Prosodic Detail
and Topic Structure
in Discourse

Margaret Kendall Zellers
Clare College

A dissertation submitted for the degree of
Doctor of Philosophy 2011
Declaration I

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as specified in the text.

Margaret K. Zellers
Clare College

13 May 2011
Declaration II

This dissertation does not exceed the regulation length, including footnotes, references and appendices but excluding the bibliography.

Margaret K. Zellers
Clare College

13 May 2011
Abstract

The research presented in this thesis addresses the role of prosody in signalling the topic structure of longer stretches of discourse in Standard Southern British English (SSBE), from the point of view of both the speaker and the listener. My research makes several unique contributions to the investigation of prosody and discourse topic. First, I investigate prosodic signalling of topic structure taking into account a more complex model of this structure than most previous phonetic studies of topic had used, as well as taking a phonologically-informed view of the phonetic variation aspects. Second, I address a neglect of non-intonational prosodic cues in the context of discourse structure signalling. Third, I make a first experimental attempt to investigate the perception of topic structure cues by listeners in an on-line processing situation rather than by asking metalinguistic questions.

A production experiment investigated the ways in which speakers of SSBE vary their prosody in relation to the topic structure of long discourses. A written text which was controlled for the topic structure and for segmental characteristics was read aloud, and the recordings were analyzed with regard to the topic structure. The well-known cues of pitch reset (i.e. a very high pitch at the beginning of a new unit) and supradeclination (gradual lowering of pitch across a unit) were found to correlate with the global grouping of utterances into topic units. The size of the first (prenuclear) F0 fall was found to vary systematically with the internal topic structure of the topic groups. A number of non-F0 cues gave further evidence for this idea, notably the variation in local speech rate at the beginnings of utterances, which appeared in different degrees relative to the topic structure category of the utterance, and which for some speakers appeared to be used instead of the F0 fall cue. Variation in global speech rate and in the distribution of aperiodicity, as well as the distribution of rising versus falling initial pitch accents, were also found to show reliable patterns related to the different topic structure categories. A post-hoc validation investigated the usefulness of the categories adopted for the text and the degree of individual variation in how texts are organized into topics.

A perception experiment was used to further investigate the degree to which listeners make use of information about topic structure during the immediate or on-line process of listening. Subjects heard a series of utterances which had had their prosody manipulated to suggest that they contained a topic change or a topic hold, and then had to respond to following sentences which were either consistent or not with the prosody of the utterances they had heard. If the prosody and the semantic content did not match, listeners were much slower to accept the following utterance as fitting the context. This indicates that listeners are sensitive to information about the topic structure of utterances as part of the natural process of listening, and not only when asked to complete metalinguistic tasks involving marking groups or identifying topics.

The robust production results in particular suggest that topic structure marking is integral to the process by which speakers and listeners communicate. Although the set of cues investigated is far from exhaustive, both local and global prosodic characteristics of utterances provide information that the listener can and apparently does take advantage of. The occurrence of many of these variations at linguistically relevant points suggests that they are not simply side effects of biological imperatives (cf. the Biological Codes), but properly part of the linguistic system, modifying the output of the intonational phonology to provide additional cues layered on top of the already rich phrase- and utterance-level prosodic signalling. This has important implications for tonal and intonational research, in that this additional level of variation must be taken into account in investigations of how the tonal or intonational phonology is structured.
Contents

Acknowledgments 13

1 Introduction 15
  1.1 Some current issues in prosody research . . . . . . . . . . . . . . . . . . . 15
  1.2 Goals of this thesis . . . . . . . . . . . . . . . . . . . . . . . . . . . . 18
  1.3 Organization . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 19

2 Literature Review 20
  2.1 Prosody and the Speech Signal . . . . . . . . . . . . . . . . . . . . . . . . 20
    2.1.1 A complex relationship . . . . . . . . . . . . . . . . . . . . . . . . 20
      2.1.1.1 F0 and pitch . . . . . . . . . . . . . . . . . . . . . . . 21
      2.1.1.2 Other prosodic characteristics . . . . . . . . . . . . . . 26
  2.2 Intonational Phonology . . . . . . . . . . . . . . . . . . . . . . . . . . . 29
    2.2.1 Motivation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 29
    2.2.2 Accounting for the signal . . . . . . . . . . . . . . . . . . . . . . . 30
    2.2.3 Modes of description . . . . . . . . . . . . . . . . . . . . . . . . . 31
      2.2.3.1 British school . . . . . . . . . . . . . . . . . . . . . . . 32
      2.2.3.2 IPO . . . . . . . . . . . . . . . . . . . . . . . . . . . . 33
      2.2.3.3 KIM . . . . . . . . . . . . . . . . . . . . . . . . . . . 34
      2.2.3.4 Autosegmental approaches and ToBI . . . . . . . . . . . 35
    2.2.4 Linking phonetics and phonology: alignment . . . . . . . . . . . . . 40
    2.2.5 Intonational meaning . . . . . . . . . . . . . . . . . . . . . . . . . 49
      2.2.5.1 Form and function . . . . . . . . . . . . . . . . . . . . . 49
      2.2.5.2 Semantic value . . . . . . . . . . . . . . . . . . . . . . . 52
CONTENTS

2.3 Discourse Topic Structure .............................................. 54
  2.3.1 Definition of topic ............................................. 54
  2.3.2 Theories of topic structure ..................................... 55
    2.3.2.1 Hierarchical theories .................................... 56
    2.3.2.2 Categorical theories ..................................... 59
    2.3.2.3 Topic in narratives ...................................... 64
  2.3.3 Prosody and topic structure ..................................... 65
    2.3.3.1 Boundary cues ............................................ 66
    2.3.3.2 Internal cues ............................................. 69
  2.4 Implications and Research Aims .................................... 72

3 F0 Cues to Discourse Topic Structure ................................ 75
  3.1 Introduction ....................................................... 75
    3.1.1 Intonation and discourse structure ......................... 75
  3.2 Production Experiment ............................................. 76
    3.2.1 Motivation .................................................... 76
    3.2.2 Materials ..................................................... 79
    3.2.3 Hypotheses .................................................... 84
    3.2.4 Recordings ..................................................... 85
    3.2.5 Analysis ....................................................... 85
      3.2.5.1 F0 peak height .......................................... 87
      3.2.5.2 F0 fall span, length, and slope ....................... 87
      3.2.5.3 F0 peak alignment ...................................... 88
    3.2.6 Results ....................................................... 88
      3.2.6.1 F0 peak height .......................................... 88
      3.2.6.2 F0 fall range ............................................ 89
      3.2.6.3 F0 fall slope ............................................ 91
      3.2.6.4 F0 peak alignment ...................................... 92
  3.3 Discussion ........................................................ 92
    3.3.1 F0 peak height .............................................. 92
    3.3.2 F0 fall span and slope ..................................... 94
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3.3</td>
<td>F0 peak timing</td>
<td>95</td>
</tr>
<tr>
<td>3.3.4</td>
<td>Individual differences</td>
<td>95</td>
</tr>
<tr>
<td>3.3.5</td>
<td>General discussion</td>
<td>96</td>
</tr>
<tr>
<td>3.4</td>
<td>Conclusion</td>
<td>99</td>
</tr>
<tr>
<td>4</td>
<td>F0 Peak Timing and Topic</td>
<td>100</td>
</tr>
<tr>
<td>4.1</td>
<td>Introduction</td>
<td>100</td>
</tr>
<tr>
<td>4.2</td>
<td>Phonological modelling</td>
<td>102</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Motivation</td>
<td>102</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Reanalysis</td>
<td>102</td>
</tr>
<tr>
<td>4.2.3</td>
<td>Results and Discussion</td>
<td>107</td>
</tr>
<tr>
<td>4.2.3.1</td>
<td>Single pitch accent analysis</td>
<td>107</td>
</tr>
<tr>
<td>4.2.3.2</td>
<td>Two pitch accent analysis</td>
<td>110</td>
</tr>
<tr>
<td>4.3</td>
<td>Discussion</td>
<td>112</td>
</tr>
<tr>
<td>4.4</td>
<td>Conclusion</td>
<td>113</td>
</tr>
<tr>
<td>5</td>
<td>Other Prosodic Cues to Topic: Speech Rate</td>
<td>114</td>
</tr>
<tr>
<td>5.1</td>
<td>Introduction</td>
<td>114</td>
</tr>
<tr>
<td>5.2</td>
<td>Initial Observations</td>
<td>116</td>
</tr>
<tr>
<td>5.3</td>
<td>Speech Rate: Word Length</td>
<td>118</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Methodology</td>
<td>119</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Results: Word Length/Local Speech Rate</td>
<td>120</td>
</tr>
<tr>
<td>5.4</td>
<td>Speech Rate: Syllable Length in Sentences</td>
<td>126</td>
</tr>
<tr>
<td>5.4.1</td>
<td>Methodology</td>
<td>126</td>
</tr>
<tr>
<td>5.4.2</td>
<td>Results: Average Syllable Duration/Global Speech Rate</td>
<td>127</td>
</tr>
<tr>
<td>5.5</td>
<td>Discussion</td>
<td>127</td>
</tr>
<tr>
<td>6</td>
<td>Other Prosodic Cues to Topic: Phonation Quality</td>
<td>130</td>
</tr>
<tr>
<td>6.1</td>
<td>Introduction</td>
<td>130</td>
</tr>
<tr>
<td>6.1.1</td>
<td>Hypotheses</td>
<td>132</td>
</tr>
<tr>
<td>6.2</td>
<td>Aperiodicity and Topic Structure</td>
<td>133</td>
</tr>
<tr>
<td>6.2.1</td>
<td>Methodology</td>
<td>133</td>
</tr>
</tbody>
</table>
### CONTENTS

6.2.2 Results .................................................. 134
  6.2.2.1 Occurrence of aperiodicity at boundaries .......... 134
  6.2.2.2 Length of aperiodic stretches at boundaries .... 135
  6.2.2.3 Discussion ............................................ 137

6.2.3 Alternative boundary characteristics: aperiodicity and syntax .... 138

6.3 Aperiodicity and F0 ........................................ 140
  6.3.1 Analysis and results .................................... 140
  6.3.2 Discussion ............................................... 141

6.4 General discussion: non-F0 prosodic characteristics ............... 142

6.5 Conclusion .................................................. 145

7 Validation of Topic Structure Categories .......................... 146
  7.1 Topic structure in different types of discourse .......... 146
  7.2 Motivation .................................................. 148
  7.3 Validation Part 1: Topic Boundaries ....................... 150
    7.3.1 Methodology .......................................... 150
    7.3.2 Results ............................................... 152
      7.3.2.1 Version 1: line-by-line topics .................. 152
      7.3.2.2 Version 2: long-form topic ..................... 155
    7.3.3 Prosodic characteristics at validated topic boundaries .... 156
      7.3.3.1 Results: F0 analyses ........................... 157
      7.3.3.2 Results: local speech rate ..................... 160
  7.4 Validation Part 2: Topic Structure Categories .............. 161
    7.4.1 Methodology .......................................... 161
    7.4.2 Results ............................................... 162
    7.4.3 Discussion ............................................. 165
  7.5 General Discussion ........................................ 167

