2.666$ and $df=359$). (The more conservative Wilcoxon sign test had $p < .01$.) Thus, the difference between the two orders of presentation was significant by item. As for the direction of the difference, averaging over all of the tokens and all of the listeners, when the paragraph-final sentence was heard second, the listeners (correctly) chose "second" as their response 1143 times out of 1800 (64%) while when the paragraph-final sentence was heard first, the listeners (correctly) chose "first" as their response 1059 times out of 1800 (59%). In sum, both the by-subject condition and the by-item condition showed significant differences by order of presentation, meaning that there were more correct identifications when the paragraph-final sentence was heard second in the pair than when it was heard first in the pair. The results of both t-tests indicate a bias

towards identifying the second sentence heard in the pair as paragraph-final, which seems to hold marginally across all subjects and to hold more strongly across all tokens.

Thus, it does not seem adequate to simply take the tokens which had 7 or more correct judgements (which is the 95% confidence interval that the identification is better than chance), since 7 correct judgements could mean 5 correct when the sentence was heard second and only 2 correct when the sentence was heard first, which would mean that only 2 of the listeners were correctly identifying the sentences both times they heard it, with the other 3 listeners identifying it correctly in one order of presentation and incorrectly in the other order of presentation. As a more stringent criterion and as a way to consider listeners' confidence, it was decided to focus on pairs of sentences which were correctly identified by a listener in both orders of presentation. In order to do this, pairs of sentences correctly identified 3 or more times (that is, by 3 or more listeners) in both orders of presentation were identified.

There were 1800 pairs of identifications (from 18 speakers, 20 pairs of discourse-medial vs. discourse-final sentences, 5 judgements per both orders of each pair). Of these 1800 pairs, 716 pairs in total from the perception test were correctly identified in both orders of presentation. When split into sentence types by speaker (for example, when tallying the number of pairs correctly identified in both orders just for Dialogue 1, Sentence A, as uttered by speaker S1), there are 117 sentence types by speaker (out of 360, from 18 speakers * 20 sentences) which were correctly identified in both orders by 3 or more listeners out of 5, so about one third of the pairs were correctly and reliably identified by subjects in the perception test.

In sum, in testing the "phonetic differences hypothesis," the set of tokens which were perceptibly different according to discourse position was found using a perception test involving pairs of utterances from two discourse positions.

3.2.2 Examination of Utterances with High Edge Tones

Only the 117 pairs of discourse-final vs. discourse-medial sentences which were correctly identified by 3 or more listeners out of 5 when heard in both orders of presentation were examined. These pairs of sentences are reliably identifiable as discourse-medial vs. discourse-final.

First, the utterances which had either H- or H% were examined to see which of those were correctly identified. Fifty-one sentences had H- or H% (which can be calculated by adding up all of the tokens in the table in 3.1.1 except the tokens which had L-L%). Of these 51, 17 had been correctly and reliably identified. Of these 17, there were 2 pairs of corresponding utterances from discourse-medial position and discourse-final position which both contained a high edge tone. In one of the pairs, both members of the pair had H-H% and in the other of the pairs, both members of the pair had L-H%. Thus, there were 15 pairs of sentences which had either a H- or a H% (or both) in one or both members of the pair. This means that there were 13 pairs of sentences in which it might have been the choice of phonological edge tone which was cuing the listener as to whether the utterance was “paragraph-final” or not (because in the two pairs which both had the same edge tones, the listeners could not have been using the edge tones as a cue to differentiate between the two sentences in the pair because the edge tones were the same). Of these 13 pairs, there were two cases in which it was the discourse-final utterance which had a high edge tone (which is completely unexpected according to the predictions about the meanings of the tones). This means that there were 11 pairs out of the 117 pairs which were correctly and reliably identified in which the tone choice might have been the cue to identifying the discourse-medial sentence that listeners were using, rather than the phonetic realization of the tone.

Thus, there were 102 pairs out of the 117 in which both utterances had L-L% and hence must have been identified as to their discourse location by some cue other than edge

tones. On the other hand, there were only 11 pairs out of the 117 in which the discourse-medial utterance could have been identified as such by the presence of a high edge tone at the end of the utterance. (Note that although we have postulated that a high edge tone means that the utterance should be interpreted with respect to a subsequent utterance, this also implies that when the listener hears a high edge tone, the listener knows that– unless the high edge tone is signalling a request for confirmation from the hearer– this is most likely a discourse-medial token since there must be a subsequent utterance and so the discourse cannot be done yet.)

The 15 pairs of sentences which contained either a H- or a H% (or both) were discarded from the acoustic analysis. This leaves 102 pairs of utterances which were reliably and correctly identified by listeners and which had L-L% at the end of both utterances.

In sum, the set of perceptibly different pairs was narrowed down further to include only perceptibly different pairs which had low edge tones, so as to ensure that the acoustic cue used by the listeners to identify the discourse position was not the choice of edge tone but rather a phonetic difference.

3.2.3 Testing Selected Tokens for Phonological Tone Similarity

Of the 102 pairs of utterances which were reliably and correctly identified by listeners and which had L-L% at the end of both utterances, there were 13 pairs which had entirely identical sequences of tones, looking at both the pitch accents and the phrase accents. These were examined in the acoustic analysis, but in a separate group. This left 89 pairs of utterances which were reliably and correctly identified, which ended in L-L%, but did not have identical sequences of tones.

These 89 pairs of utterances were examined to see which of them had pitch accent sequences which all were in the same tone target group according to the tone similarity judgements and multi-dimensional scaling described in section 3.1.2. There were 18 such

pairs (in addition to the 13 with identical tones) in which every pair of pitch accents were either the same or else two members of a group of tone targets identified as similar by the tone similarity analysis. (There were four groups identified by the tone similarity analysis: containing a L*, unaccented, containing a !H*, or containing a H*.) This leaves 71 pairs (102 - 13 - 18 = 71) of sentences which were reliably and correctly identified, ended in L-L%, and did not have the same classes of tones in their transcriptions. In these cases, it may be a difference in pitch accent choice, pitch accent location, or phrasing differences which cued the listeners as to the discourse position. Since this study is more concerned with the sound structure and discourse functions of the edge tones, such differences and their functional marking of discourse position remain a subject for future research.

In sum, narrowing down the set of utterances under consideration to the set of pairs which were perceptibly different, had low edge tones, and had either identical pitch accents or else pitch accents from the same tone similarity class implies that the differences between the utterances in such a pair must have been phonetic differences. Thus, acoustic analysis of this small set of utterances can test the "phonetic differences hypothesis" and find phonetic correlates of discourse structure without the confound of phonological differences.

3.2.4 Reliable Mis-identifications

A similar procedure was performed to find the sentences which were reliably mis-identified in the perception test in order to see whether the phonetic cues which might be found to distinguish pairs of utterances in correctly identified pairs would be reversed in the mis-identified pairs. Again, there were 1800 pairs of identifications total from the perception test, with 314 of these being incorrectly identified in both orders of presentation by three or more listeners. When split into sentence types by speaker, there are 23 (out of 360) which were incorrectly identified in both orders of presentation by 3 or more speakers out of 5 . Looking at only those 23 pairs, there were 7 which had identical tones.

3.2.5 Summary

The acoustic analysis will concentrate on three groups of utterances. First, there are the 13 pairs of sentences correctly and reliably identified (with confidence as indicated by being identified correctly in both orders of presentation by 3 or more listeners out of 5) and which had identical sequences of tones. Second, there are the 18 pairs of sentences correctly and reliably identified which had similar tone targets in the pitch accents. And third, there are the 7 pairs of sentences which were incorrectly identified (by 3 or more listeners in both orders of presentation) and which had identical sequences of tones. Thus, testing the "phonetic differences hypothesis" will use these three sets of utterances.

3.3 Acoustic Analysis

The 13 pairs of sentences from discourse-medial vs. discourse-final position which were correctly identified, had L-L% at the end, and had identical sequences of tones were compared with each other within the pairs. The 18 pairs of sentences which were correctly identified, had L-L% at the end, and had similar sequences of tones were also compared with each other within the pairs.

The factors under consideration are fundamental frequency (F0), RMS amplitude, and duration. In order to compare discourse-medial and discourse-final sentences, the F0 trace from the discourse-medial utterance can be overlaid on the F0 trace from its corresponding discourse-final utterance and the same thing for the RMS amplitude curves. Also, the durations can be compared to each other. When the RMS amplitude and F0 curves are compared, they map onto each other quite closely, with some tendencies to divergence from complete overlap. These divergences from complete overlap would seem to be what is cuing the identification of one as discourse-final and the other as discourse-medial. The divergences are tendencies and not absolutes in that each pattern of divergence is not necessarily true of each sentence in the corpus. Furthermore, the divergences are tendencies in that the point at which the divergence begins varies from token to token. That

is, in some cases the RMS peaks match each other quite closely except in the final pitch accented syllable, in others they match each other quite closely except in the final intermediate phrase, and in still others they differ over the course of the entire utterance. In all of these cases, though, the direction of difference tends to be the same, with the discourse-medial token having a higher F0 or higher RMS amplitude than its corresponding discourse-final token. These tendencies will be discussed below.

Comparing the discourse-final RMS amplitude curve with the discourse-medial RMS amplitude curve in the token in Figure 73, it can be seen that they match up pretty well . The two traces for the example token here are aligned in time such that the peaks seem to line up the best, which was done by hand for the example figure. In the token in Figure 73, the divergence of the RMS amplitude peaks begins in the final intermediate phrase (marked by the solid vertical line in the figure), with the pitch accented syllables from the discourse-medial token having a higher peak than the pitch accented syllables from the discourse-final token.

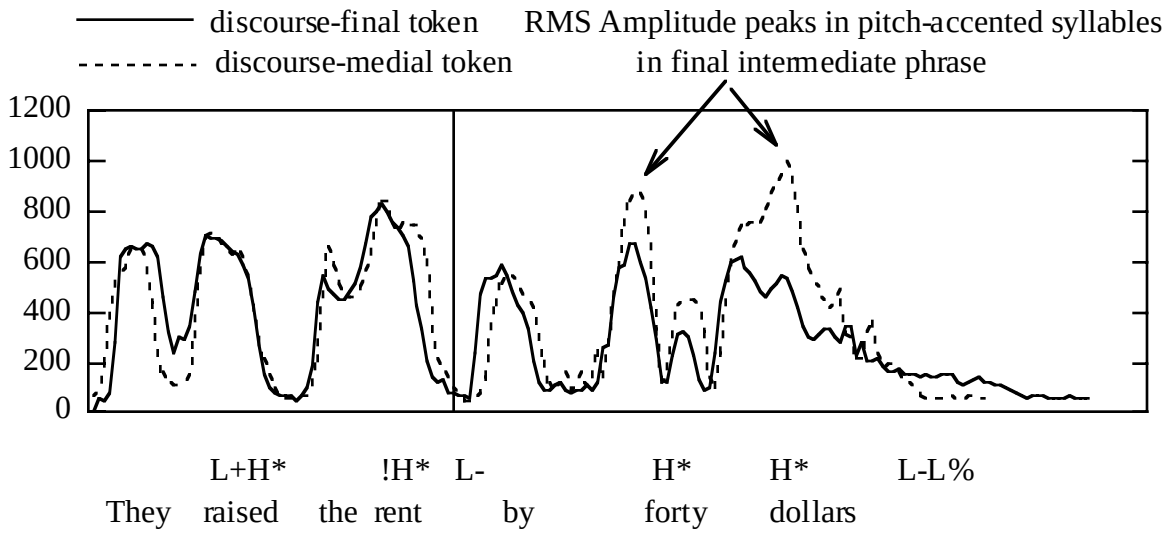


Figure 73. Example RMS amplitude traces– overlaid.

In order to make the data easier to compare, and since the peaks do not always line up in time, measurements are taken from the highest point in the pitch accented peaks for both tokens in a pair, and those two corresponding points are compared to each other. Thus, the alignment is done linguistically, not in real time, by finding the same points (the RMS amplitude peak on the pitch accented syllable) for both utterances and comparing those two points.

Comparing the discourse-final F0 trace with the discourse-medial F0 trace, it can be seen that they too match up pretty well. In the token in Figure 74, the divergence between the F0 traces is in the final pitch accented syllable.

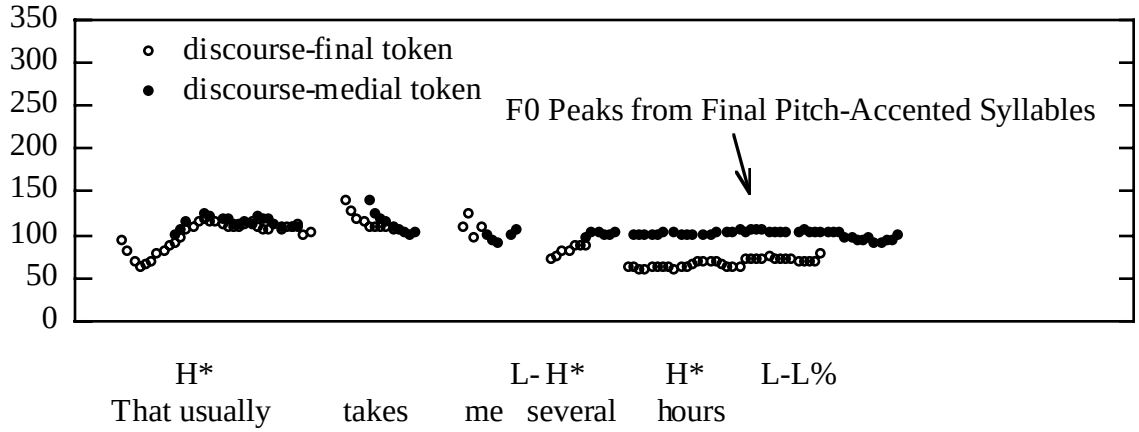


Figure 74. Example F0 traces– overlaid.

In order to make the data easier to compare, the F0 from the syllable peak (defined as the RMS amplitude maximum in that syllable) is measured, and the two corresponding points from the discourse-medial and discourse-final tokens are compared with each other.

The data will be presented in composite figures, as illustrated in the example in Figure 75. This example has all of the axes and symbols labelled, but the rest of the examples do not. All of the examples follow the same model. Each figure includes measurements from a pair of utterances, one from discourse-medial position and one from discourse-final position. The subject number is at the top left. The small graph at the top right plots the duration measurements in seconds. The bar on the left is the measurement for the discourse-medial token and the bar on the right is the measurement for the discourse-final token. The larger panel at the bottom includes both RMS amplitude measurements and F0 measurements. The RMS amplitude axis is the y-axis at the left (in arbitrary units), and the RMS amplitude measurements are connected with the solid line. The discourse-medial token is symbolized with an “m” and the discourse-final token with an “f.” The F0 axis is the y-axis at the right (in Hz), and the F0 measurements are

connected with the solid line. Again, discourse-medial is symbolized with “m” and discourse-final with “f.” The x-axis is time, but measured in terms of the pitch accents and edge tones in the utterance. The ToBI transcription is given for each token, with the pitch accent labels aligned with the peak measurements that they correspond to. The top panel plots the difference measures between the discourse-medial and discourse-final measurements for both RMS amplitude and F0, using the same y-axis for both (so the y-axis is either arbitrary units for the RMS amplitude differences or in Hz for the F0 differences). The difference between the RMS amplitude measures at each pitch accent is plotted with the white bars on the left, and the difference between the F0 measures at each pitch accent is plotted with the shaded bars on the right.