8 On-line Perception of Topic Structure Prosody .................... 170
  8.1 Introduction .............................................. 170
    8.1.1 Literature review .................................... 171
      8.1.1.1 Metalinguistic studies .......................... 171
CONTENTS

8.7 Conclusion .............................................. 208

9 General Discussion ........................................ 209

9.1 Review of results ....................................... 209

9.2 Implications for topic ................................... 212

9.2.1 Prosody and existing theories of topic .......... 212

9.2.2 Topic and discourse purpose ...................... 214

9.2.2.1 Discourse purpose, episode shifts, and individual variation in production .......... 216

9.2.2.2 Discourse purpose and perception .......... 218

9.2.2.3 Re-examining “topic structure” .............. 218

9.3 Implications for prosody .............................. 220

9.3.1 Linguistic structure ................................ 221

9.3.2 Multiple cues ....................................... 224

9.4 Implications for methodology ...................... 229

9.4.1 Production study .................................. 230

9.4.1.1 Experimental control ....................... 230

9.4.1.2 Phonologically-aware analysis ............ 233

9.4.2 Perception study .................................. 235

9.5 Conclusion .............................................. 237

Works Cited .................................................. 239

Appendix I: Emory Story .................................. 260

Appendix II: Perception Experiment Stimuli ......... 266

Appendix III: Sound Files ................................. 272
## List of Figures

3.1 F0 peak height above speaker baseline by position in utterance group . . . 89
3.2 Size of F0 falls by topic condition . . . . . . . . . . . . . . . . . . . . . . . 90
3.3 F0 fall range by level of embedding . . . . . . . . . . . . . . . . . . . . . . . 90
3.4 F0 peak alignment in Anacrusis (A) and Initial (I) conditions . . . . . . . 91

4.1 Schematics of two (prenuclear) pitch-accent models. (a) Two-accent model, 
in which early and late peaks form two categories. (b) One-accent model, 
in which all peaks are considered one category . . . . . . . . . . . . . . . . . 102
4.2 Schematic of (a) rising pitch accent and (b) falling pitch accent in relation 
to 3 syllables of target word with initial stress . . . . . . . . . . . . . . . . . 103
4.3 Typical instances of a prenuclear rise (top) and a fall (bottom). . . . . . . 105
4.4 A problematic case which was ambiguous between a rise and a fall (this 
item was ultimately discarded from the pitch accent analysis). . . . . . . . 106
4.5 PC1 for the Emory data, showing typical shapes for contours with a positive 
PC1 value (++) or a negative PC1 value (–). The mean of all 126 contours 
is shown with a solid line . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 106
4.6 Distribution of human labels along PC1 axis. Red = labelled as falls by 
human labeller. Blue = labelled as rises by human labeller . . . . . . . . . . . 107
4.7 F0 peak timing by topic condition in the single-accent model . . . . . . . 108
4.8 Timing of F0 peaks in utterances with and without an anacrusis in the 
single-accent model (anacrusis in left panel) . . . . . . . . . . . . . . . . . 109
4.9 Pitch accent by topic structure category in the two-accent model . . . . . . 110
LIST OF FIGURES

4.10 Pitch accent group distribution with and without anacrusis in the two-accent model .................................................. 111

5.1 F0 fall rate data for (a) speaker F05, and (b) speaker F04 .................. 116

5.2 Syllable length by interval number; the interval number represents the sequential location of the syllable in the overall recording. .................. 120

5.3 Word length by topic structure category for (a) speaker F05, and (b) speaker F04 .................................................. 121

5.4 Mean fall range and word length for the four topic structure categories, aggregated by speaker. Each different symbol/color combination represents one speaker. .................................................. 123

5.5 F0 fall range and word length, aggregated across speakers, marked with topic structure category .................................................. 124

6.1 Presence or absence of aperiodicity preceding different topic structure categories. .................................................. 135

6.2 Length of aperiodic stretches preceding topic structure categories .... 136

6.3 a. Presence or absence of aperiodic stretches across categories (in percent of preceding sentence tokens with aperiodicity). b. Length of aperiodic stretches across categories. .................................................. 139

6.4 Length of aperiodicity by pitch movement .................................................. 141

7.1 Screenshots of topic boundary task in (a) line-by-line format and (b) long form. .................................................. 151

7.2 (a) Clustering of line-by-line labellers in a hierarchical cluster analysis. (b) Results of a multidimensional scaling analysis of line-by-line labellers. Labeller 16 was excluded from the final analysis due to having performed drastically differently from all other labellers on the task, as these analyses show, while labellers 2 and 6 gave (apparently) almost identical characterizations of the text. .................................................. 154
LIST OF FIGURES
7.3

10

F0 peak alignment (delay ratio) by topic structure category. A = anacrusis, I
= initial (no anacrusis) (a) shows the validated new Topics compared to the
other categories (with no changes); (b) shows the validated new Topics (T)
compared to the other items formerly labelled new Topics (M = mid). . . . 158

7.4

Distribution of F0 fall range using only validated topic boundary utterances
for the new Topic condition. (Compare with figure 3.2.) . . . . . . . . . . . 159

7.5

Speech rate (in word length) across the topic structure categories, with only
validated new Topics included in that category. (Compare with figure 5.3.) . 160

7.6

Screenshot of topic structure categories task . . . . . . . . . . . . . . . . . 162

7.7

Plots of multidimensional scaling analysis for category data. (a) With all
four category labels. (b) With Additions and Elaborations combined into
one category. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 164

8.1

Response times for “acceptable” ratings by match or mismatch with prosody
in pilot study. (Statistics not reported due to small number of participants.) . 183

8.2

Schematic of “high” and “low” topic-change pitch contours (magenta and
pink, respectively) compared to topic-hold pitch contour (black). . . . . . . 186

8.3

Procedure for perception experiment . . . . . . . . . . . . . . . . . . . . . 189

8.4

Pause detection latencies by subject. Dark grey bars show items with topicchange prosody, light grey bars show items with topic-hold prosody. . . . . 193

8.5

Pause detection latencies in topic-change (T=Topic) and topic-hold (C=Continuation)
intonation conditions by item . . . . . . . . . . . . . . . . . . . . . . . . . 194

8.6

Acceptability of sentences in conditions where they matched and mismatched
the prosody of the previous utterance . . . . . . . . . . . . . . . . . . . . . 195

8.7

Effect of match or mismatch of prosody on speed of acceptance of followup sentences . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 196


List of Tables

3.1 Most common target words for each segmental condition ............... 83
3.2 Experimental design: condition combinations, with topic structure condi-
tions in columns, and sentence position conditions in rows ................. 84
5.1 Combination of F0 fall and local speech rate cues to differentiate four topic
structure categories ................................................................. 122
5.2 Mean fall range and word length by category for each group of speakers
(shown with ID of representative speaker). ‘*’ indicates that a category dif-
fers significantly from the other categories (within the speaker group) to the
p<0.5 level; ‘˙’ indicates a marginally significant difference, 0.05<p<0.08
in all cases. ................................................................. 125
5.3 Fixed effects from linear mixed model (ASD) ................................. 127
6.1 Syntactic categories of Continuations with examples (Continuations shown
in italics). ................................................................. 138
7.1 Confusion matrix showing majority-voted categories (down) compared to
original labels (across). ................................................................. 163
8.1 Mean response value for pre-tested items; df=34 in all cases. C = topic hold
(i.e. Continuation); T = topic change (i.e. new Topic). .......................... 181
8.2 Sentence rating responses in pilot study (statistics not reported due to small
number of participants). ................................................................. 183
8.3 Example stimuli for perception experiment (full stimulus list provided in
Appendix II). Italics indicate the location of the prosodic manipulation. . . 185
8.4 Linear mixed-effects model for acceptable sentence response times. . . . . 197
8.5 Linear model of memory task . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 198
Acknowledgments

Just as “it takes a village to raise a child,” it takes one to complete a PhD. This dissertation would not have been possible without the help and encouragement of many people.

First and foremost I want to thank my supervisor, Brechtje Post, for her endless support and unwavering confidence in me. I could not have had a better supervisor!

Many of the ideas in this thesis grew, directly or indirectly, out of conversations I had with a variety of other people. Mariapaola D’Imperio provided excellent input on my mess of production data, particularly during my stay in Aix, and her challenges to my assumptions led to the F0 peak timing analysis. Sven Mattys gave me invaluable help with the design of the perception experiment as well as with diagnosing some of its problems. Richard Ogden taught me just about everything I know about CA, and Rein Ove Sikveland fielded a constant barrage of questions about how to interpret it all. Oliver Niebuhr was always ready to have another discussion of the “big picture,” or just to have ideas bounced off of him. Francesco Cangemi was a foil to me and my ideas, in the best sense. Statistics have never been my strong point, and the accuracy of the analyses in this thesis is thanks to the input of Chris Cummins, Michele Gubian, Antje Heinrich, and Annett Jorshick. Rolf Carlson, Mirjam Ernestus, Carlos Gussenhoven, Sarah Hawkins, David House, Jacques Koreman, Bob Ladd, Francis Nolan, Radek Skarnitzl, Jan Volín and John Williams all provided helpful comments on a variety of interpretative aspects.

A host of other people also contributed on a more day-to-day level, particularly as I was in the final throes of completing the thesis. I would have starved many times over without the constant streams of food, encouragement and distraction provided by Naomi Hilton and Caroline Williams. Naomi, besides sometimes apparently holding the administrative end of Sound to Sense together by her teeth, was always there with a listening ear during some of
the really tough times I had. Caroline kept me moving even when I wanted to grind to a halt, and somehow seemed to think I was doing her a favor by turning up at her house and letting her cook dinner for me. My PhD-room companions past and present, particularly Cat Davies, Jeff Hanna, Danny Weston, Bettina Beinhoff, Elaine Schmidt, Calbert Graham and Ena Hodzik, helped keep me going on an everyday basis, and were always ready to head out to the pub and try to talk about something other than linguistics. Marco Piccolino-Boniforti provided various assistance ranging from a quick Python script to a 700-mile lift back to Cambridge. Ruth Cumming inspired me with her beautifully-designed and flawlessly-presented research, not to mention her rapid PhD completion, and took the time to proofread part of my thesis in an amazingly meticulous manner. (Naomi, Caroline and Bettina between them proofread the balance of what was left.) The Roberts, Vanderburgh, and Rerecich families provided my homes-away-from-home back in the USA, and reminded me that there’s life outside of Cambridge.