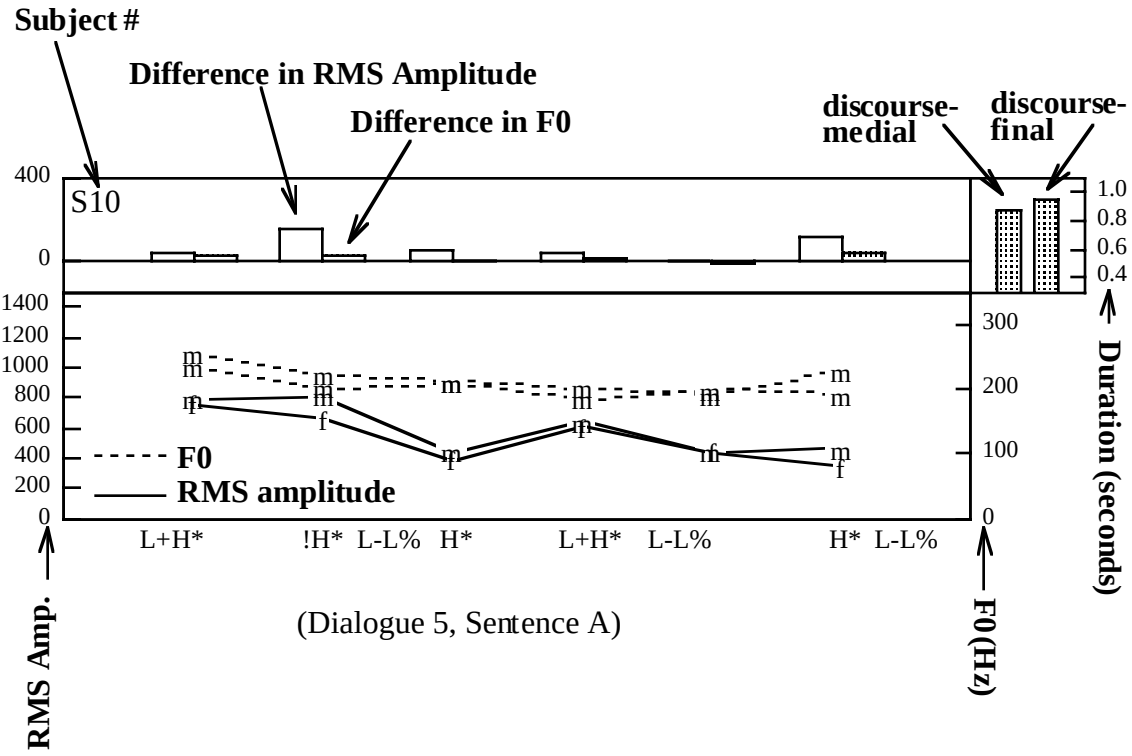
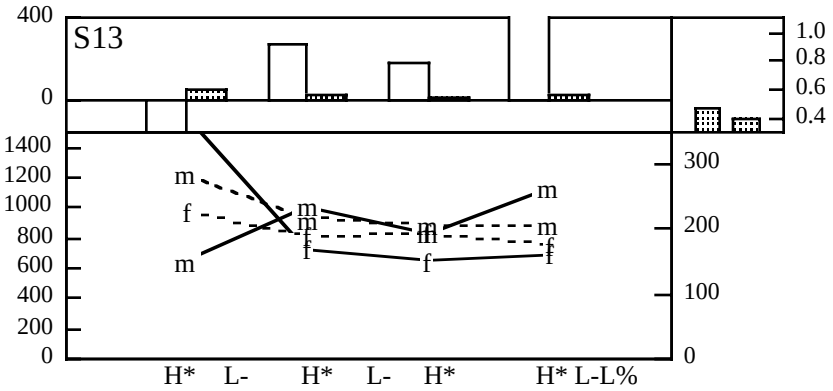


Figure 75. Example composite figure (labelled)

The following composite figures give the RMS amplitude measures, F0 measures, duration measures, and difference measures for each of the 31 pairs which were reliably and correctly identified, ended with low edge tones, and had either identical (13 pairs) or similar (18 pairs) of pitch accents. (There was a problem with one of the files, so only 30 composite figures are shown.) After all of the composite figures are presented, the differences between the tokens in each pair are summarized. A discussion of the size of the required difference in order for that case to be counted as "different" is discussed.



His friends had to drive him home at the end of the evening
(Dialogue 3, Sentence A)

Figure 76. Composite figure of measurements for Dialogue 3, Sentence A.

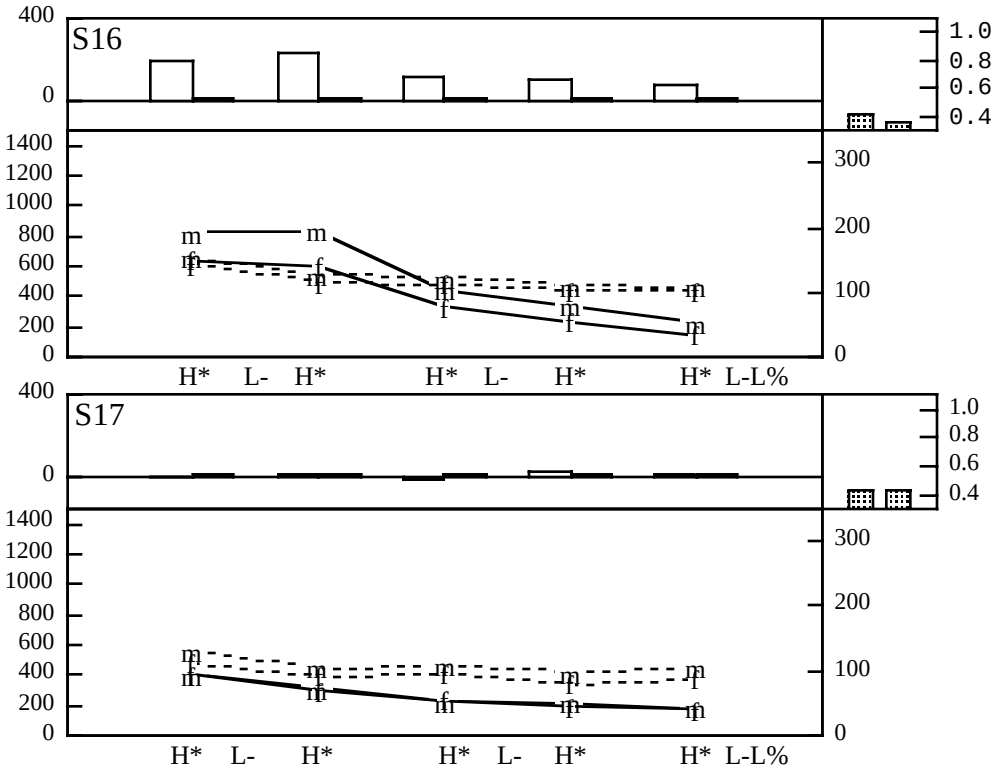
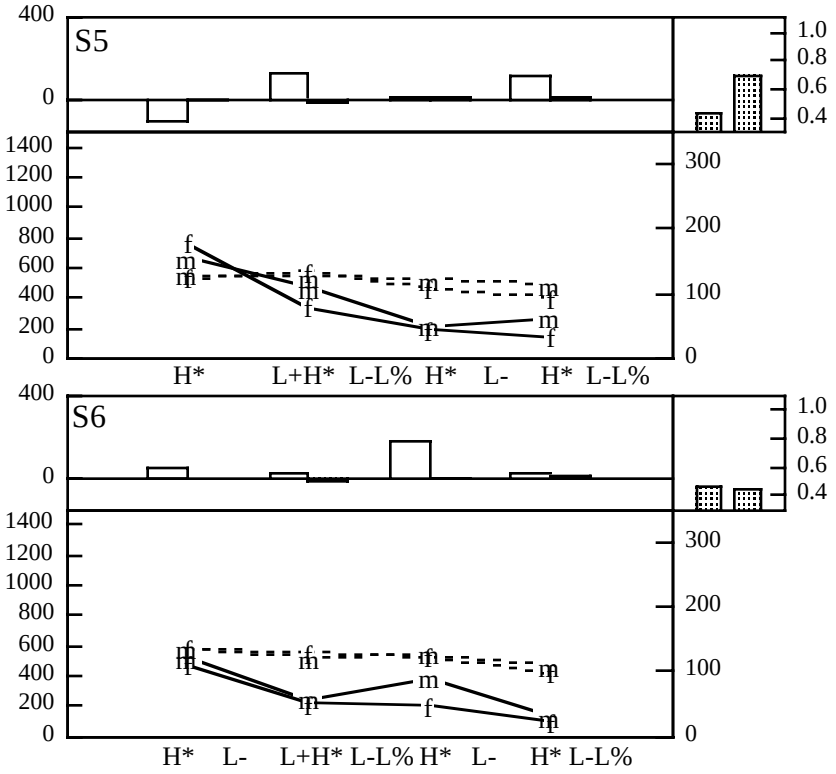
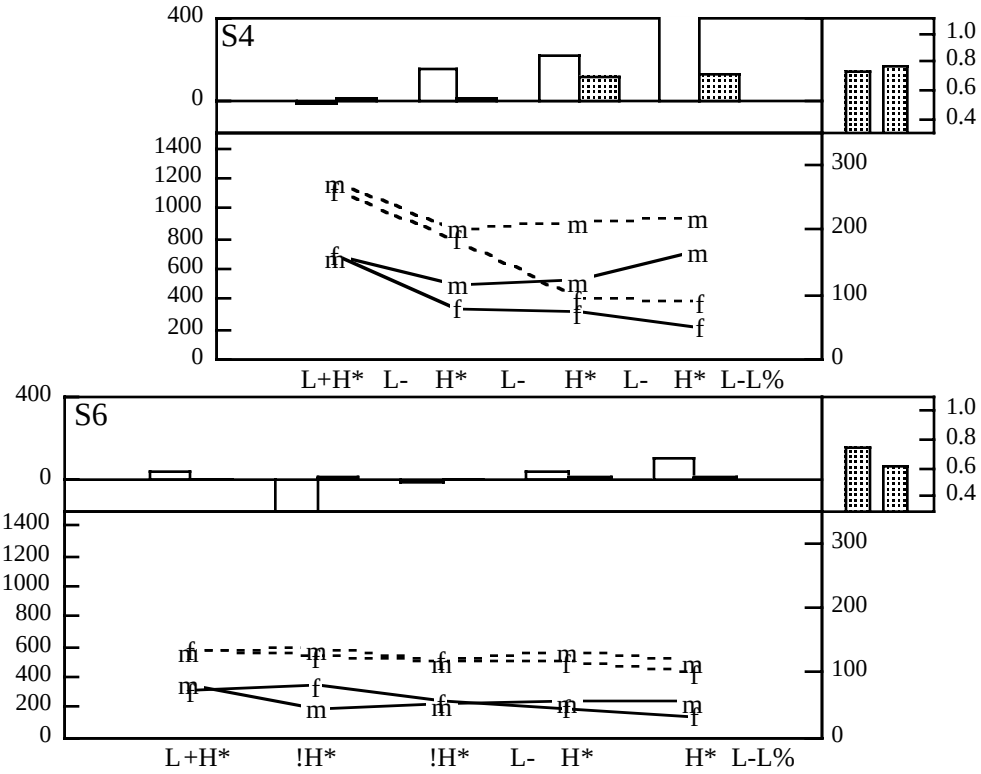


Figure 77. Composite figures for measurements for Dialogue 3, Sentence B.



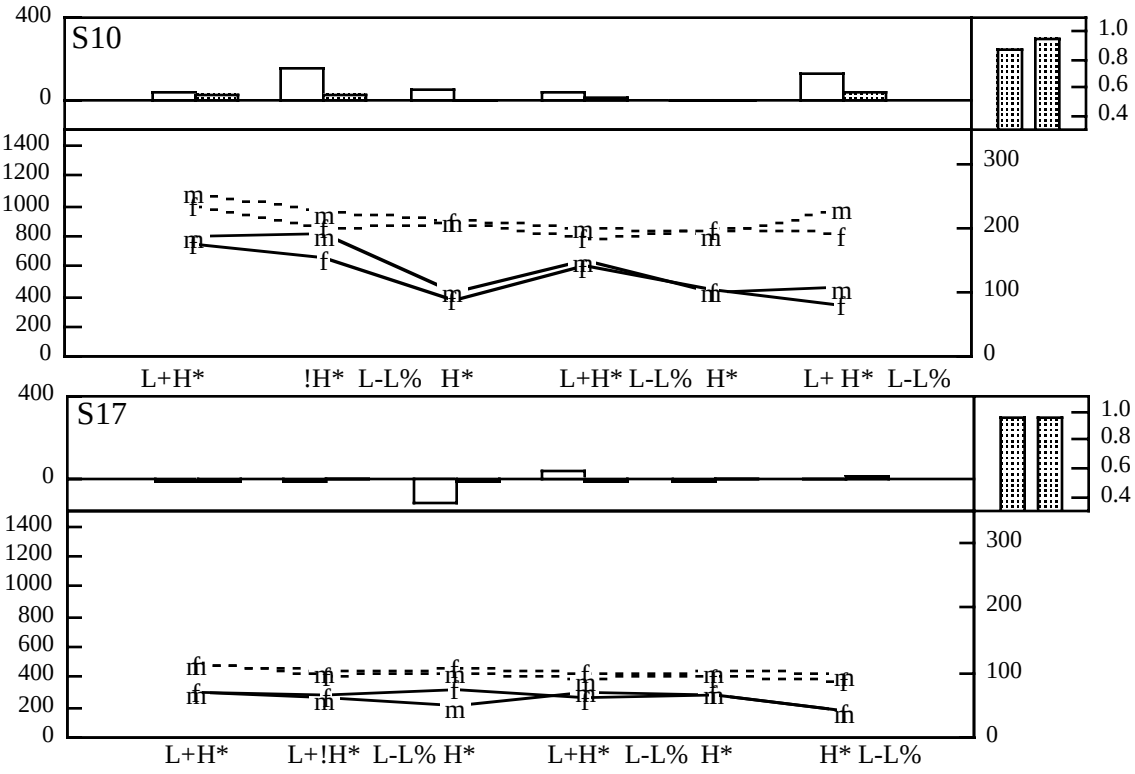
When I called him at 10 last night his roommate said he was sleepin
(Dialogue 4, Sentence A)

Figure 78. Composite figure of measurements for Dialogue 4, Sentence A.



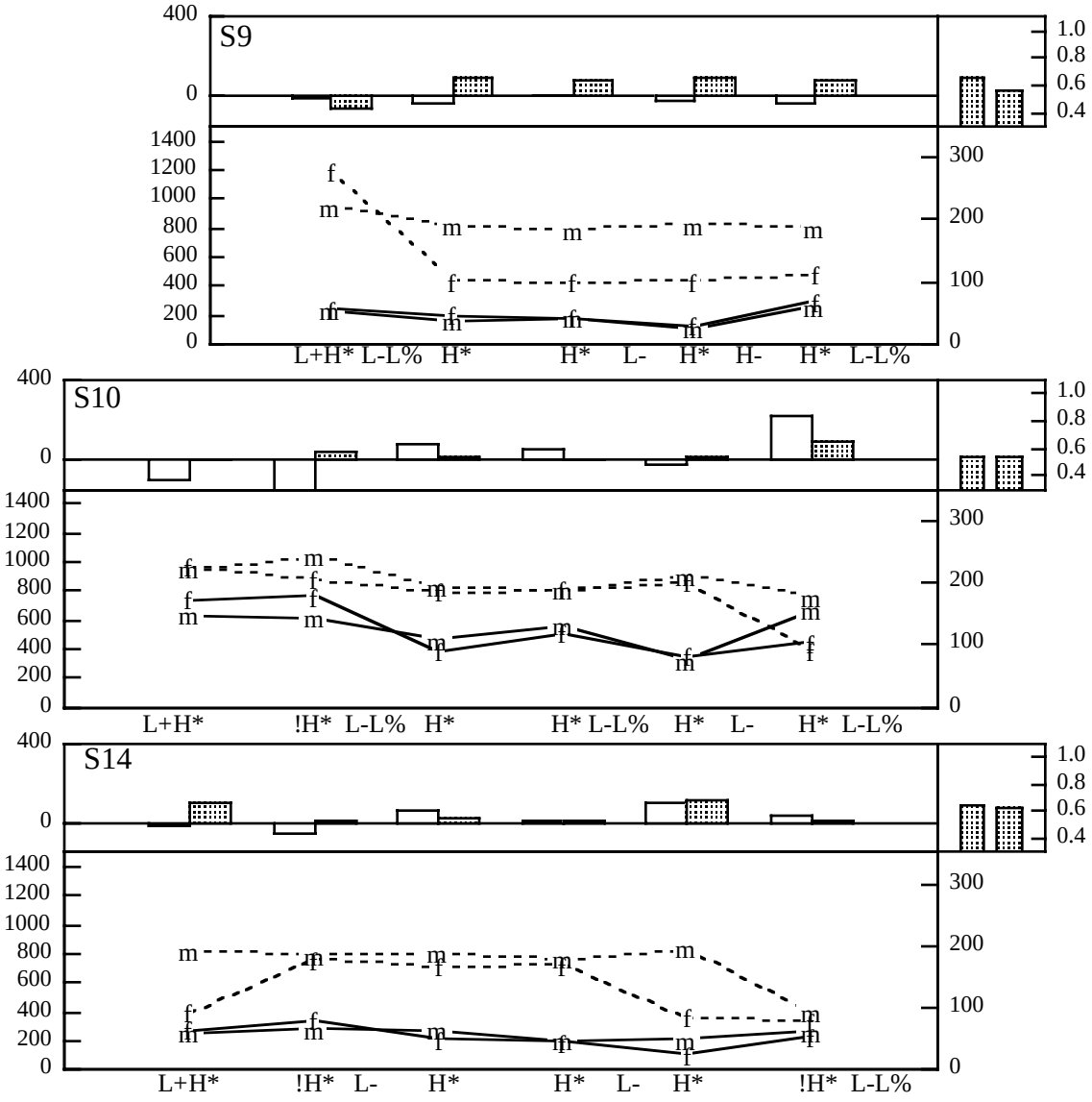
I didn't see his name on the sign-in list from the study session
(Dialogue 4, Sentence B)

Figure 79. Composite figures of measurements for Dialogue 4, Sentence B



Linguistics 201, which I'm taking right now, counts as my social science.
(Dialogue 5, Sentence A)

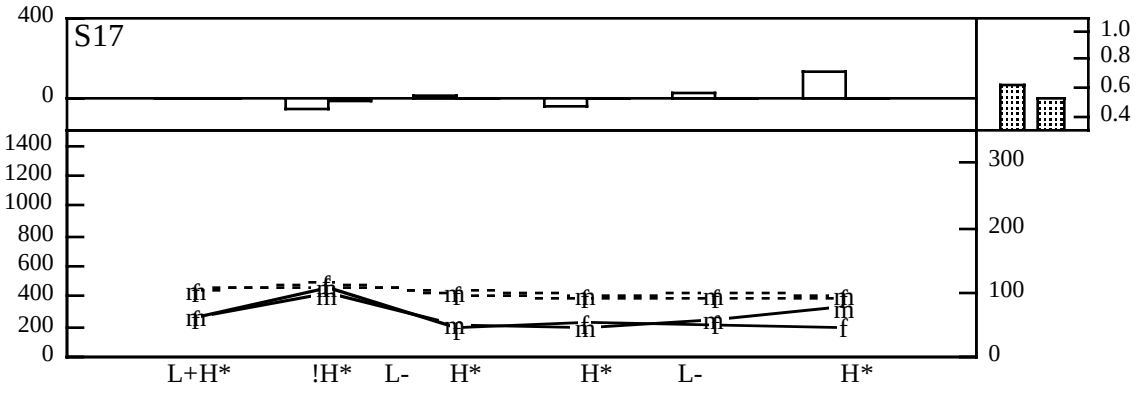
Figure 80. Composite figures of measurements for Dialogue 5, Sentence A



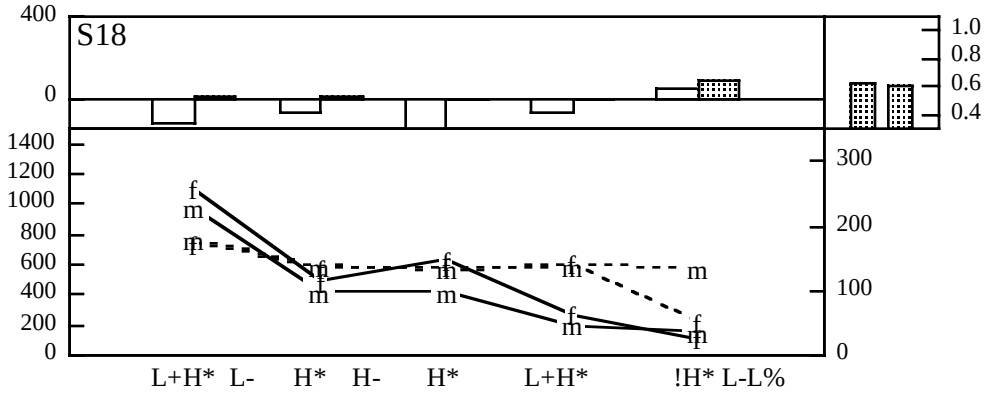
(to be continued)

Figure 81. Composite figures of measurements for Dialogue 5, Sentence B

Figure 81. continued



Next Fall I'm going to take Intro Anthropology
(Dialogue 5, Sentence B)



Also, she's involved in lots of clubs and activities.
(Dialogue 6, Sentence A)

Figure 82. Composite figure of measurements for Dialogue 6, Sentence A

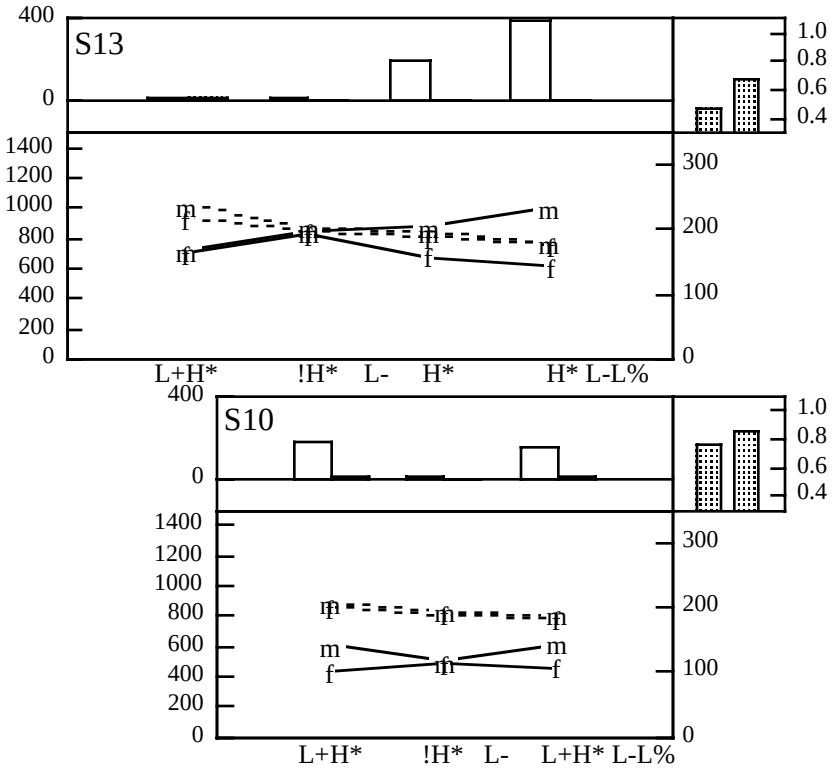
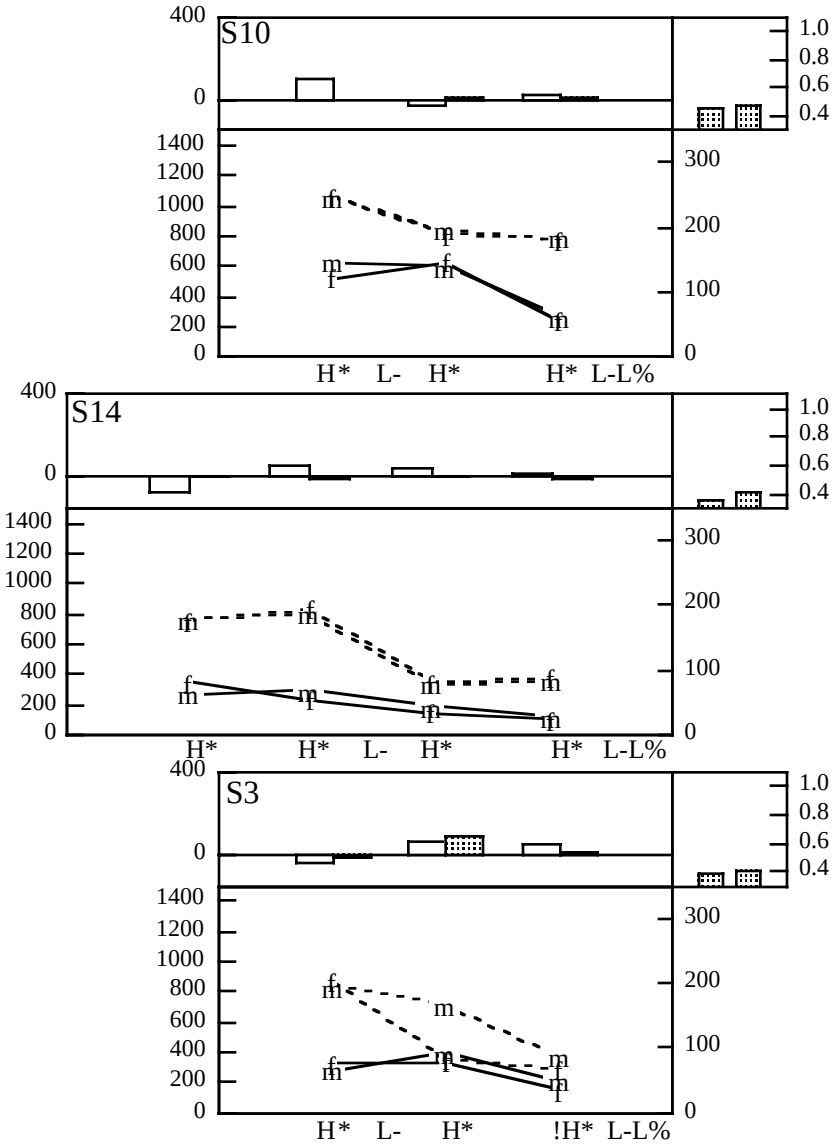


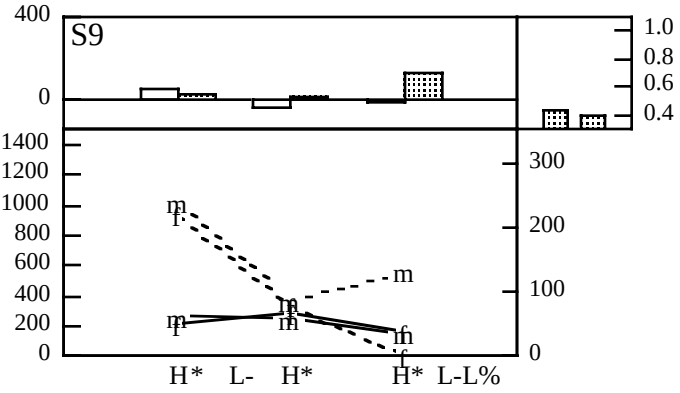
Figure 83. Composite figure of measurements for Dialogue 7, Sentence A



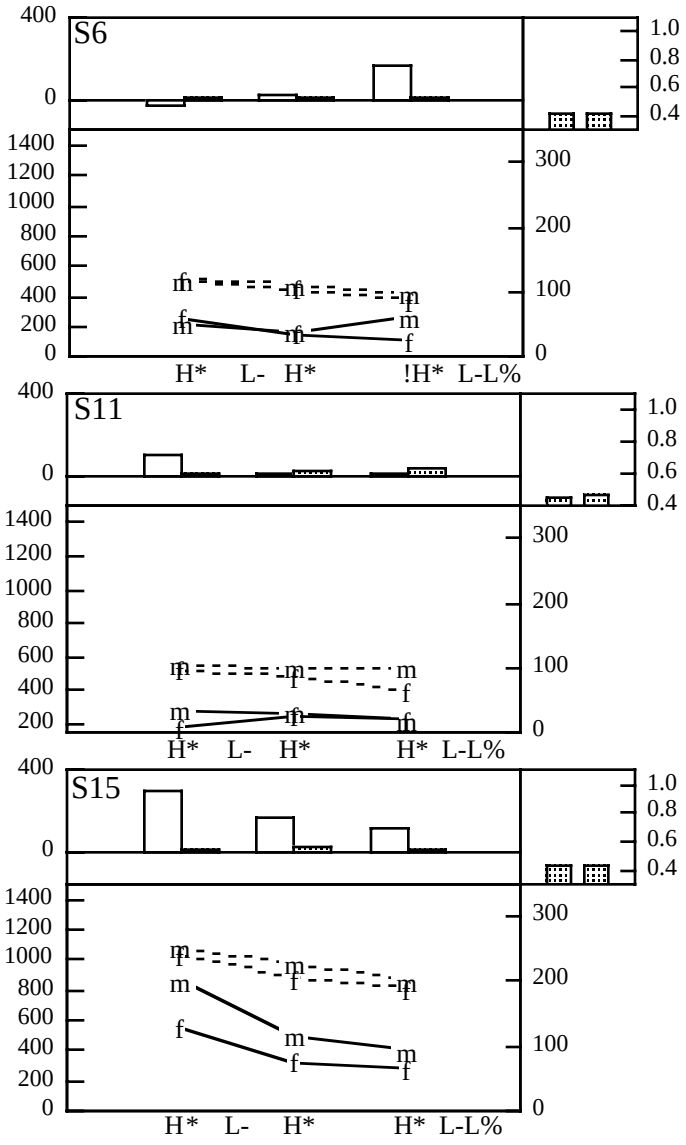
(to be continued)

Figure 84. Composite figure of measurements for Dialogue 8, Sentence A

Figure 84. continued



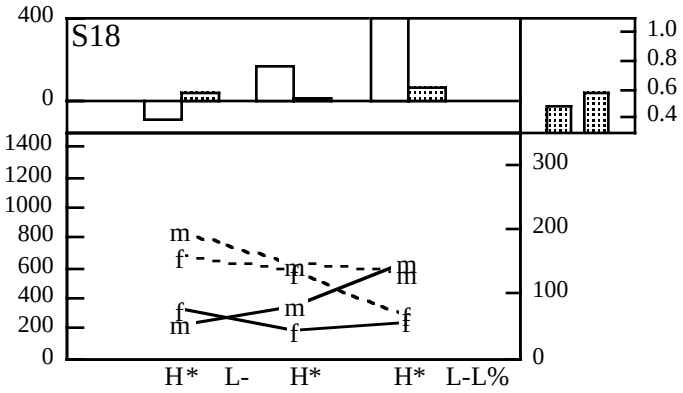
They take me about an hour each
(Dialogue 8, Sentence A)



(to be continued)

Figure 85. Composite figure of measurements for Dialogue 8, Sentence B.