I was fortunate in having my PhD funded by the EC Marie Curie Research Training Network Sound to Sense (MRTN-CT-2006-035561). S2S provided more than just financial support; it also gave me a sense of camaraderie on the PhD journey and a welcome into the research community. The friendship of many of the S2S fellows, particularly Barbara Schuppler, Bogdan Ludusan, and Olesya Rauch, among the others already mentioned, was a major source of emotional support.

There are many people whom I haven’t named here but who nonetheless have enriched my life with their love and support, for which I can only say, “Thank you again!”
Chapter 1

Introduction

1.1 Some current issues in prosody research

Over the last half-century, the study of prosody and especially intonation has expanded considerably from early work by e.g. Palmer (1922), Armstrong & Ward (1926), and Wells (1945) in the British school, and Bloomfield (1933) and Pike (1945) in the American. While intonation was once considered to be an expression of speaker emotion without direct reference to the linguistic system, we now know that many aspects of intonation are structured in consistent and apparently arbitrary ways, just as other aspects of language are. The study of intonational structure and meaning has become an important part of the development of phonological theory.

It is clear that not all aspects of intonation are linguistically structured. Although we observe consistent patterns in the signal which relate to pragmatic meanings (specifically how this works is still a matter of debate), many kinds of variation are better explained by direct reference to non-linguistic aspects of communication. On one hand, a great deal of prosodic variation between speakers can be explained by physiological means; a speaker with a larger larynx (e.g. an adult man) has a lower-pitched voice than a speaker with a smaller larynx (e.g. an adult woman or child). This kind of variation is not linguistically meaningful and can only be considered communicative in the sense that it provides information about the speaker, which is not useful in all contexts (although it might be helpful for example in a phone conversation when visual or other identifying information is unavailable). Prosody can also provide information about a speaker’s emotional state; speakers who are excited
or angry (i.e. in a state of arousal) are likely to make use of larger pitch movements and higher overall intensity than speakers who are calm or even bored. While this information is normally overlaid onto the linguistic signal, it is not generally thought to be a part of the linguistic structure.

On the other hand, there are aspects of intonation which have been clearly shown to belong to individual linguistic systems, generally having the function of changing the pragmatic interpretation of utterances. One obvious use is the meaning change of an utterance from a statement to a question by varying the intonation, for instance by changing the final pitch from low to high, as is commonly observed in English and a variety of other languages. A somewhat more subtle linguistic use makes the same question/statement distinction by shifting the location of a pitch peak from early to late in the vowel in Neapolitan Italian (D’Imperio 2000). Aside from a consistently-recognizable change in form mapping to a specific function, this kind of change can also be seen to be linguistic because the intonational variation signals a meaning change in some languages that is achieved by other means in different languages. For example, in English, syntactic manipulations such as changes in word order, the adoption of the auxiliary “do”, or the filling of a content slot in the sentence with a wh- word, all achieve the meaning transition from statement to question. In some cases this meaning transition also involves intonational change (as in yes-no questions), but this is not obligatorily the case (as in some wh-questions). Similarly, in Japanese the shift from statement to question can be achieved by adding the question marker “-ka” to the end of the verb. A variety of other meaning variations, not limited to the question-statement shift, can be accomplished by intonation as well as by other linguistic means (cf. Gussenhoven 2007).

Although it is clear that there are some meanings of intonation which are linguistically structured and some which are not, it is not always obvious where the boundary between them is to be found. One example of this may be found in the Kiel Intonation Model (KIM) for German (Kohler 1991). The KIM identifies three pitch peak categories which German listeners are able to distinguish. The early peak and medial peak have linguistic/informational meanings, but the late peak category is associated with the meaning “astonishment”, which seems more obviously to be an emotional meaning. Despite this, German listeners consistently categorize this pitch configuration as separate from the other
pitch peak configurations, not as a modification of those categories, as might normally be expected for an emotional meaning (Kohler 1987, 1991).

The question of the boundary between linguistic and non-linguistic kinds of intonational variation is therefore still very much an open one. In part it is difficult to address because there seems to be a great deal of iconicity in intonational representation, even when the forms have apparently been grammaticalized into the language. One area in which the existence of iconicity has continued to be a source of confusion is in intonational variation in long units of speech (i.e. longer than a single utterance). It is commonly known that pitch reset is a characteristic of prosodic boundaries; an utterance ends on a low pitch and the following one begins again on a high pitch. The same is true for longer units of speech; the spoken equivalent of a paragraph will end low, and an extra-high pitch reset indicates the beginning of a new long group (Lehiste 1975). In one sense, this could be seen to have physiological origins, as predicted by e.g. the Production Code (Gussenhoven 2004): the greater availability of energy at the beginning of a breath group would allow the pitch to be higher (and possibly the amplitude as well). However, this explanation is incomplete, since speakers’ breath groups do not always match up with utterances, and for paragraph-length units it would be nearly impossible to produce the whole unit over the course of one breath. Therefore it appears that these cues have been grammaticalized into the linguistic system. However, the high correlation between the biological explanation and the characteristics of the signal makes it somewhat difficult to defend this cue as being abstract rather than having an inherent tie between form and function.

The body of research on the prosody of long units and particularly of topic structure has provided an increasingly converging picture of intonation in discourse. However, the picture is still incomplete in the sense that some of the more detailed aspects of the discourse theories have not been investigated in concert with some of the more detailed aspects of the intonational phonology and phonetics. One reason for this is that there has tended to be a lack of balance in the theoretical approaches adopted. Research on the prosody of discourse structure has often been undertaken either from a discourse analysis point of view with a limited or unsophisticated view of the phonetic aspects, or from a phonetic or phonological point of view with a limited or unsophisticated view of the discourse aspects. While it is possible to confidently make the assertion that there is intonational variation in relation to
topic structure, and with more complexity than simply grouping, it is not necessarily clear what the best cues to this structure are. Also, it is difficult to say whether it would be better to characterize these kinds of variation as linguistic or non-linguistic; that is, whether they show some degree of abstraction or whether they originate naturally from physiological states (e.g. as proposed by the Biological Codes).

Intonational research has tended to center around fundamental frequency (F0) while often neglecting other prosodic aspects of the speech signal: amplitude/intensity, phonation quality/spectral characteristics, speech rate, pausing, etc. While none of these aspects has been completely ignored, they are generally not investigated with the same degree of detail as F0 when it comes to intonational theory (the exceptions here are pausing and speech rate in terms of final lengthening, which have been investigated in great detail in relation to intonational phrasing). However, it is becoming increasingly clear that these aspects of the speech signal can influence the perception of pitch in partnership with the F0, and that by neglecting them to focus only on the F0 we have at best created only a limited characterization of how prosodic variations affect the meaning of speech.

Finally, previous literature has occasionally used listening studies to identify prosodic characteristics of the signal related to topic structure, but these studies have always been metalinguistic, specifically asking speakers where boundaries lie. This is a problem for the identification of topic structure as psychologically real, since although listeners perform consistently, these tasks do not necessarily entail normal language use. Without this evidence of the psychological reality of topic, there is little point in studying it; if it does not exist in natural language settings, then we must conclude that it does not influence the development of prosodic systems. Of course the systematic variations observable do provide good reason to believe in its psychological reality, but this has yet to be demonstrated in an experimental setting.

1.2 Goals of this thesis

This thesis addresses the problems discussed above by means of a production experiment (chapters 3-6) and a perception experiment (chapter 8) dealing with prosodic variations related to topic structure. The goals of this research are as follows:
CHAPTER 1. INTRODUCTION

To investigate the role of prosodic cueing of topic structure in a manner that is aware of both the discourse structure aspects and the phonetic and phonological aspects (although the primary point of view of this thesis is a phonetic/phonological one). The production study applies a detailed model of topic structure and investigates how prosodic variation relates to this model and whether the prosodic variation motivates the existence and form of the model.

To investigate the characterization of this prosodic variation as linguistic or non-linguistic, by comparing it to what is known about other linguistic means of signalling the meanings identified. In particular, to make a claim as to the linguistic nature of discourse structure signalling in longer units, and the need to include this kind of variation in linguistic theory.

To study prosodic variation rather than only intonational variation, by taking into account a variety of prosodic characteristics of the speech signal rather than only the F0, and especially to consider how the different prosodic characteristics may relate to one another, whether these relationships appear to be necessary or more coincidental.

To demonstrate that listeners are sensitive to the prosodic cues related to topic structure, and show that they use them to guide their processing of utterances as they listen, not solely when required to identify topic structure in a metalinguistic task, but in a more natural language-use setting.

1.3 Organization

Chapter 2 reviews the literature in a variety of fields relevant to these studies. Chapters 3-4 report the methodology of the production experiment, and findings related to variations in F0. Chapters 5-6 continue to discuss the production data, this time with regard to non-F0 prosodic variation. Chapter 7 reports a post-hoc investigation of the applicability of the topic structure categories used in the production study for the text in question. Chapter 8 presents the methodology and findings of the perception experiment. Finally, chapter 9 brings together these findings in a discussion of their implications for prosodic and discourse theory.
Chapter 2

Literature Review

2.1 Prosody and the Speech Signal

2.1.1 A complex relationship

One of the goals of intonational theory is to arrive at a satisfactory description of the relationship between the complex acoustic signal and the underlying (presumably abstract) units. Pierrehumbert (1980), for example, explicitly attempts to create a theory which is phonetically (acoustically and ultimately articulatorily) transparent; that is, observation of the speech signal should be sufficient to identify the underlying structure (cf. section 2.2.3.4). Similarly, the IPO researchers (’t Hart et al. 1990, section 2.2.3.2) base their theory on a series of perceptual tests, identifying phonological similarity on the basis of perceptual similarity. However, this is not at all a simple task. To begin with, the relationship between the measurable acoustic properties of the speech signal and what is perceived by listeners is not always easy to quantify. Furthermore, experimental settings mean that cues can be tested in isolation which may never occur in isolation in actual speech. The role of these cues in influencing perception may thus be over- or under-rated. Thirdly, although prosody is easily understood in conversational interaction, it is becoming increasingly apparent that it is subject to a great degree of variation across individuals, both in production and perception. The sections below discuss some of the main acoustic features which must be accounted for in descriptions of prosody, with a particular focus on the relationship between fundamental frequency (F0) and intonation.
CHAPTER 2. LITERATURE REVIEW

2.1.1.1 F0 and pitch

Fundamental frequency (F0) is commonly considered to be the main acoustic feature which leads to the perception of pitch. Pitch is the auditory impression that certain sounds are higher or lower than each other e.g. on a musical scale. It is not equivalent to F0, although in some cases these terms are used interchangeably. In fact, the relationship between F0 and pitch is not straightforward.