Figure 85. continued



That usually takes me several hours
(Dialogue 8, Sentence B)

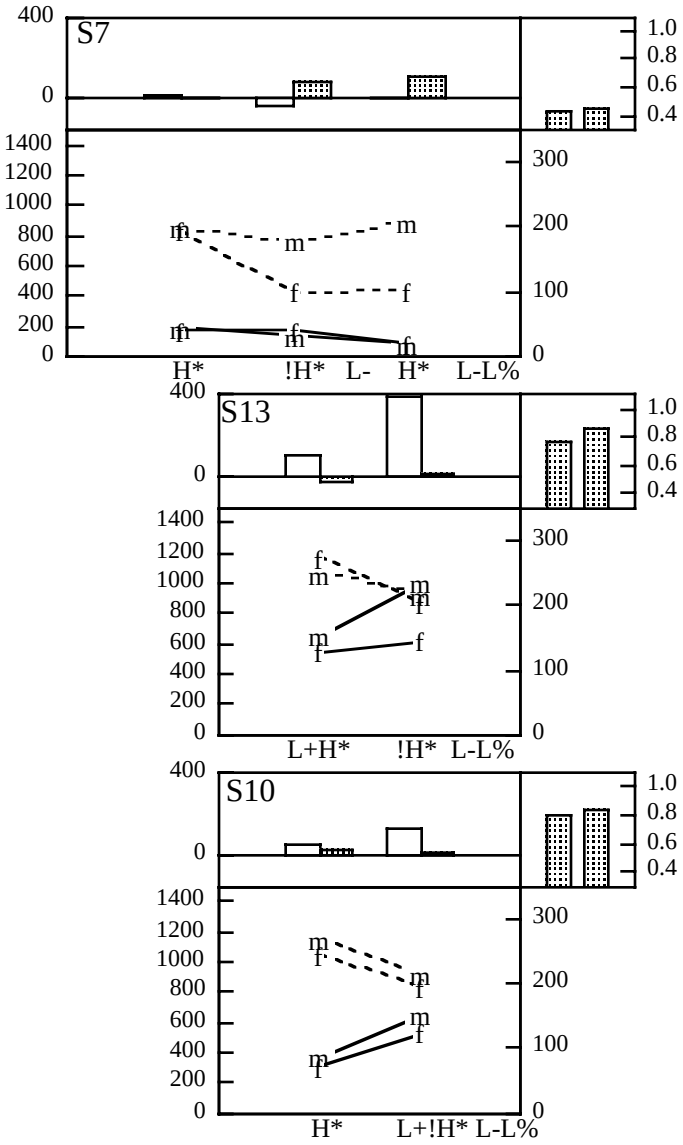
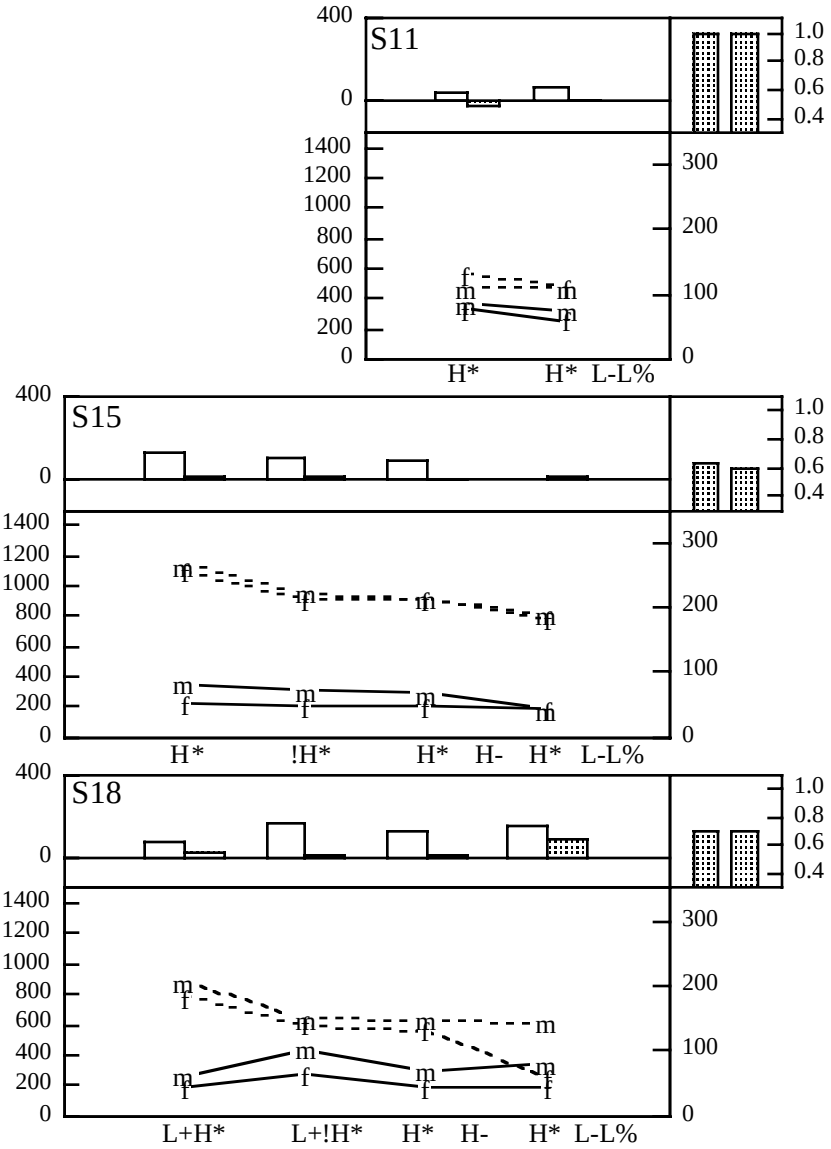


Figure 86. Composite figures of measurements for Dialogue 9, Sentence A



But she has a big exam she's studying for.
(Dialogue 9, Sentence B)

Figure 87. Composite figures of measurements for Dialogue 9, Sentence B.

The data will be summarized by showing which pairs had differences in RMS amplitude, F0, and duration such that the discourse-medial token had a higher value for RMS and F0 and a shorter duration in the final part than the discourse-final token.

However, in some cases the differences between the two tokens in a pair are quite small, seeming almost negligible, while in other cases, the differences are quite large. In order to define what counts as "different," a histogram of the RMS amplitude differences in the final pitch accented syllable is given in Figure 88.

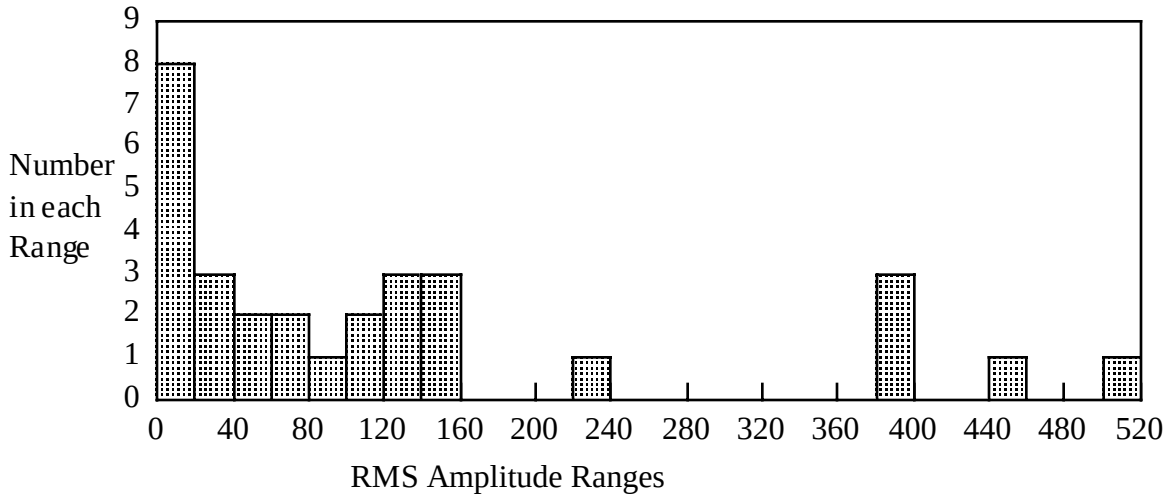


Figure 88. Histogram of RMS amplitude differences in final pitch accented syllable.

According to Figure 88, there seems to be a bi-modal distribution, with the cut-off point being at 80 (in arbitrary RMS amplitude units). There are 15 tokens whose difference in RMS amplitude is less than 80 and 15 tokens whose RMS amplitude difference is greater than 80. Thus, using this histogram as a guide, any pair whose difference in RMS amplitude in the final pitch accented syllable is less than 80 will be counted as "not different" while any pair in which the difference is greater than 80 will be counted as "different."

A histogram of F0 differences in the final pitch accented syllable is given in Figure 89. This histogram will be used to determine what counts as "different" in the F0.

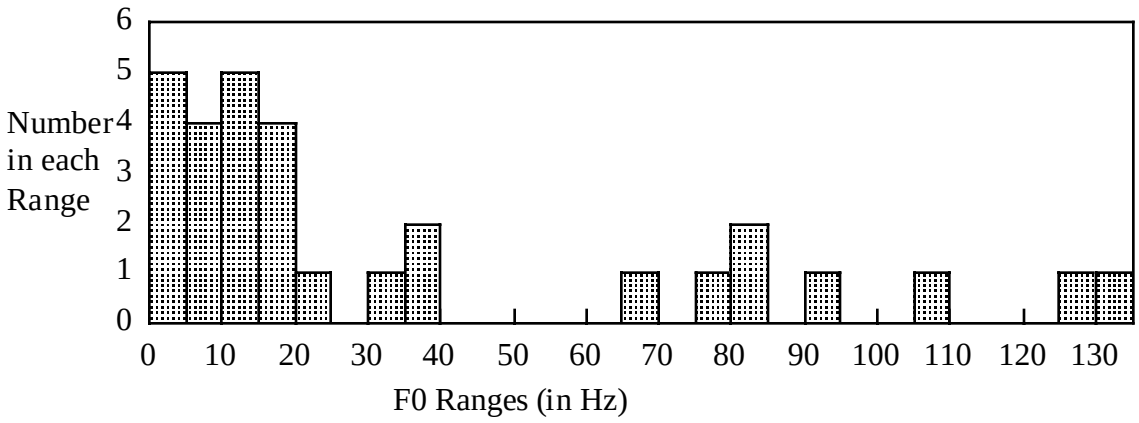


Figure 89. Histogram of F0 differences in final pitch accented syllable.

According to Figure 89, there seems to be a break in the distribution of F0 differences at 25 Hz. There are 19 pairs which have a difference in F0 in the final pitch accented syllable of less than 25 Hz, and these pairs will be counted as "not different." There are 11 pairs which have a difference in F0 in the final pitch accent greater than 25 Hz, and these pairs will be counted as "different."

In terms of the duration, the distinction will be made between a positive difference, in which the final part of the utterance in the discourse-final token is longer than in the discourse-medial token, and a negative difference, in which the final part of the utterance in the discourse-final token is shorter than in the discourse-medial token.

Furthermore, a general trend towards a difference in RMS amplitude over the course of the entire utterance will be marked impressionistically, as will a general trend towards a difference in F0 over the course of the utterance (and not just in the final pitch accented syllable).

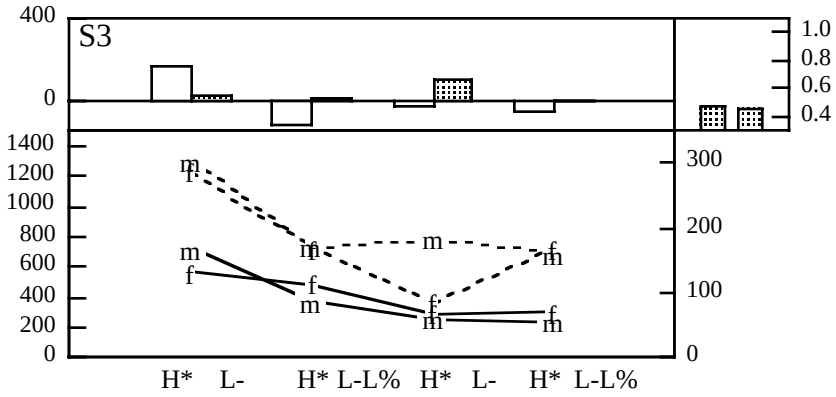
dialogue #, sentence, speaker #	difference in RMS amplitude in final pitch accented syllable	difference in F0 in final pitch accented syllable	duration in discourse-final longer than in discourse-medial	overall difference in RMS amplitude	overall difference in F0
D3 A S13	X	X		X	X
D3 B S16				X	
D3 B S17					X
D4 A S5	X		X		
D4 A S6					
D4 B S4	X	X	X	X	X
D4 B S6	X				
D5 A S10	X	X	X	X	X
D5 A S17			X		
D5 B S9		X			X
D5 B S10	X	X			X
D5 B S14					X
D5 B S17	X				
D6 A S18		X			
D7 A S13	X		X	X	
D7 A S10	X		X	X	
D8 A S10			X		
D8 A S14			X		
D8 A S3			X	X	X
D8 A S9		X			
D8 B S6	X				
D8 B S11		X	X	X	X
D8 B S15	X		X	X	X
D9 B S18	X	X		X	X
D9 A S7		X	X		X
D9 A S13	X		X	X	
D9 A S10	X		X	X	X
D9 B S11			X		X
D9 B S15				X	
D9 B S18	X	X	X	X	X

Figure 90. Summary table of differences in acoustic measures.

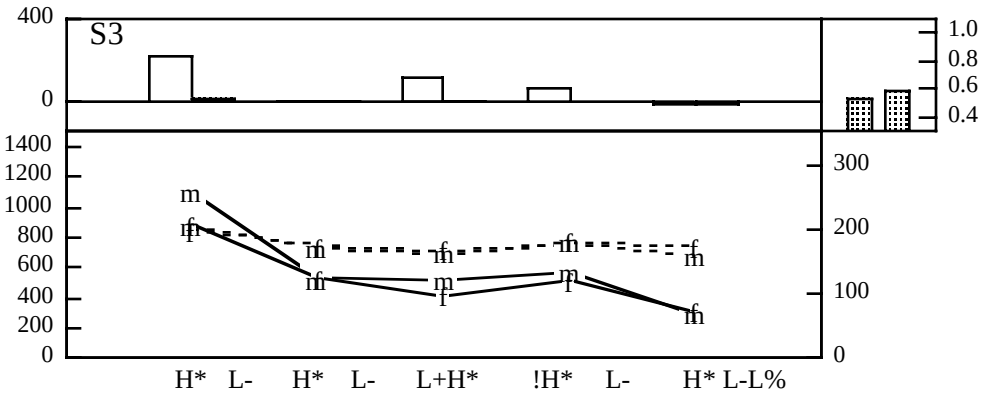
According to the summary table in Figure 90, it is not always the case that a difference in F0 in the final pitch accented syllable is accompanied by both a difference in RMS amplitude and difference in duration (final lengthening), as would be predicted by the

hypothesis that all of these acoustic factors are due to the more general control mechanism of “vocal effort.” However, it is the case that the 29 out of 30 of these pairs show at least 1 of the acoustic phonetic differences which were measured, and that although there are differences in the exact implementation from token to token (especially in where in the utterance the difference begins), the differences do tend to be in the same direction– higher RMS amplitude in discourse-medial than in discourse-final position, higher F0 in discourse-medial than in discourse-final position, and shorter duration in the final part of the utterance in discourse-medial than in discourse-final position.

The figures below give composite figures of the acoustic measures for the reliably mis-identified tokens. These cases would be expected to have reversals of the differences in the acoustic cues as compared to the pairs which were reliably and correctly identified. This is not always the case, though, as seen in the figures below. In some of the cases, such as for example Dialogue 1, Sentence B, Speaker 3, it might be that the steep rise in F0 to the final pitch accented syllable which occurs in the discourse-final token is responsible for the mis-identification. A similar rise (though not so steep) occurs in Dialogue 8, Sentence B, Speaker 9. In the other cases, though, it is not entirely clear what acoustic cue caused the reliable mis-identification.