Voice F0 can be defined as the rate at which a speaker’s vocal folds open and close during the production of voiced sounds, measured in Hertz (Hz, cycles per second\(^1\)). To an extent this is controlled by an individual’s anatomy: adult male speakers have a larger vocal apparatus and therefore a lower average F0 than adult female speakers or children. However, the rate at which the vocal folds vibrate can also be affected by the tension of the muscles around the larynx and by the subglottal pressure, or the pressure at which air is being expelled from the lungs (cf. Titze 1989, Pickett 1999). Different combinations of laryngeal muscular tension and subglottal pressure can result in the same F0, although they may lead to different spectral characteristics for the resulting sound.

Constant variations in speech articulation mean that the measurement of F0 does not normally result in a smooth, continuous line. In speech, the vocal folds are not always vibrating; the presence of a phonetically voiceless segment leads to a gap in the F0 trace (or, depending on how the measurement is calculated, spurious and/or discontinuous values). Also, changes in the shape of the vocal tract (i.e. as a result of segmental articulation) can affect the movement of the laryngeal muscles or the air-pressure balance on either side of the vocal folds, which can affect the rate at which they vibrate. This can lead to what are called microprosodic effects or pitch perturbations: small variations in the F0 which do not contribute to intonational/tonal meaning, although they might be helpful in segmental perception (Liberman 1961; Silverman 1984, 1990).

Although F0 is the primary measurement used in studying pitch, F0 itself may not be as strong a cue to pitch perception in speech as other factors, particularly the harmonic structure of the signal. The concept of a “residue” pitch, a low pitch perceived on the basis of a series of high harmonics, was first introduced by Schouten (1970). Ritsma (1967) found

---

\(^{1}\)This measurement assumes that speech sounds are periodic; that is, that they repeat with complete regularity. This is not strictly true but is a useful approximation.
that the third, fourth and fifth harmonics in particular affected subjects’ judgment of pitch in the range of human voices (i.e. 100-400 Hz). Other research has since suggested that changes in pitch are best perceived when low harmonics are present in addition to F0 (Moore & Glasberg 1988; Carlyon & Shackleton 1994). However, F0 has remained in common use as the acoustic measurement for the percept of pitch, probably because it is simpler to measure F0 than to calculate the additional effects of other spectral characteristics; and in fact the results related to F0 have proven to be reasonably reliable.

Despite the aforementioned discontinuities in the signal, humans appear to perceive pitch as a more or less continuous characteristic, with relatively smooth transitions. That is, microprosodic variations do not influence the perception of the overall pitch contour, although large jumps are still perceptible and may be meaningful. The output of the auditory system is thought to pass through a “smoothing” device during perception. This has been modelled theoretically as a low-pass filter (Viemeister 1979) or as a temporal integration system which gives a weighted average for a sliding perceptual window (Moore et al. 1988; Plack & Moore 1990). This means that small variations in the signal can get “ironed out” without losing the larger changes that are necessary for perception. Furthermore, the system is likely to be flexible and therefore adaptable to different perceptual tasks, so small variations which might go unnoticed in a casual speech situation are still perceptible if the listening task requires that they be identified (Moore 2009).

In addition to the discontinuities in the F0 contour itself, another difficulty in the study of intonation from a perceptual point of view is that the pitch percept is not linearly related to the F0. That is, an increase of 10 Hz in the F0 does not have the same perceptual effect if it is a change from 100 to 110 Hz versus a change from 200 to 210 Hz. A variety of scales have been proposed to describe the relationship between changes in F0 and changes in pitch, all of which make use of logarithmic transformations of the F0 value, but the details of how they are calculated are not the same. The semitone (st) scale, commonly known from music, is simply the logarithm of the F0 (as measured in Hz). Another measure, mels (Stevens, Volkman & Newman 1937), is also essentially logarithmic, although it uses a nearly linear scale to transform pitch values below 500Hz (i.e. in the range of human voice F0). The Bark scale (Zwicker & Terhardt 1980) and the ERB scale (cf. Moore & Glasberg 1983) are based on perceptual experiments using masking noise; both identify the
threshold bandwidth of masking noise that affects perception of pitch at a given F0 value. The values given by these experiments also lead to more or less logarithmic values at higher frequencies, but like mels, the relationships tend to be closer to linear below 500 Hz.

Most of the experimental work on these scales has been done in the context of pitch perception independent of language. Some intonational researchers have tested the scales in the context of language but their results vary. Rietveld and Gussenhoven (1985) found that the Hz scale was acceptable in the context of comparing prominence, although Nolan (2003) suggests that this may be because F0 and amplitude can co-vary or be confounded in the case of assessing prominence (cf. Kochanski et al. 2005, who find amplitude to be more important than pitch in conveying prominence); and indeed, variations in amplitude can affect the perception of pitch even with the same F0 (Moore 2003). Patterson and Ladd (1999), on the other hand, found that both the ERB scale and the logarithmic (i.e. semitone) scale were more closely correlated with listeners’ judgments about the emotional content of utterances than was a Hz scale.

Neither Rietveld and Gussenhoven (1985) nor Patterson and Ladd (1999) compared intonation contours across categories. Nolan (2003) asked listeners to imitate intonation contours in their own voice range, and then compared the results of the different scales to see which yielded the least error between the normalized values of the stimulus and the speakers’ imitations. He found that the semitone or ERB scales were the best models of “intonational equivalence” (2003: 774), and suggests that this is because the task in the experiment (imitating intonation contours) was a more “musical” task than the identification of prominence. The results of these experiments therefore seem to converge on either semitones or ERB as appropriate ways of expressing changes in magnitude of the pitch percept in a linguistic context.

House (1990) also investigated how pitch is perceived in speech contexts as opposed to in general auditory perception. He started from the observations of a number of previous studies which had indicated that pitch perception in speech did not always follow the same rules as studies conducted with pure tones. In part this was because of the nature of speech sounds, which always consist of complex tones; but the difference was greater than could be accounted for by this. For example, the perception of F0 movements in vowels when tested
by means of a just-noticeable difference (JND) task\(^2\) appeared to be different from when perception of pitch differences was tested using single-vowel stimuli with dynamic F0s (Rossi 1971, 1978; Klatt 1973). Furthermore, Rossi (1971, 1978) had found that movements of 2-4 st might not be perceived as movement in some contexts, but rather as a level tone with a pitch value comparable to the value about two-thirds of the way through the vowel in question.

House’s (1990) research suggests that a critical question in the perception of pitch is the amount of spectral information or spectral change (i.e. from a segmental point of view). Stevens and Blumstein (1981) found that the point of maximum new spectral information coincides with the increase of signal intensity at the vowel onset. House therefore conducted a number of tests investigating how listeners perceive pitch when the pitch movements coincide with different portions of the vowel(s); in other words, under variation of the amount of new spectral information. He found that in steady-state vowels, listeners use pitch movements to categorize items into intonation groups. However, when consonantal information and therefore changes in the spectral information are included in the stimuli, listeners are far more likely to rely on the differences between the endpoints of the pitch movements. In other words, they appear to be recoding the pitch movement into a series of pitch levels.\(^3\) It seems, therefore, that listeners are able to use movement for classification, but that the presence of spectral change disrupts the use of movement as a cue.

House interprets this as possibly resulting from the limitations of short-term memory: since spectral changes must also be taken into account, the pitch movement is recoded into a (simpler) level, meaning that less information needs to be held in the memory. He also notes that different portions of the tonal contour might be optimized for perception of movement versus levels:

\[\text{\"Vowel onset could provide the perceptual mechanism with an important boundary immediately after which the perception of tonal levels would be favored while the perception of tonal movement would be favored during areas of spec-}\]

\(^2\)Investigations of just-noticeable differences (JNDs) aim to identify the smallest variation in pitch, loudness, etc. which can be perceived. Listeners are presented with pairs of stimuli and asked to say whether the two items in the pair are the same or different. The JND is the smallest interval between the two stimuli at which listeners reliably indicate a difference.

\(^3\)This is not the case for all listeners; while the majority of listeners shifted to an endpoint-based perception mode, some continued to use the pitch movements, and others ended up being unable to classify the different stimuli into groups at all.
House also extends his research on tonal perception from single syllables to the perception of prosodic boundaries, where he finds once again that pitch levels seem to be more important than pitch movement. The two strongest phrasing cues he identifies are a falling configuration at the end of a phrase, and the change of tonal level from low to high across a phrase boundary; but the change of tonal level is the stronger cue (1990: 98), although there is no indication of whether this might work differently in different sentence types. The phrasing cues are consistent with previous findings by Bruce and Granström (1989), who demonstrated that this was also the case for syntactic disambiguation in Swedish. House (1990) suggests that the best way to think about this combination of cues is in terms of dual perception:

“The perception of tonal movements should...probably not be seen as a binary choice between levels or movement configuration but rather in terms of perceptual constraints which entail a perceptual optimization and not necessarily a perceptual imperative” (1990: 100).

In particular, the type of encoding can depend on the kind of structure involved; House proposes that local information is cued by pitch movements or configurations, while global intonational information is conveyed by pitch levels, again due to the way this information is encoded in short-term memory. He sees this as an argument against strict linearity in perception.

The data collected by House lead him to the conclusion that pitch levels have primacy over pitch movements in the speech mode of auditory perception. On this basis, he argues that pitch movements should be encoded as levels unless they meet three conditions: pitch movement throughout the steady-state portion of the vowel; the start of the pitch movement occurring during the initial portion of the vowel (i.e. coinciding with the peak of new spectral information); and a vowel length of 100ms (or a relatively long vowel in context, as this can vary depending on speech rate) (1990: 134). He further notes that pitch movements which are too fast or too slow may also be encoded as level tones.

The implication that levels are relevant to perception of pitch in linguistic contexts lends important support to intonational theories based on levels, such as autosegmental theories.
(cf. section 2.2.3.4), although the evidence for this remains incomplete in two ways. First of all, in House’s initial experiments he found a great deal of individual variation between listeners, with some listeners continuing to use pitch movements as their primary categorization cue even in cases where greater degrees of spectral information occurred. It is of course possible that this was a task-based effect; since the listeners were not being asked to identify segmental information it would presumably be possible to ignore the spectral changes and continue to listen more closely to the pitch. On the other hand, this may be a case of different listener strategies in general; even if one strategy is preferred by a majority, the other is still possible. If this is the case then it is not valid to make theoretical claims about phonology on the basis of perceptual recoding into levels.

Related to this is the second way in which the study is incomplete: it was conducted only on native speakers of Swedish. While House notes that his intention is to use Swedish to make claims about language in general, it is possible and even likely that different languages would use different strategies, especially since different speakers of the same language are apparently using different strategies. It is not physiological or general perceptual limitations that require the re-encoding of pitch movements into levels; this may be a preferred method for Swedish speakers but not occur in other languages. The requirement for listeners to simplify intonational information may also be different in a pitch accent language, like Swedish, than in an intonation language, like English, where the information load carried by pitch may be less dense. This remains purely speculation for the moment; however, it is obvious that not all portions of the pitch contour are created equal in terms of their perceptual relevance and even their perceptual accessibility, and it is essential to take this into account when investigating the perception of intonation in any context.

2.1.1.2 Other prosodic characteristics

The relationship between F0 and pitch is well-studied, if complex, and it is relatively easy to make generalizations about the relationship between the two, both in isolation and to a certain extent within language. However, other acoustic characteristics of the signal can also affect the pitch that is perceived. There is a growing body of evidence that suggests that speech perception is likely to be a special mode of auditory perception (cf. Repp 1984), and that certain mechanisms that are clearly observable in the perception of pure tones are
not necessarily present or used in speech perception. Loudness/amplitude has already been mentioned briefly above; it is well-established in the literature that perception of pitch can vary with loudness, so that pitches below 2000 Hz seem lower when loudness is increased, while pitches above 4000 Hz seem higher under that condition (cf. Moore 2003). Increased amplitude is considered to be a correlate of prominence at least in English (Beckman & Pierrehumbert 1986; Kochanski et al. 2005) although not in all languages; this means that in considering what pitch listeners are perceiving it may also be important to take into account variations in amplitude or other phonetic characteristics across the signal. Changes in phonation quality such as creaky voice have been associated with low pitch in English (cf. Dilley et al. 1996) though occasionally with high pitch in other languages (Gordon & Ladefoged 2001), but given the often-aperiodic signal in this context, it is difficult to quantify what pitch is perceived.

Although pitch features seem to be the most important in terms of intonational perception in pitch accentuation, other prosodic characteristics have a great deal of relevance in terms of phrasing. Perhaps the most notable among these are pausing and variation in phonation quality at prosodic boundaries. Variation in pausing at phrase boundaries of different strengths has been reported by Lehiste (1975) among many others, and is fairly uncontroversial as a boundary cue; Swerts (1997) demonstrated, for example, that longer pauses correlate well with stronger discourse boundaries.

Variation in phonation quality, particularly the use of glottalization at phrase boundaries, is also well-recognized, although it has proved somewhat harder to quantify. In part this is because of difficulty with the term glottalization: although it implies an articulatory setting, it is also used in acoustic contexts. In this case it refers to a variety of characteristics that the signal can have which give it the auditory impression of glottalization, including aperiodicity, creak, and diplophonia (cf. Gerratt & Kreiman 2001, Redi & Shattuck-Hufnagel 2001). Despite this confusion, there is a general consensus about the presence of glottalization at phrase boundaries. Catford (1977) and Wells (1982) both report the presence of creak at the ends of declarative utterances in Received Pronunciation (RP). Expanding on these studies, Henton and Bladon (1988) investigated the presence of creak in utterance-final position in RP and Modified Northern (British) English (MN). They found that in both dialects, utterance-final syllables and penultimate syllables showed a high incidence of
creak, although there were gender and dialect differences, with men using creak more than women, and MN men using creak more than RP men.

A number of studies have investigated the production of glottalization at phrase boundaries in relation to segmental acoustics. Pierrehumbert and Talkin (1992) found that production of glottalization in /\l/ contexts in English was common in stressed syllables, but in unstressed syllables, the glottalization or /\l/ was not realized unless there was a phrase boundary. Similarly, Pierrehumbert (1994) found that the glottalization of /t/ and /p/ in final position, as well as glottalization of vowel-initial words, depended on the syllable structure as well as the prosody of the phrase in which they occurred. Dilley et al. (1996) found that the degree to which word-initial vowels showed glottalization increased in relation to the strength of the prosodic boundary they followed. Vowels at the beginnings of intonational phrases showed more glottalization than vowels beginning intermediate phrases that were IP-medial, which in turn showed more glottalization than word-initial vowels in other positions. They suggest that glottalization at phrase boundaries can be considered as a strengthening phenomenon, similar to the segmental hyperarticulation effects found by Fougeron and Keating (1995) at phrase boundaries. Redi and Shattuck-Hufnagel (2001) provide data further supporting this claim. Their speakers produced a greater degree of glottalization in utterance-final words than on words at the ends of intonational phrases that were utterance-medial, as well as a greater degree of glottalization at IP-final boundaries than at boundaries of smaller units like intermediate phrases (cf. also Huffman 2005). However, as Redi and Shattuck-Hufnagel (2001) note, speakers use a wide variety of strategies to produce the acoustic effect of glottalization. As other authors have pointed out (e.g. Pierrehumbert & Talkin 1992, Fougeron & Keating 1995), this variation between speakers could be problematic in terms of accepting glottalization as a consistent perceptual cue.

Another complicating factor in studying the incidence of glottalization at phrase boundaries is that glottalization or creak are associated with a number of other segmental or prosodic characteristics that occur independently of prosodic phrasing. For instance, syllable-final voiceless stops in English may lead to the production of creaky voice on preceding consonants if not to a complete allophonic substitution by a glottal stop (cf. Gordon & Ladefoged 2001).4 Creaky phonation by definition also has an F0 which is lower than that

---

4As a syllable-boundary phenomenon, this is still related to phonological grouping, and is therefore probably
of modal speech, and this may be used by listeners as a perceptual cue to identify glottal-
ization, as reported by Gerratt and Kreiman (2001); similarly, some speakers may use a
lowering of intensity instead of glottalization in comparable speech contexts (cf. Dilley &
Shattuck-Hufnagel 1995), which means that it is necessary to be cautious in identifying the
perceptual effects of these acoustic variations. There is also a confound with the presence of
creak; Dilley et al. (1996) report that glottalization was more common in contexts following
pauses, but since pause length can be a cue to boundary strength (cf. Swerts 1997), it may
be difficult to identify separate contributions of the two cues in these contexts.

2.2 Intonational Phonology

2.2.1 Motivation

The existence of categorical perception is a way of demonstrating the existence of linguis-
tic structure in sound perception. In particular, it is a way of showing that listeners group
sounds into meaningful categories, identifying, for example, all stops with a voice-onset
time (VOT) of greater than 30ms as voiceless (Abraham & Lisker 1967). These distinc-
tions, in turn, are used by listeners to identify the words produced by a speaker. The mean-
ingfulness of these distinctions is therefore in their link to an abstract category rather than
their specific phonetic characteristics (although of course the phonetic characteristics can
also be meaningful in and of themselves). This kind of meaningfulness as linked to an
underlying abstract category is what we can term a “phonological” distinction: the set of
sound contrasts that is meaningful in a given language. From a segmental point of view, we
are interested in whether listeners distinguish between /ð/ and /ð/ as categories (a contrast
which is meaningful in English but not in Spanish, since the latter does not have phonemic
voiced plosives), or between /n/ and /n/ (a contrast which is meaningful in Spanish but not
in English, which does not have a phonemic palatal nasal). Intonational phonology seeks to
describe the same kinds of contrasts in pitch patterns at a similar level of abstraction. For
purposes of this thesis I will focus primarily on studies about intonation languages; that is,
languages which do not use pitch for lexical contrasts. Tone languages have been shown
to have intonational characteristics as well (e.g. Xu 1999 for focus in Mandarin), but the

not completely independent of higher-level prosodic phrasing.
intonational systems for these languages may be different since they must also allow for variations based on tone. However, the review will in some cases cover pitch accent languages (e.g. Swedish), which have both intonation-like and tone-like pitch characteristics.

2.2.2 Accounting for the signal

Intonational theories must account for at least two types of phenomena in the speech signal: prominence phenomena and boundary phenomena. Most theories, when dealing with prominence, also have a rhythmic or metrical aspect, which can be considered as contributing compositionally with the first two. The metrical aspects are primarily relevant in terms of the placement and grouping of pitch accents.

Prominence phenomena in intonation languages comprise what are known as pitch accents: pitch movements associated with an accented syllable. The rules of accentuation are determined in part by metrical organization, and in part by the semantic structure of an utterance; the inventory of accents available varies by language and also by theoretical model. Ladd (1996) identifies these two elements pre-theoretically as “relative prominence” and “tune” (1996: 9), although “tune” of course refers to both pitch associated with accentuation and that associated with boundaries.

Boundary phenomena concern the organization of speech into groups of varying sizes, from syllables or feet up to utterances or longer units, although not all authors believe the larger groupings are possible (e.g. Vaissière 2005). In terms of intonation, the kinds of groups that are of interest are primarily those surrounding individual pitch accents (an accentual phrase, AP) and those which may bring together groups of accents (an intonational phrase, IP). Boundaries can constrain the scope of an accent’s meaning, since the semantic content affected by an intonational meaning is not necessarily equal to the segmental content modified by an intonational event. It is likely that boundaries also serve to divide language up into “digestible” units which can be processed as a whole (cf. Sturges & Martin 1974; Reeves et al. 2000).

---

5 I here refer to pitch accents in intonation languages in particular. Many languages use pitch accents to create lexical distinctions, e.g. Swedish, but while these accents may be similar to intonational accents, their function is different.
2.2.3 Modes of description

Early intonational theories tried to account for the meanings of whole-sentence pitch contours; Jones (1909) divided up contours into two basic tunes: a fall, assigned to statements, commands, and wh-questions, and a rise, assigned to yes/no questions and requests (cf. also Armstrong & Ward 1926). However, most theories currently in use acknowledge that it is possible and even desirable to account for sentence contours as a series of distinct parts. Palmer (1922) was an early advocate for the discreteness of intonational events, dividing intonational contours (i.e. what Jones 1909 had treated as a single unit) into the functional elements of head, nucleus, and tail. More recent analyses in the British tradition, such as those by Kingdon (1958), O’Connor and Arnold (1961), and Halliday (1970), have drawn on this insight.

Modern intonational theories, while agreeing that sentence contours are composed of a set of discrete parts, can be roughly divided into two groups; those which describe the units of intonation as contours or pitch movements, and those which describe the units of intonation as tones or targets, with pitch movements interpolating between them. The British school has preferred a contour-based system (e.g. O’Connor & Arnold 1961; Crystal 1969; Halliday 1970; cf. Cruttenden 1986) and pitch movements were also adopted by the IPO in Eindhoven for Dutch (’t Hart, Collier & Cohen 1990) and, to an extent, the Kiel Intonation Model (KIM) for German (Kohler 1991). On the other hand, descriptions based on a series of level tones became more popular in the United States (e.g. Pike 1945) and have led more recently to the relatively mainstream autosegmental-metrical (AM) analysis (Bruce 1977; Pierrehumbert 1980), which is now being extended to a number of other languages. Proponents of contour theories argue that contours are more perceptually viable from an impressionistic and an experimental point of view, while supporters of tone/level theories point out the simpler descriptive system that these make possible. To a certain extent, the different models of intonation have used similar terminology and are translatable one to another (since the phenomena to be described are basically the same). In some cases, this means that the same terminology is used with slightly different implications; in what

---

6 ToBI (Silverman et al. 1992) is both a modified version of Pierrehumbert’s (1980) analysis as well as providing a terminology for AM descriptions of intonation. It has been adapted as, for example, GToBI (German), Grice & Benzmüller 2005; ToDI (Dutch), Gussenhoven et al. 2003; Sp-ToBI (Spanish), Díaz-Campos et al. 2002; K-ToBI (Korean), Beckman & Jun 1996; and J-ToBI (Japanese), Venditti 1995.
follows I point out some of these different uses when relevant.