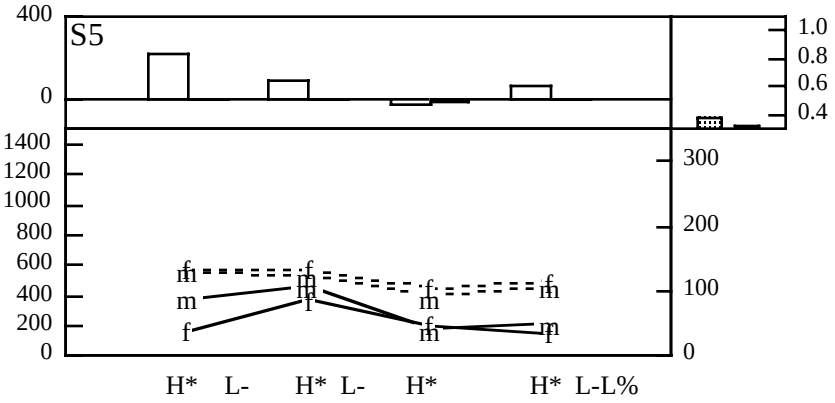


It's o.k. to have an advisor who's away, but only for one quarter
(Dialogue 1, Sentence B)-- mis-identified

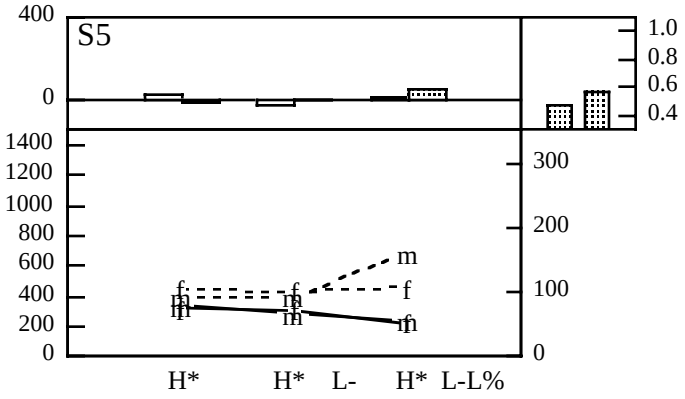


Also, she's involved in lots of clubs and activities.
(Dialogue 6, Sentence A)-- mis-identified

Figure 91. Composite figure of mis-identified tokens for speaker S3

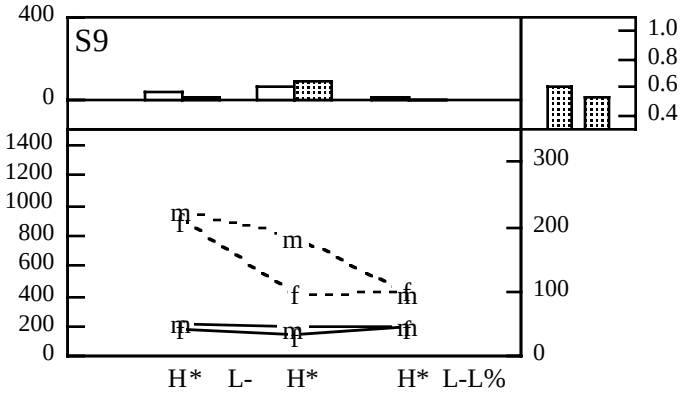


His friends had to drive him home at the end of the evening.
(Dialogue 3, Sentence A)--mis-identified

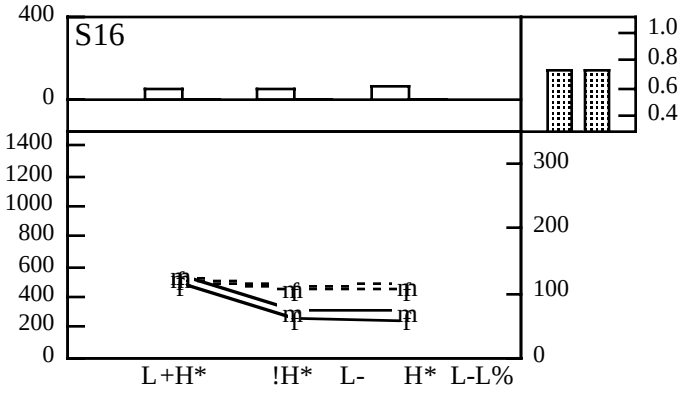


But she has a big exam she's studying for.
(Dialogue 9, Sentence B)--mis-identified

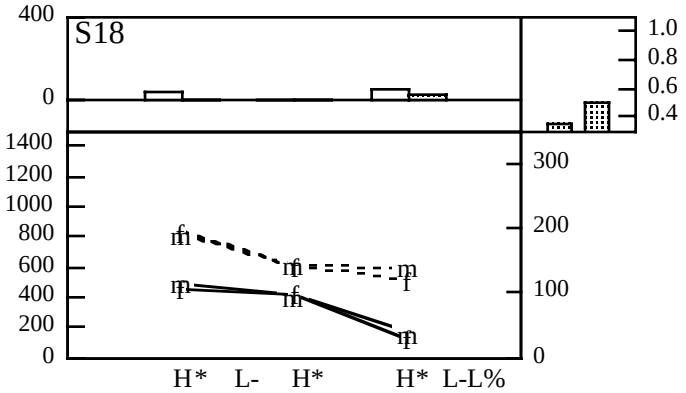
Figure 92. Composite figure of mis-identified tokens for speaker S5.



That usually takes me several hours.
(Dialogue 8, Sentence B)-- mis-identified



They raised the rent by forty dollars.
(Dialogue 7, Sentence A)-- mis-identified



They take me about an hour each.
(Dialogue 8, Sentence A)-- mis-identified

Figure 93. Composite figure for mis-identified tokens

In summary, the RMS amplitude peaks of pitch accented syllables for discourse-final vs. discourse-medial pairs tend to diverge in the direction of having a higher RMS amplitude peak for discourse-medial tokens than for discourse-final tokens (shown in the figures as a positive difference between the measurements). In cases in which the RMS amplitude peaks do indeed diverge, the divergence begins at different locations, sometimes happening over the entire utterance, sometimes beginning in the final intermediate phrase, and sometimes only occurring in the final pitch accented syllables. Thus, it seems that one acoustic cue to discourse position is for the entire utterance to be louder in discourse-medial position. Another potential cue is a "final fade-off" in which the RMS amplitude range contracts either in the final phrase or in the final pitch accent. In general, then, the tendency is for the RMS amplitude peaks of pitch accented syllables to be higher for discourse-medial tokens than for the corresponding discourse-final tokens.

The F0 peaks of pitch accented syllables for discourse-final vs. discourse-medial pairs also tend to diverge in the direction of having higher F0 values for discourse-medial tokens than for discourse-final tokens (shown in the figures as a positive difference between the measurements). Again, in cases in which the F0 values do diverge, the divergence begins at different locations, sometimes happening over the entire utterance, sometimes beginning in the final intermediate phrase, and sometimes only occurring in the final pitch accented syllables. Thus, one acoustic cue to discourse position is for the entire utterance to be in a higher pitch range in discourse-medial position. Cases with a difference in F0 at the final pitch accented syllable correspond roughly to Liberman and Pierrehumbert's (1984) findings on the final pitch-accent in a list of words being lower than predicted from a model of the declination over previous pitch-accents.

Incidentally, the presence of a difference in the final pitch accented syllable (and even more strongly, the fact that the difference in F0 between the two tokens extends beyond the syllable peak to the end of the voiced region, as exemplified in Figure 75

above, in which the difference in F0 extends to the end of the F0 traces) argues against the "constant sentence-final L" hypothesis described by Maeda (1976), Boyce and Menn (1979), and Liberman and Pierrehumbert (1989). This hypothesis states that the F0 reaches the same low point at the end of every sentence ending with a L edge tone. However, as Beckman and Pierrehumbert (1986) note, some of the studies which have found the low F0 point at the end of an utterance to be an invariant target used materials with highly simplified discourse structure, meaning that each of the sentences is elicited as a mini-discourse. Thus, the invariance in F0 may have been due to not enough variance in the materials. Supporting this point is Ladd and Terken's (1995) finding that overall raising of the pitch range also slightly raises the speaker's final F0 low. The study in this thesis, with its manipulation of the discourse structure, tends to support Beckman and Pierrehumbert's (1986) hypothesis as well as Ladd and Terken's (1995) finding, because the discourse structure was manipulated, and informal observation showed that the utterance-final low points varied (and the value at the final pitch accented syllable varied as well), as a function of the position in the discourse structure.

Another divergence between the sentences in a discourse-medial vs. discourse-final (but otherwise identical) pair is that the F0 from the first pitch-accented syllable in the sentence differs between discourse-medial tokens and discourse-final tokens. The F0 from the first pitch-accented syllable in the discourse-medial sentences tends to be higher than the F0 from the first pitch-accented syllable in the discourse-final sentences. This may serve to create a bigger contrast over the course of the utterance, thus perhaps lending the impression of a greater decrease in F0 over the utterance. This result is seen in examination of actual F0 traces, but does not in general appear in the summary plots which have the F0 from the syllable peak measured. This may be because perhaps the location of the F0 peak does not match the location of the RMS amplitude peak (and in these cases, at the very beginning of the utterance, there does seem to be a well-defined peak, contrary to the

situation later in the utterance in which it is not always clear where the F0 peak is). The examples in Figures 94 - 96 exemplify this difference, using the full F0 traces compared with each other, not the summary plots of the peaks. In these figures, the x-axis is time and the y-axis is F0 in Hz.

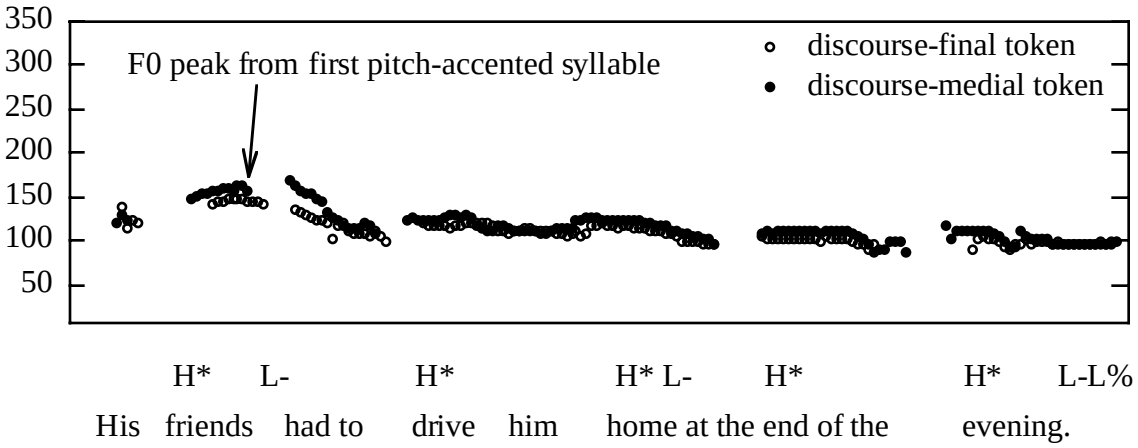


Figure 94. Example F0 traces– overlaid.

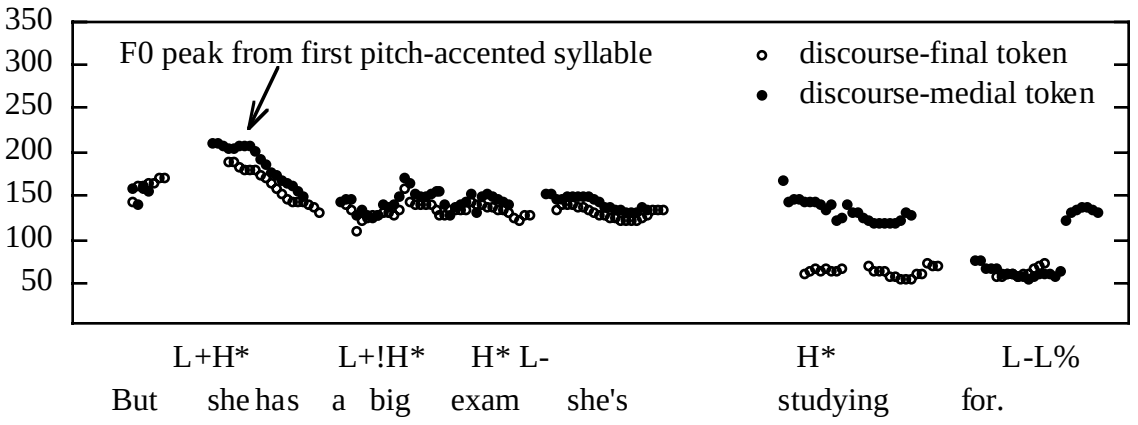


Figure 95. More example F0 traces– overlaid.

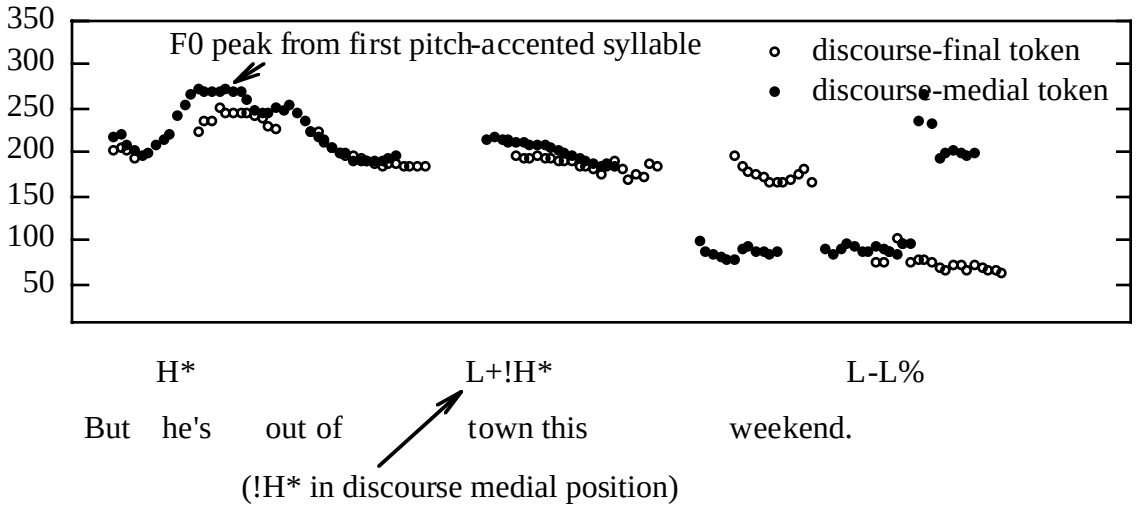


Figure 96. More example F0 traces– overlaid.

The trend is for the first pitch accented syllable in the utterance (which may or may not have been the first pitch accent in the discourse segment) to have a higher F0 value in discourse-medial position than in discourse-final position. This tendency is very slight, though, sometimes being a matter of only a few hertz, and only when measured at the actual peak in F0, not at the syllable peak. This difference may be giving the impression of a greater decrease in F0 in the discourse-medial utterances.

3.4 Summary

In sum, several results were found in this set of experiments. The first result was that the large majority of utterances in the corpus did terminate with low edge tones, as predicted by the "same edge tones hypothesis."

The next result is that the transcriber's tone space can be mapped out using multi-dimensional scaling. Furthermore, the mapping of the transcriber's tone space allowed for groupings of tones into similar tone target classes. This could be used in testing the "phonetic differences hypothesis," to have a method of comparing tones for similarity and

hence avoiding the phonetic comparison of tokens belonging to different categories.

Another result is that listeners can identify about 60% of the tokens correctly by discourse position, but that there was a slight bias in identification by order of presentation. Using a more stringent test of examining only tokens which were correctly identified by 3 or more listeners (out of 5) in both orders of presentation, there remained 117 pairs of tokens (out of 360) which met the requirement. Thus, it was shown that the "phonetic differences" in the "phonetic differences hypothesis" were perceptible to listeners, at least part of the time. Out of these 117 tokens, 15 pairs had either a H- or a H% (or both), leaving 102 pairs. Thirteen of the 102 pairs of utterances which were reliably and correctly identified and which had a L-L% at the end of the utterance had identical sequences of tones. The 89 remaining pairs were examined, and 18 of those were found to have sequences of pitch accents which were from the same tone target group, as found in the tone similarity mapping. Thus, there was a small set of utterances which could be examined to test the "phonetic differences hypothesis," because they had perceptible differences which were not due to differences in phonological category. Furthermore, 7 pairs of sentences which were reliably mis-identified were also examined. Thus, three groups of utterances were used in the acoustic analysis, the group which was reliably and correctly identified, had L-L% at the end, and had identical tones, the group which was reliably and correctly identified, had L-L% at the end, and had tones from similar tone classes according to the results of the tone mapping, and the group which was reliably mis-identified, had L-L% at the end, and had identical tones.

An acoustic comparison of the RMS amplitude and F0 of these groups of tokens revealed that the correctly identified ones mapped onto each other, with patterned divergences from each other. The first type of divergence was that the RMS amplitude peaks in pitch-accented syllables tended to be lower in discourse-final position than in discourse-medial position, beginning at varying positions in the utterance (either

throughout the utterance, in the final intermediate phrase, or in the final pitch accented syllable). The second type of divergence was that the F0 at the syllable peaks in pitch accented syllables tended to be lower in discourse-final position than in discourse-medial position, again beginning at varying positions in the utterance (either throughout the utterance or in the final pitch accented syllable). The third type of divergence was that the duration from the final pitch accented syllable to the end of the utterance tended to be longer for the discourse-final tokens than for the discourse-medial tokens. And the final type of divergence was that the F0 of the first pitch-accented syllable in the utterance was lower in discourse-final position than in discourse-medial position. These results do bear out the prediction that utterances which are the same except that one allows only itself to be popped from the focus stack while the other allows both itself and its dominant DSP to be popped from the focus stack will have different phonetic implementations of their phonological tones.

Thus, in a corpus in which there were pairs of sentences with the dominance relationships the same but in different orders, the phonological edge tones were the same in a large proportion of the pairs (although some other aspect of tonal selection distinguished many of the correctly identified pairs), but the phonetic implementations of tones in the utterances were different, which might be used as a cue as to how many discourse segments to pop off the focus stack.

CHAPTER 4

DISCUSSION AND CONCLUSION

4.1 Implications of Experiment Results

Two predictions about intonation and discourse were being tested. The first hypothesis, the "same edge tones hypothesis," was about the use of phonological intonational edge tones to express dominance relations in the intentional and attentional structures. This hypothesis was tested in discourses for which the set of dominance relations was the same but for which the dominance relationships occurred in different linear orders. In this case, if the intonational edge tones really do have as meanings the expression of dominance relations, then the phonological intonational edge tones are predicted to be the same for corresponding utterances from the two different discourses, and given the structure of the discourses used, with sister DSPs, the edge tones would be predicted to be low tones. (The converse prediction, which was not tested, would be that if the dominance relations are different, then the phonological edge tones chosen might be different from each other and include more L-H%, H-L%, or even H-H% sequences.) The second hypothesis, the "phonetic differences hypothesis," was about the use of phonetic implementation of tones to express how much should be popped off of the focus stack. This hypothesis was tested in discourses in which the set of dominance relations was the same but for which the dominance relationships occurred in different linear orders. What is

crucial is that in one case a daughter DSP allows only itself to be popped from the focus stack because its sister has not yet been satisfied but in the other case the daughter DSP allows both itself and the mother DSP to be popped from the focus stack (because it is the last of the daughters to be satisfied). In this case, if the phonetic implementation of phonological intonational tones expresses how many of the focus spaces can be cleared out of the focus stack, then the phonetic implementations of the tones are predicted to be different for corresponding utterances from the two different discourses. The results of the experiments did in general support these hypotheses, indicating that the interpretation of intonational tones as expressing local dominance relations and the interpretation of the phonetic implementation of intonational tones as expressing more global hierarchical intentional structures are plausible interpretations which deserve further study.

This study has made several contributions to our understanding of the relationship between intonation and discourse structure. The first contribution is in the actual design of the corpus. Setting up the corpus such that a logical structure with carefully controlled dominance relations was translated into natural language discourses allowed for testing of the predictions, even though this meant that what was recorded was read speech, which has been shown to differ from spontaneous speech. One potential problem with the way the experiment was set up is that there was no check to make sure that the discourse structure as perceived by participants in the experiment did match the intended discourse structure. This could have been tested by independent annotation of the discourse structure from text alone and from the text with the speech.

The implications of a study in which the intentional structure is controlled is that studies on the meanings of intonational tunes should examine how intonation varies as a function of variations in the discourse structure. Another method, in which it is not necessary to explicitly control discourse structure, used for example by Hirschberg and Nakatani (1996), is to annotate the discourse structure from a corpus of speech (Hirschberg

and Nakatani used a corpus of direction-giving monologues) and then to use the consensus labelings of the annotated discourse structure in examining the sound structure. In either method, though, the discourse structure is considered as a factor in the analysis. As an analogy, traditional field work in segmental phonology carefully varies lexical meanings and morphological paradigms in discovering the sound patterns and the phonemic system. The meanings are varied and the sound changes corresponding to those variations are noted in order to discover the functional categories. In the current study, the discourse structure was kept the same and the test was whether the tones which were produced were the same. Future studies in intonational phonology should also attempt to do the opposite, namely, to vary discourse structure by minimal units in systematic ways when trying to discover the tonemic system. If the "meanings" in the phonology of intonation correspond to discourse structures such as dominance relations, then those are exactly the elements which should be varied in order to discover the functional categories in the grammar of intonation.

The second contribution of this study concerns the prediction that if the set of dominance relations is the same for two discourses, even if the dominance relations are realized in a different order, they should have the same intonational edge tones, and those tones should be low tones because of the sister-of relation between the two subordinate DSPs. This prediction was borne out in the data. A large proportion of the pairs of utterances had L-L%, with only a few having either H- or H% (or both). The tokens with high edge tones might be accounted for by hypothesizing that the participants in the experiment had a different interpretation of the discourse structure than the discourse structure that was intended by the experimenter. This might be further explored by having subjects naive to the purposes of the experiment segment the discourses to see whether the presence of high edge tones corresponded to dominance relations in the discourse in which the phrase marked by a high edge tone needed to be interpreted with respect to a subsequent phrase. Another option to account for the presence of the high edge tones is as the speaker

indicating the need for reassurance from the experimenter that the utterance had been satisfactorily uttered (see, e.g., Ayers, 1994, on the use of high edge tones for seeking back-channeling from the interlocutor). Thus, there may be explanations for the presence of the few high edge tones which occurred, but the majority of discourse-segment-final phrases which did not have subordinate discourse segments were marked by low edge tones (L- and L%).

In order to avoid phonetic comparisons of tokens which were phonologically different, a method was developed to map out tone similarity. In this method, the transcriber was asked to give similarity judgements for every pair of tone categories. These similarity judgements were used to produce a multi-dimensional scaling solution. The maps which were produced as the result of the multi-dimensional scaling solution did have groupings of tones which seemed plausible to interpret as similar tone targets. For the pitch accents, a one-dimensional map seemed adequate and for the phrase accents and boundary tones, a two-dimensional map seemed best. Thus, the multi-dimensional scaling solution did produce groups of similar tones, so this method of mapping out a transcriber's "intuitive tone space" can be considered to have succeeded (although further refinements regarding combinations of tones and comparing combinations of pitch accents and phrase accents to bitonal pitch accents are needed).

Then, the groupings of similar tone targets from the multi-dimensional scaling solution could be used in the acoustic analysis in order to test which pairs of utterances had similar sequences of tones and so were comparable in the acoustic analysis.

The third contribution concerns the prediction that the phonetic implementation of the phonological intonational tones is tied to more global intentional factors such as whether the set of daughter DSPs for some dominant DSP has been completely satisfied or whether there are more sub-topics remaining to be addressed. This prediction was partly borne out

in the data, although there are still some uncertainties in this area. The results concerning this prediction came from two areas, perception and acoustics.

As far as perception goes, the question was whether discourse position was perceptible to listeners, more specifically, whether two utterances which were from different discourse positions (and allowed different amounts of material to pop off of the focus stack) were perceptible to listeners as being from different discourse positions. In some cases the utterances were correctly identifiable as being from different discourse locations. Only about 60% of the utterances were identified correctly by listeners. There was also a slight bias towards identifying the utterance heard second in the pair of utterances as "paragraph-final." Furthermore, a more stringent criterion which accepted only utterances which were identified correctly by 3 or more listeners (out of 5) in both orders of presentation resulted in only 117 pairs out of 360 which were reliably and correctly identified. This means that listeners were correctly and reliably identifying the discourse position of an utterance in a pair of utterances about one third of the time. This somewhat low value may be due to a few factors. For example, this was read speech, and some of the subjects felt uncomfortable reading out loud. Thus, they may not have been able to express discourse intentions as clearly as they could have in spontaneous conversation. Furthermore, the fact that it was read speech from a printed dialogue and that the other conversational participant was the experimenter, who obviously already knew the full intentional structure, may have caused the speakers to not convey their intentions as carefully as they might have otherwise. Furthermore, in natural conversations, speakers use not just intonation but also other factors such as cue words (Grosz and Sidner, 1986) to indicate intentional structure, while this experiment was somewhat artificial in this respect in that the exact same sequences of words were used in both discourse locations. Thus, if speakers had been allowed to change the wording and to add cues such as "and finally," then of course listeners in the perception test would be expected to perform better

than they actually did. On the other hand, this rather low number of correct and reliable identifications may provide a clue to the nature of interruptions or to the nature of "But I'm not done talking about our first subject yet" misunderstandings. That is, if the more global intentional structure can be conveyed through the phonetic implementation but not always reliably, then this means that the speaker's more global intentions are not always clearly conveyed to the listener, and the listener may try to change the subject (pop the dominant DSP) before the speaker is ready. Thus, although the results from the perception tests indicate that discourse structure is not always clearly conveyed by prosody, there may be explanations for this stemming from the nature of the experimental conditions.

The acoustic analyses focussed on cues the listeners might have been using when they reliably identified the discourse position of a pair of utterances, when the utterances had the same string of phonological tones. This was a relatively small proportion of the correctly identified pairs, indicating that choice of intonational tone may be an important cue to discourse finality. The acoustic analysis focussed rather on those cases in which the medial and final tunes were matched and asked how listeners could distinguish between discourse-medial and discourse-final utterances. In order to answer this question without confounds from different pitch accents, phrase accents, or boundary tones providing cues, this acoustic analysis was limited to only those utterances which were correctly and reliably identified by listeners, which did not have a H- or a H% at the end of the utterance, and which had either identical sequences of pitch accents or else similar sequences of pitch accents based on the tone similarity mapping. There were 13 pairs of utterances from the corpus which matched these requirements and had identical tones and an additional 18 pairs of utterances from the corpus which matched these requirements and had similar pitch accents.

The results of acoustical analysis of these pairs of utterances were that for the most part, when RMS amplitude curves from pairs of sentences were overlaid on each other and

when F0 curves from pairs of sentences were overlaid on each other, they did match up quite well, with a few patterned divergences from each other. The first difference was that the RMS amplitude peaks for pitch accented syllables tended to be lower in discourse-final utterances than in discourse-medial utterances. The second category of divergences was that the F0 tended to be lower in the syllable peaks of pitch accented syllables in discourse-final utterances than in discourse-medial utterances. The third category of divergences was that the duration of the final part of the utterance tended to be longer in discourse-final utterances than in discourse-medial utterances. And the final category of divergences was that the F0 of the first pitch accented syllable in the sentence tended to be lower in discourse-final utterances than in discourse-medial utterances, perhaps in order to provide a greater change over the course of the utterance. These tendencies show some similarities to earlier findings as discussed in section 1.5. For example, some of the phonetic factors found by other researchers include low terminal F0 and differences in maximum F0 and RMS amplitude in the phrase. The current study did find some tendencies towards differences in "terminal F0" in that the F0 of the final pitch accented syllable tended to be lower in discourse-final position than in discourse-medial position (athough whether "terminal F0" is analogous to the final pitch accented syllable is not clear). Also, there were differences in F0 and RMS amplitude when comparing phrases, such that discourse-medial utterances tended to have higher F0 and RMS amplitude than discourse-final utterances. Thus, there are some correspondences between the current work and the earlier work on phonetic correlates of discourse structure, if the earlier work is interpreted so as to allow for analogy between "terminal" or "discourse-segment final" and the discourse-final position in the current study. In sum, there were some phonetic factors which differed between the two corresponding utterances in a pair, indicating that the phonetic implementation of the phonological tones can in some cases serve to cue differences in the global intentional structure.

Thus, the phonetic implementation is providing a functional difference in this case. In general in the division of labor between phonological category differences and phonetic implementation, it seems that the phonological category differences bear the responsibility of indicating meaning differences or functional differences. This case is different, however, because the discourse function of the phonetic implementation is actually to cue the more global intentional structure, which is the underlying logical structure of the information in the discourse. This puzzle, of the phonetic differences in this case having a crucial functional difference, requires further examination.

4.2 Directions for Future Research

Although the results of these experiments did in general support the predictions, they also open up new areas of research and raise new questions in the area of intonation and discourse. The areas of research which arise from this study fall into four general areas– further explorations of discourse structure, further examinations of phonological tones, broader explorations of the phonetic implementation of the phonological tones, and cross-linguistic research on these issues. Each of these general areas will be discussed in turn.

4.2.1 Discourse Structure

First, further exploration is needed in the area of discourse structure. For example, more research is needed on the focus stack. In studies on discourse structure which use simply structured discourses, a rather neat and clean model of the focus stack can be used in which sister DSPs are never simultaneously in the focus stack. That is, as each sub-topic is addressed, it is popped completely off of the focus stack and cannot be used in further interpretation. However, it is not clear if this holds true in more complex, more realistic conversations as well, or whether there are effects such as the recency of a topic allowing it to be referred to or not. Furthermore, there is another option which could result in popping a whole sub-topic off of the focus stack besides simply satisfaction of all dominated DSPs,

and that is the possibility of the "practically un-answerable question." (Roberts 1996) In this case, conversational participants decide to abandon the topic without satisfying all of the DSPs (throwing up their hands, as it were, in an intuitive sense, and continuing to the next topic), and this possibility must be allowed for as well in a more realistic model of the focus stack.

Further explorations of the focus stack may also illuminate the relationship between the use of intonational cues and use of other cues such as referring pronouns and referring expressions, similar to Nakatani's (1997) study. Such studies may be able to determine whether the use of different types of cues play off of each other or not. That is, does the use of more cue phrases and more referring expressions reinforce the use of intonational cues, and if so, how are they related to each other?

Furthermore, additional studies on discourse structure may also be helpful in determining what the practical limits are on locality in the focus stack. That is, can only one discourse segment intervene and still allow for locality of a daughter DSP and its dominating DSP? Or can more discourse segments intervene and still allow for the dominance relation between the now non-adjacent daughter DSP and its dominating DSP to exist and to be recognized by conversational participants? Is there a practical upper limit on how many intervening daughters there are which still allow for locality between a daughter DSP and its dominating DSP, and does this limit vary from person to person?