### 2.2.3.1 British school

The British school (O’Connor & Arnold 1961; Crystal 1969; Halliday 1970; cf. Cruttenden 1986) studied intonation primarily in an impressionistic way, with researchers making observations about similarities or differences between different intonational patterns. This tradition describes intonation as a series of accents realized by changes in pitch, lengthening, and intensity. Pitch accents are syllable locations where there is a large change in pitch from the context syllables, with a pitch movement either towards or away from the pitch of the accented syllable. The pitch movement can be either a step (i.e. relatively discrete) or a contour (i.e. a smoother movement).

Pitch accents are local events which contribute to the overall contour or “tune” of the entire intonation group. While in some cases the meaning of a whole tune can be identified, in most cases the different parts of the tune contribute to the meaning in identifiable ways, making it useful to separate out the different components.7 In particular, variations in the nucleus can change the meaning of tunes.

The nucleus, a notion which exists across intonational theories, has been defined as the final pitch accent in an intonation group (e.g. by Pierrehumbert 1980 and ToBI analyses developing her framework); in the British tradition, however, it is defined as the most prominent pitch accent, which is not obligatorily final (cf. Cruttenden 1986: 42).8 A final pitch accent may be realized with low prominence, making the penultimate accent the nucleus. The greater phonetic prominence of the nucleus is linked to greater prominence in terms of meaning. In this analysis, the nucleus and the part of the tune that follows it are the most important part of the meaning of the intonational contour (within the intonation-group); the structure of the nuclear accent can thus be said to be left-headed. Cruttenden (1986) describes the relationship between the portion preceding the nucleus and the nucleus itself as being similar to the relationship between a prefix and a root morpheme, where the most important part of the meaning is carried by the root. Although the prefix still has something

---

7Discreteness and compositionality are relevant in all of the theories that will be discussed here; for contrast, see Couper-Kuhlen & Selting (1996), who argue that compositionality of intonational meaning does not exist.

8This refers to phonological, not phonetic prominence; non-focus-bearing nuclear pitch accents do not tend to differ acoustically from non-nuclear pitch accents (cf. Ladd 1996).
to contribute, it modifies the root rather than adding some entirely independent meaning.

One corollary to this analysis of intonational tunes is that nuclear tones and prenuclear tones do not have the same status in the theory; they serve different functions, and the inventory available for each is not necessarily the same. The series of prenuclear tones preceding a nuclear pitch accent is sometimes called the “head”, and the choice of head makes only relatively small modifications of meaning. In fact, Cruttenden (1986) argues that the choice of head is based more on a specific language’s preference than on the need to make any large-scale meaning variations. Therefore the analysis of nuclei has received much more attention and is much more detailed in the British tradition than the analysis of prenuclear/head phenomena. The difference between prenuclear and nuclear accents will be relevant for the investigation of pitch accent distribution in chapter 4.

### 2.2.3.2 IPO

The IPO model of intonation, presented in great detail in ’t Hart, Collier & Cohen (1990), is another contour model, although it takes a different approach to the British tradition. The IPO approach is based on a large body of perceptual data, and aims to use perceptual limits to narrow down what kind of phonetic variation can be meaningful to listeners. Specifically, they claim that

> “...the F0 changes that do contribute essentially to the perception of the speech melody are just those changes that are programmed and voluntarily executed by the speaker” (1990: 40).

Discarding the impressionistic approach used in the British tradition, the IPO researchers used extensive experimentation to narrow down what constitutes meaningful versus non-meaningful pitch change for listeners. Their approach can be considered “bottom-up” in that they looked for phonetic patterns first and then tried to group and/or differentiate these patterns with perceptual testing; they specifically resist any “linguistic categorization” of the patterns they identify.

The IPO researchers chose to use pitch contours rather than level tones as a descriptive method because of the lack of “invariant melodic properties” of levels-based approaches: that is, they believe the characteristics of a level tone cannot be defined specifically enough.
The assumption of the IPO researchers is that the characteristics of pitch movements themselves are not accidental and cannot be determined on the basis of the characteristics (e.g. slope, shape) of the end points of the movements, as in an AM approach (section 2.2.3.4). A contour-based approach, they argue, can take into account characteristics of contours that are invariant even across widely differing segmental material, such as their direction, timing in relation to syllable boundaries, slope, and size. These last two are defined in generic and binarily-distinguishable ways rather than gradiently, as fast/slow and full/half. These features are seen as being inherent to contour categories.

There is a certain amount of distinction between “Prefix” (prenuclear) and “Root” (nuclear) pitch movements, although not as much as in the British tradition. In the IPO grammar of Dutch intonation, different pitch movements are considered to be available in Prefix, Root, or Suffix (postnuclear) position, without being essentially different from one another. However, there are constraints as to the order in which some pitch movements may occur; a given movement may only be acceptable as a Root following a particular Prefix or combination of Prefixes. The set of Roots as a subset of all possible pitch movements is this system’s method of accounting for the similarity of nuclear accents observed by researchers in the British tradition. As in that body of research, the Root (nuclear accent) is seen as being the primary feature on the basis of which intonational contours should be grouped. However, an important distinction between the British tradition and the IPO analysis is that while the British tradition groups contours by meaning, the IPO does not assume that contours have an intrinsic meaning. The similarity they refer to is a similarity of form, by which they mean the perceptual equivalence of F0 contours. They believe that additional pragmatic and situational concerns contribute so much to the choice of an intonation contour in context that it is impossible to narrow down a meaning for any given contour.

2.2.3.3 KIM

The Kiel Intonation Model (KIM; Kohler 1991, 1995) is another experimentally-based intonational phonology with contours as the atomic unit. The model seeks to identify “independently-motivated” (Kohler 1995) intonational categories that may apply regardless of the language studied.

The KIM falls somewhere in between a contour-based model and a tone-based model,
in that it identifies intonational features as peaks and valleys, but defines these terms on the basis of pitch movements. It is similar to the IPO approach as well as autosegmental theories in that the timing of pitch movements is highly relevant to the identity of intonational categories. In the KIM, timing is defined in terms of the F0 maximum or minimum rather than the contour as a whole, even though the contour is considered to be the relevant intonational feature. The KIM therefore combines features of both contour-based and levels-based approaches, and can be thought of as a “configurations” or pattern-based approach (Oliver Niebuhr, personal communication). Since the underlying phonological unit is a configuration, a speaker’s goal is to produce the perceptual effect of a given configuration. This may be done either in a canonical way (i.e. by forming the appropriate F0 contour) or by making use of some other cue that leaves the listener with the perceptual effect of the pitch configuration despite it not actually being present in the signal. Kohler (1995) gives the example of F0 fall truncation in short syllables: since the syllable is short, the F0 contour is cut off before the pitch reaches the speaker’s baseline, so the contour is in a sense incomplete (cf. Grabe et al. 2001 for English). However, the listener is able to make use of regularities in timing relative to the short vowel in question and thus can “recover” the intended contour despite the truncation. This reference to perceptual effects allows the KIM to account for variation in the mapping of F0 contours to the phonological categories. In its methodology, the KIM is therefore like the IPO in terms of having an experimental-perceptual basis.

2.2.3.4 Autosegmental approaches and ToBI

Autosegmental-metrical (AM) approaches to intonation have their roots in research done in the 1960s and 1970s on African tone languages (cf. Goldsmith 1976, 1990). The term “autosegmental” refers to the way the phonological representation is structured; instead of phonology consisting of a linear string of elements each with a different set of features, as in traditional generative phonology, it is made up of a series of levels or “tiers” of representation, and each tier carries the specification for certain features. The insight of this approach was that by locating tones on a separate tier and coming up with a set of rules for relating them to the segmental content, the description of the tonal structure of a language could be greatly simplified. A string of tones could be associated to a string of tone-bearing units (TBUs; these could consist of syllables, morae, or some other phonological unit depending
on the language), with the aid of a series of conventions describing how the tones were to be assigned to the TBUs, and what strategy to adopt in the case of there being a mismatch between the number of TBUs and the number of tones. This process of linking tones to TBUs is known as association; association and its counterpart alignment are discussed further in section 2.2.4.

Autosegmental-metrical theory also allowed for the insight that words with different numbers of syllables (or other TBUs) had the same tonal representation. As in example 2.1, the underlying tonal representation $H \downarrow L$ could be applied to two different words, one with two syllables and one with three, to come up with the surface representations of $H \downarrow L$ and $H \downarrow L \downarrow L$, by first associating the tones to the syllables left-to-right, and then spreading the last tone (L) to the third syllable to meet the requirement that all tone-bearing units have a tone associated to them.

(2.1) \[
\begin{array}{ll}
CV & \quad CV \\
H & \quad L \\
\end{array}
\]

\[
\begin{array}{lll}
CV & \quad CV & \quad CV \\
H & \quad L & \quad \nearrow
\end{array}
\]

Bruce’s (1977) dissertation applied some of these same rules to the description of the lexical pitch accent distinction in Swedish. The particular insight offered by his work was to identify word accent not with two separate pitch movements, but to simplify this definition by pointing out that “a pitch obtrusion in connection with the stressed syllable” (1977: 11) is all that is necessary for word accent to be present. The difference between the two word accents available in Swedish is that Accent I has an F0 peak before the stressed syllable and tends to have a steeper fall, while Accent II has an F0 peak within the stressed syllable, and a more variable slope. These conditions hold even when the effects of sentence accent (i.e. the nuclear accent) and terminal juncture (i.e. boundary phenomena) are taken into account. Bruce describes the accent patterns as a series of High and Low values which are context dependent in terms of their actual realization; both word accents are a sequence of High and Low, but in Accent I, the stressed syllable carries Low pitch, while in Accent II the stressed syllable has High pitch (see 1977: 134 for a detailed diagram). In other words, the two
accents are distinguished phonologically by the timing of the pitch movement in relation to
the stressed syllable; the underlying description is the same.