Also, it seems crucial to explore the nature of compositionality in the compositional account of tone meaning. How are the meanings composed? This is a question which is being worked out in formal semantics for lexical and compositional meanings. Now, this issue needs to be worked out in pragmatics as well, in the area of discourse structure. If it is true that different intonational tones have different "meanings" in the sense of dominance relations and other discourse structures, then how are the meanings of strings of such tones composed from the meanings of all of the individual tones? Certain intonational contours in

English seem to be idiomatic and even are referred to by names in the intonational literature, such as the "calling contour," the "surprise-redundancy contour," and the "uncertainty contour." If we want to maintain that intonational meanings really are compositional, then there are a few options. One option is that these are stylized contours, analogous to frozen idiomatic expressions. Just because there are frozen idiomatic expressions whose meanings are not compositional, that does not mean that the compositional semantic approach is a doomed enterprise. Similarly in the intonation, if it turns out that these contours are stylized idiomatic expressions with a fixed pragmatic meaning, that still does not doom the enterprise of finding compositional meanings for individual tones. Another option regarding compositionality is that the meaning comes from one of the components of the contour, and not the entire contour. As Pierrehumbert and Hirshberg (1990) point out with respect to the "incredulous" or "uncertainty" interpretation of the L*+H L- H% contour (discussed by Hirschberg and Ward (1992)), it may be that the contour interpretation is "more properly associated with the L*+H pitch accent rather than the entire contour" (p. 295), which they demonstrate by sustituting in a H phrase accent and showing that there is still the conveyance of uncertainty. This type of preliminary finding calls for more study on the stylized, well-recognized contours to see if they are more readily analyzable as frozen idiomatic expressions or if the meaning of the whole really is a product of the meaning of the parts and the way that they are composed (which is the compositional account).

Also in the area of further explorations of discourse structure is the need for the study of more naturalistic discourses. The discourses used in this set of experiments were rather artificial in the sense of being based on a carefully controlled logical structure and being read out loud by speakers. There are, however, questions about what happens in more spontaneous conversations and whether the results found here can be supported by studies of more natural discourses. For example, natural discourses generally involve more than one switch between speaker and listener, and so there is also the issue of turns and

turn-taking in discourse. Thus, further exploration of more natural discourses will allow for more understanding of the interaction between turns in discourse and the handling of discourse segments and discourse segment purposes. Since conversational participants may have different intentional structures, and since the goal of the discourse is to communicate the intentional structure to one's interlocutor and to arrive at a synthesis of the interlocutors' goals, then how is the pushing and popping of DSPs in the focus stack negotiated by the participants? Also in the area of turn-taking in conversation is the question of how intonation plays in to the overall negotiation of turns. Studies such as Kreiman (1982), Schaffer (1984, 1983), Cutler and Pearson (1986), and Stephens and Beattie (1986) have examined the use of intonational cues to turn taking in conversation. Such investigations have laid the groundwork, giving some hints about useful methodologies. For example, Stephens and Beattie (1986) make a distinction between agreement and disagreement conditions, and they found that the turns are more reliably identified in perception experiments when they are extracted from the disagreement condition, perhaps due to the speakers having a greater stake in holding their turn in an argument. Such studies, while providing a useful background, require further extension by taking into account the phonological tones when discussing phonetic cues. In general, then, studies of more naturalistic conversations are needed in order to investigate the interactions between turn-taking and the negotiation among conversational participants of the intentional structure, and also in order to investigate the use of intonational cues on turn-taking.

4.2.2 The Phonology of Intonation

The next area which deserves further exploration is the area of the phonemic system of tones and determining what it is in general and which tones are present in specific instances. That is, careful phonemic analyses varying the intentional and attentional structures may help to clear up some consistently problematic and debated issues in the phonology of intonation. For example, do the nuclear tone and the phrase accents and

boundary tones vary separately as functions of different aspects of discourse structures, or do they form a complex, varying all together? If it turns out that minimal variations in the discourse structure result in minimal variations of nuclear tones and that other minimal variations in the discourse structure result in other minimal variations of the edge tones, then this would provide support for Pierrehumbert's analysis of nuclear accents as independent of the edge tones. On the other hand, if minimal variations in the discourse structure affect both the nuclear tones and the edge tones, then this would provide support for Gussenhoven's (1984) analysis of a complex of tones (see section 1.3).

Also, the tokens which had different phonological tones in the two different discourse positions need further examination. Investigation of whether there were patterned differences in the choices of pitch accents in the two different discourse position may shed more light on Nakatani's (1996, 1997) results on the discourse functions of pitch accents.

Furthermore, extensions of the method described in this thesis for mapping out a transcriber's "intuitive tone space " may be useful in comparing between transcribers and in further studies on inter-transcriber agreement. Determining whether two transcribers have similar tone maps or not and along what dimensions they differ may help to illuminate the differences between transcribers' transcriptions.

4.2.3 The Phonetics of Intonation

The area of phonetic implementation of phonological tones also merits further study. For example, it would be interesting to see how finely gradient the phonetic implementation can be. That is, how many degrees of "finality" can be expressed (defined as how many focus spaces can be popped from the focus stack at one time)? The older work which used the categories of sentences and paragraphs seems to imply that there are two degrees of finality. However, this newer understanding seems to implicate a theoretically unbounded amount of focus spaces which may be popped from the focus

stack at any one time. So the question is what the practical limits are of the phonetic implementation, and how many degrees of finality can be expressed.

Still in the area of phonetic implementation there are also further questions about the acoustic factors. The factors explored here were RMS amplitude and F0, but further studies are needed. Other acoustic phonetic factors which were not studied here but which have been found to be phonetic correlates of discourse structure include more detailed study of final lengthening, subsequent pauses, and the use of different modes of voicing such as laryngealization. A pilot study to the present study also found that a sharp, audible inhale used by some speakers at the end of an utterance served as a reliably perceived cue to discourse-medial position (indicating that more was to come in the discourse). Thus, phonetic factors besides just F0 and RMS amplitude need to be examined further in order to have a more comprehensive study of the phonetic correlates of discourse structure.

Also, laryngealization has not been examined here but is often present at the ends of utterances. This voice quality is variously referred to as laryngealization, glottalization, glottal fry, creak, or creaky voice. It has been postulated as a phonetic correlate of discourse boundaries by Lehiste and Wang (1975) among others. It was also hypothesized in section 3.3.2 that the frequent occurrence of pitch-halving in the final pitch accent in the sentence was due to changes in voice quality causing perturbations in the waveform. Thus, one desirable measure in the study of the phonetic correlates of discourse structure might be the identification of laryngalization in the signal, and if possible, also the identification of more vs. less laryngealization in the signal. There are two general methods of identifying laryngealization in the speech signal, identification by visual and auditory inspection of the signal and automatic identification using measures of perturbation. Several investigators have resorted to visual and auditory examination of the signal. For example, Henton and Bladon (1988) used visual inspection of the spectrogram for widely spaced striations and visual inspection of the spectrum for “a very fine-grained quasi-harmonic structure which

effects a low-amplitude modulation of the overall shape." (p. 16) Similarly, Pierrehumbert (1994) used pair-wise comparisons between waveforms, looking for "an irregular excitation pattern involving extra-long pitch periods." (p. 45) Pierrehumbert (1994) used a sign test to investigate the pair-wise comparisons, but another option might be to use Bard et al.'s (1996) method of magnitude estimation, in which subjects associate a numerical judgement with a physical stimulus according to perceived ratios between an initial stimulus and the stimulus in question. Such a method might be adapted to the perception of laryngealization by presenting listeners with modal voiced tokens and also tokens with different voice qualities. Listener judgement scales could then be used to rank the tokens according to least-to-most laryngealized. A refinement of the method of visual identification of laryngealization can be found in Hagen (1997), in which a hidden-Markov-model / neural net hybrid was trained to detect glottalization. Although this seems to be an automatic classification of glottalization, it must be remembered that the system is trained on labelled data, in which a classification system for different types of laryngealization (glottalization, diplophonia, etc.) is used to label a large corpus of training data. The labels, or descriptive features, were applied to the training corpus by hand, meaning that the process here was visual identification of glottalization followed up by training of the HMM/NN system and then automatic identification. Such an approach is useful because it eventually allows for automatic identification of glottalization (after the training period), but it must be understood as being based on the visual/auditory method of identification of laryngealization. These, then, are some potential ways of identifying laryngealization in the signal by visual and auditory inspection, as well as some potential ways of analyzing the identifications once they are made. The other general approach to identifying laryngealization is through automatic perturbation detection. As Pierrehumbert (1994) suggests, irregularity in glottal pulses might be one of the hallmarks of laryngealization. The question is how irregularity in the signal is to be identified automatically. For example,

Titze (1992, chapter 11) discusses perturbation measures such as jitter (cycle-to-cycle fluctuations in the period) and shimmer (cycle-to-cycle fluctuations in the amplitude). However, as work such as Bielamowicz et al. (1996) points out, “The fact that jitter (a measure of aperiodicity) lacks robustness with respect to aperiodicity seems to us a fatal paradox inherent in this measure.” (p. 134) They further note that “The fact that measures of aperiodicity apparently cannot be reliably applied to signals that are even slightly aperiodic leads us to question their utility in analyzing vocal quality...” (p. 134) Such paradoxes are inherent even in more mathematically sophisticated methods of signal analysis, such as Behrman and Baken’s (1997) analysis of the correlation dimension of the signal (in this case of the electroglottographic signal). The correlation dimension is a tool from nonlinear dynamics which, in the simplest understanding, is “an estimate of the complexity or irregularity of a system in space or time,” (p. 2371) and which “defines the number of essential degrees of freedom of ... the dynamical system.” (p. 2372). Even this very sophisticated mathematical modelling method, though, runs into the problem that it works well for signals which are nearly periodic and have an absence of irregularities, but presents problems when applied to data with many non-stationarities (such as data from pathological voice qualities). As Behrman and Baken point out, it may be the non-stationarities themselves which are of interest, meaning that the utility of the correlation dimension in the identification and analysis of various voice qualities may be questionable. Thus, even methods which have attempted to automatically identify laryngealization (or other voice qualities) have not succeeded in finding an algorithm. And in fact, Kreiman and Gerratt’s (1996) discussion of the reliability of identification of voice qualities by listeners points out some problems with the assumption of a common perceptual space shared by listeners, which is necessary in order to use rating scales to identify tokens which have a particular characteristic quality and which can be examined for a common acoustic factor. Kreiman and Gerratt note that the perception of voice quality is a function of both the

listener (including the listener's experience, training, and perceptual strategy) and the voice stimulus, not just a function of the voice stimulus itself. Thus, they suggest modelling quality as "a function of both a particular signal and a specific listener." (p. 1794) They recommend analysis by synthesis as a technique for doing so, because analysis by synthesis will allow listeners to construct a signal which matches the natural voice quality under observation. Thus, analysis by synthesis allows for the modelling of "the relationship betwen a particular voice and listener on a particular occasion as a unique relation" while cutting out the rating process as an intervening step. (p. 1794) And there has been some succes at synthesizing glottalization, although the results were only subjected to perceptual experiments by the experimenters themselves. Pierrehumbert and Frisch (1997) found that glottalization could be synthesized either by lowering the fundamental frequency, lowering the amplitude, or reducing the open quotient of the glottal pulse, which flattened the spectrum. These parameters were adjusted by "trial and error to match the parameters to each talker's voice quality" (p. 15), which would correspond to Kreiman and Gerratt's (1996) suggesion of using the method of analysis by synthesis. Pierrehumbert and Frisch found that of the three parameters, adjustment to a lower F0 alone produced the perception of a glottal stop, although adjustment of all three parameters did produce a more natural sounding stimulus. Thus, the study of laryngealization has been plagued by problems of reliable identification of voice qualities both using visual and auditory inspection of the signal and using automatic methods of perturbation analysis. Given these problems and some of the inherent problems in identifying voice qualities and automatically identifying non-stationarities in the signal, it seems that the most promising line of research on laryngealization in speech would be to use methods of analysis by synthesis, varying parameters suspected of creating the perception of laryngealization in order to try to match the resulting stimulus with unquestionable cases of laryngealization in natural tokens. The results of such research would be the identification of acoustic factors

which lead to the perception of laryngealization. If this method is successful, then the examination of changes in voice quality utterance-finally might also be pursued, but until laryngealization itself as a phenomenon is better understood, the use of laryngealization as a phonetic cue to discourse structure will not be understandable either.

The issue of resynthesis and analysis by synthesis also raises another point, namely, that resynthesis can be used as a method to test whether the acoustic cues discussed in section 3. 3 are indeed reliable. As Cutler and Pearson (1986) point out, the use of resynthesis to impose sets of acoustic features on otherwise identical utterances will create a range of controlled stimuli for perception tests. That is, it seemed like lowered RMS amplitude peaks in the final intermediate phrase of the utterance were one cue to the discourse-position of an utterance in the experiments described in section 3.3 If so, then it ought to be the case that raising the RMS amplitude peaks in the final intermediate phrase of discourse-final utterances and lowering the RMS amplitude peaks in the final intermediate phrase of discourse-medial utterances should cause a reversal of the listener judgements actually obtained if the resynthesized stimuli are presented to new listeners. Similarly with the finding that a lower F0 on the final pitch-accented syllable of the utterance marked an utterance as discourse-final. In this case, resynthesizing the utterances such that a discourse-final utterance's final pitch accented syllable had a raised F0 while a discourse-medial utterances' final pitch-accented syllable had a lowered F0 ought to cause a reversal of the listener judgements actually obtained if the resynthesized stimuli are presented to new listeners. Thus, resynthesis may be used as a tool to test the acoustic analysis by reversing the cues and testing to see whether listener judgements are reversed when listening to the resynthesized stimuli.

Also in the area of phonetic implementation is the question of not just what the acoustics are for the phonetic implementation of global structures in the intentional hierarchy but also, now that there is a bit more of a picture of what it is that people are

doing and when they are doing it, what the articulations are which produce the reduction of amplitude and the drop in F0 at the end of the sentence.

4.2.4 The Cross-Linguistic Perspective

The final area which deserves further exploration is the cross-linguistic perspective and an examination of what happens in typologically different languages. This study has examined only English, in which there is an inventory of pitch accent shapes which can be chosen from to convey different pragmatic meanings, and which associate with particular syllables in the utterance, but do not change the lexical meaning. In other languages, such as more densely specified tone languages in which lexical meanings can be changed by choice of tone but in which there is still a pragmatic intonational component, it would be interesting to investigate which aspects of discourse structure correspond to which parts of the intonational tune.

4.3 Conclusion

The goal of this dissertation was to examine the phonetic correlates of discourse structure without confounding the discourse function of the phonetic implementations with the meanings of the phonological edge tones. The ToBI transcription system (Beckman and Ayers, 1994), was used as the formal model of the intonational system. The Grosz and Sidner (1986) model of linguistic structure, intentional structure, and attentional structure was used as the formal model of the discourse structure. Using these models in designing the experiments and analyzing the data helped to avoid some of the pitfalls present in earlier work on the phonetic correlates of discourse structure, including phonetic comparisons across phonological categories and comparisons of utterances in different positions, but only within discourse segments and not across different hierarchical discourse structures.

Pierrehumbert and Hirschberg's (1990) interpretation of high edge tones as meaning that the utterance is to be interpreted with respect to a subsequent utterance was extended to mean that a high edge tone signals that more information needs to be added to

the attentional space before any information can be removed. (Similarly with the specific interpretation of a high phrase accent, which signals that the phrase in question does not form such a coherent unit with the subsequent phrase as it would if there were a low phrase accent.) Their interpretation of low edge tones as meaning that the utterance need not be interpreted with respect to a subsequent utterance was extended to mean that a low edge tone signals that the current focus space can be popped off of the focus stack. Thus, the functional categories of H vs. L edge tones have contrasting discourse meanings as well.

Furthermore, Hirschberg and Pierrehumbert's (1986) interpretation of the phonetic implementation of L edge tones ("final lowering," or contraction of the pitch range) as functioning to indicate how complete the utterance was in terms of its place in the discourse hierarchy was extended to mean that the phonetic implementation of phrases with low edge tones signal how deep into the focus stack the conversational participants can go when popping focus spaces.

Thus, the tone categories' discourse meanings (such as whether to push or pop focus spaces into or out of the focus stack) are to express the dominance relations which are part of the intentional hierarchy. These relations are "local" in the sense that they hold between two utterances, although they are not necessarity "local" in the sense of utterance adjacency in the linear order of utterances. On the other hand, the discourse functions of phonetic implementations of tones are to express whether all of the daughter DSPs of a dominant DSP have been satisfied, and hence how many DSPs can be simultaneously popped from the focus stack. (Intuitively, addressing the question of how big of a sub-topic has been closed.) These relations are "global" in the sense that they affect groups of DSPs, not necessarily just two DSPs. The issue of whether a set of daughter DSPs has been completely addressed cannot be expressed in the focus stack, because sisters are (presumably) never in the focus stack at the same time. This is an issue which can only be addressed by examining the intentional hierarchy as a whole, and seeing the overall

organization of dominance relations. Thus, the choice of phonological edge tone seems to express local discourse meanings while the way that the tones in the phrase governed by that edge tone are implemented seems to express more global discourse meanings.

This understanding of the phonological intonational edge tones and their relationship to the local discourse structure vs. the phonetic implementations and their relationship to the global discourse structure leads to two testable hypotheses regarding intonation and discourse structure. The first hypothesis (the "same edge tones hypothesis") is that local dominance relations in the discourse structure are expressed by the choice of edge tones. Thus, sentences with the same local dominance relations should have the same edge tones. Furthermore, if the DSPs are sisters and do not dominate other DSPs, then the edge tones for both sisters should be low. The second hypothesis (the "phonetic differences hypothesis") is that the amount of DSPs which can be popped from the focus stack is expressed by the phonetic implementation of the tones. Thus, if only the single utterance can be popped from the focus stack (because it has sister DSPs which have not yet been addressed), then its pitch accents should be implemented differently than the pitch accents in an utterance which allows both itself and its dominant DSP to be popped from the focus stack (because all of its sister DSPs have already been addressed).

These hypotheses were tested experimentally using a corpus in which discourse structure was tightly controlled. The corpus was recorded by native speakers, and the test sentences were extracted. This resulted in pairs of sentences which were identical lexically and which had the same dominance relations to the dominating DSP but which were from different locations in the intentional hierarchy. One of the sentences was earlier in the hierarchy, allowing only itself to be popped from the focus stack when it was uttered, while the other sentence was later in the hierarchy, allowing both itself and its dominant DSP to be popped from the focus stack. These test sentences were prosodically transcribed using the ToBI labelling system. Their edge tones were examined to see whether there were

any high edge tones present. In order to do the acoustic analysis, first the transcriber's tone similarity judgements were elicited and multi-dimensional scaling was used to produce a map of the transcriber's intuitive tone space. Within these maps, groups of tones which clustered together were sought with the understanding that those groups had similar tone targets and hence might be easily confusable, so should perhaps be considered to be similar tone targets and hence comparable phonologically. A perception test was run in which the test sentences were played in pairs to listeners to see if the listeners could identify the discourse position from which the sentences were extracted. And finally, the pairs of sentences which were reliably and correctly identified by listeners, had low edge tones in both members of the pair, and had either identical tone sequences or else similar tone targets (according to the tone similarity mapping) were acoustically analyzed. The acoustic analysis concentrated on RMS amplitude comparisons between utterances and F0 comparisons between utterances. Thus, the sentences were produced by subjects and recorded, intonational tones in the utterances were transcribed and compared, tone similarity was mapped out, a perception test was run, pairs of utterances which met certain criteria of identifiability and tone similarity were selected, and an acoustic analysis was done on the selected pairs. These were the steps that were used to experimentally test the two hypotheses.

The two hypotheses were borne out by the results of the experimental study. That is, there were very few examples of high edge tones at the ends of the utterances under consideration. Out of the 117 utterance pairs which were correctly and reliably identified by listeners in the perception test, there were 11 which had a high edge tone at the end of the discourse-medial utterances (either H- or H% or both), and hence could be interpreted as having the choice of tone category functioning as an acoustic cue to discourse position. Out of these 117 utterance pairs, there were 102 which had low edge tones at the end of both utterances. Within the utterances which were reliably and correctly identified by listeners,

had low edge tones, and had similar sequences of pitch accents (and hence could be compared, because they belonged to the same phonological category), there were tendencies towards patterned acoustic differences between the sentences in the pairs, which could be interpreted as having the phonetic implementation of the tones functioning as an acoustic cue to discourse position (as predicted by the hypotheses). The first tendency was that the RMS amplitude peaks for pitch-accented syllables tended to be higher for discourse-medial tokens than for the matched discourse-final tokens, beginning at different points in the utterance. The second trend was that the F0 (measured at the RMS amplitude peak) tended to be higher for discourse-medial tokens than for the corresponding discourse-final tokens, also beginning at different points in the utterance. The third tendency was that the F0 (measured at the F0 peak) for the first pitch accented syllable tended to be higher for discourse-medial tokens than for the corresponding discourse-final tokens, indicating the setting of a lower pitch range for the discourse final tokens. Thus, there were phonetic correlates of discourse structure present even when comparing only within phonological categories. Thus, the prediction about the same set of dominance relations in the intentional hierarchy being expressed with the same tones (and, in fact, with low edge tones) was borne out, as was the prediction that the different order of the dominance relations in the intentional hierarchy should be expressed through the phonetic implementation of the tones.

In conclusion, the relationship between local discourse structure and the choice of phonological tones was tested experimentally, as was the prediction about the relationship between global discourse structure and the phonetic implementation of tones. Both of these hypotheses were verified. The first hypothesis was verified by the predominant use of low edge tones and the rare use of high edge tones in the recorded utterancers, which served to mark that a discourse segment did not need to be interpreted with respect to a subsequent sister discourse segment when the two had no dominance relation between them. With

respect to the second hypothesis, some qualifications on the generality of the result are necessary. The discourse medial and discourse final utterances were often indistinguishable to listeners (which may be an artifact of the recording situation), and then when they were distinguishable as to discourse position, they often had different intonational tunes. But in some cases, the two utterances did have the same tunes and were still distinguishable as to their discourse position, indicating that the utterances were distinguished by differences in the phonetic implementation (phonetic correlates of discourse structure). The phonetic implementation thus could be used to mark elements of the global intentional structure, such as whether all of the sister discourse segments had been expressed yet. Thus, local aspects of the discourse structure (dominance relations in the intentional hierarchy) are conveyed by the phonological structure (the choice of tone) and global aspects of the discourse structure (whether sets of daughter discourse segments have been exhaustively completed or not) are conveyed by the phonetic structure (different phonetic implementations of the same tune). This study thus synthesizes results from phonemic studies of intonational meaning with results from phonetic studies of discourse markers.

LIST OF REFERENCES

Ayers, Gayle M. (1996) *Nuclear Accent Types and Prominence: Some Psycholinguistic Experiment*s. PhD Dissertation. Ohio State University.

Bard, Ellen Gurman, Dan Robertson, and Antonella Sorace. (1996) Magnitude Estimation of Linguistic Acceptability. *Language* **72** pp. 32-68.

Beckman, Mary E. and Gayle M. Ayers (1994) Guidelines for ToBI Labelling.

Beckman, Mary E., Jan Edwards, and Janet Fletcher. (1992) Prosodic Structure and Tempo in a Sonority Model of Articulatory Dynamics. In G.J. Docherty and D.R.Ladd (eds.) *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*. Cambridge University Press: Cambridge.

Beckman, Mary E. and Janet B. Pierrehumbert. (1986) Intonational Structure in Japanese and English. *Phonology Yearbook* **3** pp. 255-309.

Behrman, Alison and R.J. Baken. (1997) Correlation Dimension of Electroglottographic Data from Healthy and Pathologic Subjects. *Journal of the Acoustical Society of America* **102** pp. 2371-2379.

Bielamowicz, Steven, Jody Kreiman, Bruce R. Gerratt, Marc S. Dauer, and Gerald S. Berke. (1996). Comparison of Voice Analysis Systems for Perturbation Measurement. *Journal of Speech and Hearing Research* **39** pp. 126-134.

Bolinger, Dwight. (1958) A Theory of Pitch Accent in English. *Word* **14** pp. 109-149.

Boyce, Suzanne and Lise Menn. (1979) Peaks Vary, Endpoints Don't: Implications for Intonation Theory. *Proceedings of the Fifth Annual Meeting of the Berkeley Linguistics Society*.

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

Cutler, Anne and Mark Pearson. (1986) On the Analysis of Prosodic Turn-Taking Cues. In Catherine Johns-Lewis (ed.) *Intonation in Discourse.* Croom Helm: London.

Grosz, Barbara and Julia Hirschberg. (1992) Some Intonational Characteristics of Discourse Structure. *Proceedings of the 2nd International Conference on Spoken Language Processing*. Banff, Canada. pp. 429-432.

Grosz, Barbara J. and Candace L. Sidner. (1986) Attentions, Intentions, and the Structure of Discourse. *Computational Linguistics* **12** pp. 175-204.

Gussenhoven, Carlos. (1984) *On the Grammar and Semantics of Sentence Accents*. Foris: Dordrecht.

Gussenhoven, Carlos, and Toni Rietveld. (1988). An Experimental Evaluation of Two Nuclear-Tone Taxonomies. *Linguistics* **29** pp. 423-449.

Hagen, Astrid. (1997) *Linguistic Functions of Glottalizations and their Language Specific Use in English and German*. PhD Dissertation. Friedrich-Alexander-Universitat Erlangen-Nurnberg and MIT.

Halliday, M.A.K. (1967) *Intonation and Grammar in British English*. Mouton: The Hague.

Henton, Caroline and Anthony Bladon. (1988) Creak as a Sociophonetic Marker. In Larry M. Hyman and Charles N. Li (eds.) *Language, Speech and Mind*. Routledge: London.

Herman, Rebecca. (1996) Final Lowering in Kipare. *Phonology*. **13** pp. 171-196.

Herman, Rebecca, Mary Beckman and Kiyoshi Honda. (1996). Subglottal Pressure and Final Lowering in English. In H. Timothy Bunnell and William Idsadi (eds). *Proceedings of the Fourth International Conference on Spoken Language Processing.* Philadelphia. pp. 145-148.

Hirschberg, Julia and Christine H. Nakatani. (1996) A Prosodic Analysis of Discourse Segments in Direction-Giving Monologues. *ACL*.

Hirschberg, Julia and Gregory Ward. (1992) The Influence of Pitch Range, Duration, Amplitude and Spectral Features on the Interpretation of the Rise-Fall-Rise Intonation Contour in English. *Journal of Phonetics* **20** pp. 241-251.

Jannedy, Stefanie. (1997) Acquisition of Narrow Focus Prosody. In Antonella Sorace, Caroline Heycock, and Richard Shillcock (eds.) *Proceedings of the GALA '97 Conference on Language Acquisition.* Edinburgh. pp. 245-250.

Kingdon, Roger. (1958) *The Groundwork of English Intonation*. Longman: London.

Kreiman, Jody. (1982) Perception of Sentence and Paragraph Boundaries in Natural Conversation. *Journal of Phonetics* **10**, pp. 163-175.

Kreiman, Jody and Bruce R. Gerratt. (1996) The Perceptual Structure of Pathological Voice Quality. *Journal of the Acoustical Society of America* **100** pp. 1787-1795.

Ladd, D. Robert (1996) *Intonational Phonology*. Cambridge University Press: Cambridge.

Ladd, D. Robert and Jacques Terken. (1995) Modelling Intra- and Inter-Speaker Pitch Range Variation. In Kjell Elenius and Peter Branderud (eds.) *Proceedings of the International Conference of Phonetic Sciences*. volume 2, pp. 386-389.

Lehiste, Ilse. (1979) Perception of Sentence and Paragraph Boundaries. In B. Lindblom and S. Ohman (eds.) *Frontiers of Speech Communication Research.* Academic Press: London.

Lehiste, Ilse. (1975) The Phonetic Structure of Paragraphs. In A. Cohen and S.G. Nooteboom (eds.) *Structure and Process in Speech Perception*. Springer-Verlag: New York.

Lehiste, Ilse and William S-Y Wang. (1977) Perception of Sentence Boundaries With and Without Semantic Information. *Phonologica 1976.* pp. 277-283.

Liberman, Mark and Janet Pierrehumbert. (1984) Intonational Invariance Under Changes in Pitch Range and Length. In M. Aronoff and R. Oerhle (eds.) *Language Sound Structure*. MIT Press: Cambridge.

Liberman, M. and I. Sag. (1974) Prosodic Form and Discourse Function. *Papers of the Tenth Regional Meeting of the Chicago Linguistic Society* . Chicago. pp. 416-427.

Maeda, Shinji. (1976) *A Characterization of American English Intonation*. PhD Dissertation. MIT.

Nakatani, Christine. (1997) Discourse Structural Constraints on Accent in Narrative. In Jan P.H. van Santen, Richard W. Sproat, Joseph P. Olive, and Julia Hirschberg (eds.) *Progress in Speech Synthesis*. Springer: New York.

Nakatani, Christine. (1996) Integrating Prosodic and Discourse Modelling. In Yoshinori Sagisaka, Nick Campbell, and Norio Higuchi (eds.) *Computing Prosody*. Springer: New York.

O'Connor, J. D. and G. F. Arnold. (1973) *Intonation of Colloquial English*. 2nd ed. Longman: London.

Palmer, Harold. (1922) *English Intonation, with Systematic Exercises*. Heffer: Cambridge.

Pierrehumbert, Janet. (1994) Prosodic Effects on Glottal Allophones. In Osamu Fujimura and Minoru Hirano (eds.) *Vocal Fold Physiology: Voice Quality Control*. Singular Publishing Group, Inc.: San Diego.

Pierrehumbert, Janet. (1990) Phonological and Phonetic Representation. *Journal of Phonetics* **18** pp. 375-394.

Pierrehumbert, Janet. (1980) *The Phonology and Phonetics of English Intonation.* PhD Dissertation. MIT

Pierrehumbert, Janet B. and Stefan Frisch. (1997) Synthesizing Allophonic Glottalization. In Jan P.H. van Santen, Richard W. Sproat, Joseph P. Olive, and Julia Hirschberg (eds.) *Progress in Speech Synthesis* . Springer: New York.

Pierrehumbert, Janet and Julia Hirschberg. (1990) The Meaning of Intonational Contours in the Interpretation of Discourse. In Philip R. Cohen, Jerry Morgan, and Martha E. Pollack (eds.) *Intentions in Communication*. The MIT Press: Cambridge, MA.

Pitrelli, John F., Mary E. Beckman, and Julia Hirschberg (1994) Evaluation of Prosodic Transcription Labeling Reliability in the ToBI Framework. *Proceedings of the 3rd International Conference on Spoken Language Processing*. Yokohama. pp. 123-126

Roberts, Craige. (1998) The Place of Centering in a General Theory of Anaphora Resolution. In Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince (eds.) *Centering Theory in Discourse*. Clarendon Press: Oxford.

Roberts, Craige. (1996) Information Structure in Discourse: Towards an Integrated Formal Theory of Pragmatics. In Jae-Hak Yoon and Andreas Kathol (eds.) *Papers in Semantics*. The Ohio State University Working Papers in Linguistics, no. 49. pp. 91-136.

Schaffer, Deborah. (1984) The Role of Intonation as a Cue to Topic Management in Conversation. *Journal of Phonetics* **12** pp. 327-344.

Schaffer, Deborah. (1983) The Role of Intonation as a Cue to Turn Taking in Conversation. *Journal of Phonetics* **11** pp. 243-257.

Stephens, Jane and Geoffrey Beattie. (1986) On Judging the Ends of Speaker Turns in Conversation. *Journal of Language and Social Psychology*. **5** pp. 119-134.

Titze, Ingo. (1994) *Principles of Voice Production*. Prentice Hall: Englewood Cliffs, New Jersey.

Walker, Marilyn, Aravind Joshi, and Ellen Prince. (1998) Centering in Naturally Occurring Discourse: An Overview. In Marilyn A. Walker, Aravind K. Joshi, and Ellen F. Prince (eds.) *Centering Theory in Discourse*. Clarendon Press: Oxford.