For Bruce, the use of a tone-based analysis rather than a contour-based analysis better
captures the regularities across different utterances, and the F0 contour is the result of a
set of production rules applied to a tonal sequence. However, the nature of these tones is
less specified than in later work, most notably that of Pierrehumbert (1980; developed with
followed Bruce’s analysis and used it to describe the intonational system of (American)
English. She, like the IPO researchers as well as Bruce, particularly wanted to create a
descriptive system which was phonetically transparent, although she states that she believes
that studying F0 is simply a shorthand for the physiological manipulations that underlie it,
and which are actually foundational to intonational variation.9 In Pierrehumbert’s approach,
tonation is comprised of a metrical representation of a text to which tones (either high (H)
or low (L)) are linked. As in Bruce’s (1977) analysis, H and L do not have specific pitch
values, even for an individual speaker, but instead reflect their context and to an extent the
degree of emphasis at that point in the utterance. This means, among other things, that the
same F0 value could represent a H tone at one point in an utterance and a L tone at another.
Tones can appear alone or in combination with another tone to form pitch accents. They
are annotated in the form H* (a monotonic high pitch accent) or L*H (a bitonal rising pitch
accent in which the low tone is most closely connected to or associated with the stressed
syllable). Also part of the intonational grammar are boundary tones, of which there are two
varieties: an intermediate boundary or phrase tone, which can fall between pitch accents
within an intonational phrase, annotated H-; and a phrase boundary tone, which occurs at
the end (and possibly the beginning) of intonational phrases, annotated L%.

Pierrehumbert (1980) presents her intonational system in the form of a finite-state gram-
mar, like the IPO researchers, but aside from choosing tones rather than contours for the
representation, there is another crucial difference: in the Pierrehumbert system, all possible
combinations of pitch accents are theoretically permissible within an IP. This differs from
other systems in that prenuclear accents may have the same theoretical status as nuclear

9The IPO researchers also take an interest in articulation, specifically claiming that the meaningful variations
in pitch are just those produced by movements of the laryngeal muscles (as opposed to e.g. pulmonic control);
however, they believe that perception, not articulation, is primary in intonational organization.
accents. The nuclear accent in the Pierrehumbert system is simply the last pitch accent in an intonational phrase. Therefore, a given accent, say H*, is “the same thing” whether it occurs as the first pitch accent in an intonational phrase or the last. This does not mean that the nuclear position is completely the same as prenuclear ones; the nuclear accent is still considered to be at least potentially the most prominent, and it may escape rules that apply to the rest of the phrase, such as downtrends (cf. Ladd 1996). However, the idea that the set of nuclear accents and the set of prenuclear accents is the same does have important implications for theories of intonational meaning: any “meaning” discovered to attach to a certain pitch accent would have to apply both in nuclear and in prenuclear contexts (cf. Niebuhr & Zellers, to appear).

The original intonational grammar proposed by Pierrehumbert (1980) has continued to undergo development since that time. Beckman and Pierrehumbert (1986) give a detailed exposition of a slightly modified version of Pierrehumbert (1980), addressing some criticisms of the model as well as applying it to lexical accent in Japanese and contrasting the prosodic structures of the two languages. Work by Pierrehumbert and Beckman and colleagues, as well as by Ladd and colleagues, has served to refine somewhat the list of possible pitch accents (cf. Pierrehumbert & Beckman 1988; Ladd 1989; Silverman et al. 1992; Ladd & Schepman 2003; Dilley et al. 2005).\footnote{The most notable debates here have been (1) the status of H*+L, and (2) the distinction between L+H* and H*. In Pierrehumbert’s (1980) original description, H*+L did not represent a falling pitch movement, but rather a high peak which licensed downstep on a following high peak. However, the inconsistency of this notation with the relatively straightforward interpretation of the other pitch accents has led to a great deal of criticism, notably from Ladd (1983). In ToBI, the H*+L pitch accent was dropped altogether from the inventory of pitch accents, and the symbol ! preceding a High tone was adopted to show that the accent had undergone downstep (Silverman et al. 1992). The distinction between L+H* and H* has been under debate for some time, particularly given anecdotal evidence that ToBI labellers are least consistent on this distinction (cf. Syrdal & McGory 2000). Ladd & Schepman (2003) suggest that the purported difference in this case is only one of scaling, not of meaning, and that therefore the two pitch accents should be collapsed into one. They propose the label (L+H)* in this case, but this convention has thus far not been adopted into the standard ToBI annotation scheme.} Silverman et al. (1992) combine a number of these theoretical refinements into a transcription system known as ToBI (Tone and Break Indices). As its name suggests, the ToBI labelling conventions take into account not only the combination of High or Low tones in the intonational grammar, but also the strength of boundaries between different lexical items, which gives an indication of how the speech is phrased (cf. Jun 2005).

Although the ToBI model is generally described as a theory of English intonation, the
majority of the research furthering it has been conducted on American English, and in some cases researchers are unspecific about the differences in intonation between different varieties of English. However, some work has been done specifically on other varieties of English, most notably Gussenhoven’s (1983, 2004) work on Standard Southern British English (SSBE). Gussenhoven’s theory is consistent in organization with the British school’s model while following AM descriptive principles. Intonation contours are modeled as a series of High and Low tones which can occur in various combinations to form the resulting pitch contours. Similarly to at least the original version of the Pierrehumbert theory, all pitch accents which are available in nuclear position are also available in prenuclear position, although Gussenhoven’s model additionally posits a tritonal accent, H*LH, which is available in prenuclear position only. A major difference between the two models is that Gussenhoven’s system, consistent with the structure posited by the British tradition, does away with the intermediate phrase tone H-/L-, which he sees as an unnecessary complication, particularly given the dubious status of the intermediate phrase in experimental settings (cf. White 2002). Gussenhoven (2004) argues that there is no proposed purpose for the intermediate phrase that cannot be accounted for with the combination of the prosodic word and the intonational phrase.

Another, perhaps more subtle, difference is in the relationship between the underlying tones and the phonetic realization of those tones. While Pierrehumbert (1980) explicitly states that the English tonal grammar she posits does not require tonal deletion rules and that therefore the underlying representation is completely transparent in the surface form, Gussenhoven’s (2004) model is slightly less transparent in that tones that belong to one unit phonologically may group differently with other tones in the surface representation; for example, the H*LH prenuclear accent tends to be realized as a local H*L, with the following H drifting right to be realized immediately preceding the start of the next tonal element.

Another difference is that all of the pitch accents in Gussenhoven’s model are composed of a starred tone with an optional trailing tone; none have leading tones, unlike in the ToBI analysis where both leading and trailing tones are possible in bitonal accents. This left-headedness is again consistent with the analysis of the British school. While an “off-ramp” analysis (Gussenhoven 2004: 319) may be theoretically more elegant in that it mirrors the
metrical foot structure in the tonal structure, it is unclear whether this kind of description is preferable in terms of transparency of the analysis.

Gussenhoven and Rietveld (1991) conducted an experiment comparing the Pierrehumbert and Gussenhoven models. Their results suggest that Gussenhoven’s model is a better fit for the way listeners perceive intonation. However, this result was not so strongly tied to the main tonal elements as it was to considerations of pitch range and permissible modifications (e.g. stylizations) of contours. This suggests that the exact details of the models of pitch accent structure are less important than the ways in which they can meaningfully vary their form, since the Gussenhoven model is richer in additional tonal morphemes which may contribute to the intonational structure. Perhaps this means that the intonational system does not need to be as strictly syntactic as might be suggested by the finite-state model proposed by Pierrehumbert (1980). However, further testing of this hypothesis has yet to take place.

2.2.4 Linking phonetics and phonology: alignment

One concept that has been extremely influential in autosegmental models of intonation in particular is the notion of alignment, or the temporal relationship between an F0 contour (or other prosodic information) and the segmental stream. In particular, F0 peak alignment has been proposed as being relevant to the signalling of new topics in discourse, as will be addressed in more detail in chapters 3-4.

Alignment goes beyond the concept of association borrowed from tone languages (cf. Goldsmith 1976, section 2.2.3.4). Specifically, alignment was Bruce’s (1977) answer to the two forms of lexical pitch accent available in Swedish: the unifying characteristic of word accents was the pitch peak, but the difference between Accent I and Accent II could be defined in terms of where the F0 peak fell in relation to the word, and in particular, in relation to the stressed syllable.

The notion of alignment has been the source of significant debate in the intonational literature in the autosegmental framework. In particular, there has been a great deal of experimental work done attempting to determine whether alignment is part of the phonology of a language or part of the phonetic system. Pierrehumbert’s (1980) analysis appears to conflate the terms association and alignment as both describing the process by which pitch
accents are linked to the appropriate (metrically strong) syllables in an utterance. The two terms have long had different meanings, however, and it is important to keep them separate.

“Association” refers to the system by which tonal elements are attached to tone-bearing units (cf. section 2.2.3.4); whether this means assigning a High or Low tone to each syllable in a word in a tone language, or whether it means assigning a pitch accent to a prominent or metrically strong syllable (or mora) in an intonation language. Tone association, therefore, falls within the realm of a language’s phonology. It is concerned with linking two tiers of underlying representation, but its output may also be an underlying representation in the sense that it does not define a specific F0 point; it is just a step closer to the surface production.

“Alignment”, on the other hand, is much more closely linked to the surface representation, regardless of whether it is ultimately considered to be phonological or a phonetic effect. A strong theory of alignment refers to a specific point within the tone-bearing unit (TBU) where a tone is affixed; the essential assumption when discussing alignment is that there is a great degree of regularity in how this affixation is accomplished. Whereas association assigns a particular pitch accent (say H*, represented by a peak in F0) to a particular (prominent) syllable, alignment determines where in the syllable the F0 peak falls; for example, the F0 peak may be aligned just after the onset of the syllable nucleus.

Bruce’s (1977) account of Swedish word accents may thus be considered as treating “association” rather than “alignment.” His model includes a temporal reference for the linking of the tone to any syllable which is classifiable for tone; this is the specific alignment (1977: 132). Therefore, the distinction between the two accents is one of association. The difference in alignment may of course be the source of the perceptual difference, regardless of the underlying analysis. Therefore, making the distinction between alignment and association has led to a great deal of debate in the autosegmental literature on intonation.

The question of whether alignment is a separate phenomenon from association has two sub-questions. One is whether alignment and association are actually separable (i.e. whether phonetic alignment is ultimately the main perceptual cue for pitch accent identity). The other is the degree to which these rules are language-specific. The Segmental Anchoring Hypothesis (SAH, Arvaniti et al. 1998) addresses these issues. It is based on experimental evidence that tones are consistently aligned to a specific point in the segmental
structure; the strongest version of the SAH claims that every tone in a given intonational system should have a consistent alignment.

Silverman and Pierrehumbert (1990) investigated the alignment of prenuclear H* in English, finding that the F0 peak in this tone was normally aligned after the end of the stressed syllable but that prosodic context could influence the alignment in systematic ways. Specifically, the upcoming (i.e. right-side) prosodic context could cause the F0 peak to occur earlier if the context included a boundary or stress clash; their study shows that it was not simply the additional segmental material which caused the shift in position, but the prosodic structure’s impact on the segmental material. Silverman and Pierrehumbert introduced the use of “peak proportion” for measuring tonal alignment; the peak proportion is defined as the timing of the peak from the onset of the syllable rhyme divided by the length of the rhyme (1990: 87). This measure has since been adopted in many other studies on F0 peak alignment.

Prieto et al. (1995) also studied the alignment of H* accents, but in Mexican Spanish. In their case, they found extremely consistent alignment of F0 valleys, whereas peaks were strongly affected by contextual factors including speech rate. They did not find peak proportion to be a useful measure for F0 peak alignment. In both Silverman and Pierrehumbert’s (1990) and Prieto et al.’s (1995) studies, the essential finding contributing to the SAH is that there is not evidence for contour invariance; the slope and length of the F0 rise are not constant across different segmental material. This finding also lends support to tonal theories in general, since in a contour-based theory different contours would be expected to show these invariant characteristics.

Following these two studies, Arvaniti et al. (1998) investigated the alignment of F0 peaks in Greek. They found evidence for a stably aligned H tone, which could be affected by proximity of the tone to the right-side word boundary and the number of unaccented syllables following the accented syllable (i.e. tonal crowding). Their main argument is that the factors influencing tonal alignment are to do with “phonological distance” rather than “phonetic distance” (1998: 22); that is, the factors which influence whether the alignment of the F0 peak is shifted are to do with the phonological structure. The implication is that the alignment must be a part of the phonological description, since otherwise phonological factors would not be able to have an impact on the phonetic representation. This conclusion
CHAPTER 2. LITERATURE REVIEW

43
gained further support from Ladd et al.’s (1999) study showing that in English, changes in speech rate did not affect the alignment of prenuclear accents (in contrast with Prieto et al.’s 1995 finding for Mexican Spanish). Research on Dutch by Ladd et al. (2000) also indicated that phonological structure, specifically phonological vowel length, was the primary source of variations in alignment in their data, although an effect of time pressure could be observed in phonologically long but phonetically short vowels. Similar effects have also been noted in less-related languages (and with different experimental paradigms): Xu (1998) reported that the ends of rises in the second tone in Mandarin are consistently aligned with the end of the syllable, and Ishihara (2003) found that F0 peaks in Japanese were aligned with a second mora, independent of whether this mora was part of the stressed syllable itself or part of the following syllable.

This body of experimental evidence, particularly in terms of considering alignment as part of the phonological specification of a language, led to the adoption of the notion of “secondary association” (cf. Grice 1995; Pierrehumbert & Beckman 1998; Gussenhoven 2000; Grice et al. 2000), which might also be called phonological alignment. Under this proposal, not only is a tone associated with a particular tone-bearing unit, it is also secondarily associated (i.e. aligned) to a particular portion of that unit, e.g. the alignment of an H* pitch accent to the onset of the stressed vowel.

Prieto et al. (2005) adopt this notion to describe alignment differences between Central Catalan, Neapolitan Italian, and Pisa Italian. Their analysis makes pitch accents more similar to boundary tones in that they claim that pitch accents can also align (secondarily) to edges of phonological domains; the difference is in the domains in question, since the domain of the accent is the whole accentual phrase, while the domain of the boundary tone is only the boundary. Prieto et al. also propose optionality of secondary association, both in the sense that it may not occur in all languages, and that even in languages where it occurs, it may not occur in all pitch accents. This would seem to be an unnecessarily complex analysis, however, given that the language might already have general alignment conventions (cf. Atterer & Ladd 2004).

Despite the appeal of the SAH, more recent evidence indicates that it is probably not valid, primarily because its claims are too strong. A growing body of data suggests that although F0 peaks do show consistent timing patterns, they can be greatly affected by a va-
riety of other factors, including even the conventions of a language as a whole. Atterer and Ladd (2004) showed that different patterns of alignment appear across languages for prenuclear rising accents; German speakers align F0 rises later than speakers of English, Greek and Dutch, and Southern German speakers align F0 rises later than Northern German speakers. This would seem to suggest that the patterns of alignment are in fact language-specific phonetic conventions, and that the differences reflect different realization rules across languages rather than phonological structure. Furthermore, Atterer and Ladd (2004) note two details which make it difficult to analyze the differences in terms of phonology. First, the differences in alignment between English, Northern German, and Southern German are very small, and would be best described in terms of anchoring to different points of the same domain (accented syllable or vowel). This is problematic in that it implies dividing up the domain, which adds a great deal of complexity to the description. Second, Atterer and Ladd note that there is apparently a continuum of alignment, with English aligned the furthest left, then Northern German, then Southern German. They draw a parallel with segmental-level acoustics, pointing out that a continuum with a variety of language-specific values does not necessarily indicate different categories, but simply different values for the boundaries of a single category. For example, different languages have different values for the voice-onset time (VOT) boundary between voiced and voiceless segments, even though the abstract contrast can be considered essentially the same (cf. Strange 1999).

The conclusion that there is a continuum of alignment across languages may be contradicted by research on Irish dialects by Dalton and Ní Chasaide (2005); they found no evidence that the dominant rising accent of Ulster Irish was simply a realignment of the dominant high or falling accent of Connaught Irish. Their synchronic analysis, determining that the two accents have different underlying categories, does not block the possibility of a diachronic derivation from the same category. Given the alignment characteristics of the two accents, they see this as improbable, or at best requiring a complex account. However, this result could be brought into line with Atterer and Ladd’s findings by arguing that diachronically there was a slippage in the alignment of the accent, followed by a shift in the categories (i.e. the replacement of one phonetic form by another in the new category).

A large body of intonational research has focused on Germanic languages, primarily English, German, and Dutch; and much, though not all, of the data supporting the SAH
CHAPTER 2. LITERATURE REVIEW

has come from research done on these languages; a notable exception is Spanish (Prieto et al. 1995, 2005), which, however, does not appear to fit with the data for the Germanic languages. D’Imperio (2000) found that H peak alignment in Neapolitan Italian L+H* and L*+H was affected by syllable structure, with peaks being aligned differently in open and closed syllables. Welby (2006) investigated tonal alignment in French two-rise accentual phrases (i.e. LHLH). She found that only the first L and the last H in these phrases were consistently aligned with the segmental structure, contradicting a strong version of the SAH, in which all tones would have to have a specific alignment. Post (2011) also found a lack of consistent tonal alignment, and indeed of phrasing, under variations in speech rate in French; she proposes that these variations may be better accounted for by phonetic implementation mechanisms rather than by a phonological description.

Some additional evidence against the SAH comes from Prieto and Torreira (2007), who found that syllable structure and speech rate affected F0 peak alignment consistently in Peninsular Spanish (as Prieto et al. 1995 had found for Mexican Spanish). Prieto and Torreira argue that the SAH in its full version is too strong, although the idea of anchoring could still be considered valid in that there is still a synchronization with the stressed syllable. However, they believe that it is more valid to consider alignment in terms of articulation; that is, “gestural coordination of F0 gestures with the supraglottal gestures” (2007: 23). D’Imperio et al. (2007) also make this argument on the basis of data from French and Neapolitan Italian; they find a better correlation of H target (i.e. F0 peak) timing with the peak velocity of the articulatory gesture in Neapolitan Italian than with any segmental measure, whereas in French, L targets appear to be aligned with the peak velocity. However, as D’Imperio et al. point out, one difficulty with measuring F0 peak alignment in terms of articulatory events is that this entails comparing an acoustic event with an articulatory event; that is, it is a comparison between events in two different systems. Still, they believe that articulation may be a more promising source of actual alignment, since there is “stability” in terms of acoustic parameters but no exact alignment.

Welby and Lœvenbruck (2006) describe this stability in terms of an “anchorage”; instead of a specific point, they claim that tones are anchored to a region within the segmental stream. This accounts for variability in alignment; however, Welby and Lœvenbruck see anchorage as a supplement to rather than a replacement for segmental alignment, since some
tones align more stably than might be expected from their anchorage account. One possible explanation for this, put forth by de Jong (2007), is that less variability might be possible in tones bearing a higher functional load. This would appear to be consistent with the conclusion of Schepman et al. (2006) that Dutch nuclear and prenuclear accents show somewhat different variability of alignment, with nuclear accents apparently less affected by phonological considerations (i.e. stress clash and syllable membership) than prenuclear accents. D’Imperio (2000) provides further results suggesting that starred tones are less variable than others.

Although the SAH in its strong form is no longer considered to be valid, stability of F0 alignment is still in use as a criterion in identifying phonological tones. Ladd and Schepman (2003) used this as one test in a study which concluded that L+H* and H* in the ToBI system should not be differentiated. They found that the L dip between two H* accents (where the second could be either L+H* or H* in the ToBI system) was consistently aligned with the word boundary, and they argue that this consistent alignment means that the F0 movement reflected a phonological tonal activity, rather than simply a phonetic effect. Dilley et al. (2005) took this study further by demonstrating that the L and the following H* were aligned independently of one another relative to the segmental string. Unlike L, H* was not affected by the syllable structure, as might be expected given the greater stability of starred tones found in other contexts. This can be considered evidence for the existence of two separate tones with different phonological status (i.e. starred versus not starred).

The SAH is of course not the only hypothesis which has been put forward to account for alignment data. One major competitor is the target approximation theory put forward by Xu and colleagues (Xu 1998, 1999; Xu & Wang 2001; Xu & Xu 2005), based primarily on work on tones in Mandarin Chinese.\footnote{Strictly speaking, the target approximation theory assumes a completely different mechanism of linking tonal specification to phonetic production, claiming that speakers aim at a static or dynamic tonal target within a given domain (in this case, the syllable), rather than interpolating between tones as predicted by autosegmental theories (section 2.2.3.4). Xu (1998) found that the contours of all lexical tones in Mandarin have a consistent alignment with the syllable that bears them, regardless of syllable structure. One advantage of the autosegmental approach to intonation is that it is possible to draw direct parallels between the tones in tone languages and those in intonation languages, since the difference between these two kinds of languages lies in the domain of pitch use, rather than in their having completely different systems.}
and speech rate. He examined the rising tone in more detail, and found that both the onset of the rise and its peak occurred in consistent locations, again regardless of syllable structure or speech rate. Xu’s framework for modelling lexical tone assumes that the rising contour is a single dynamic target, aligned with the syllable that bears it. Xu (1999) expands these results to all the lexical tones, and demonstrates that variability between syllables with the same tone is lowest at the ends of syllables. He considers this to be evidence for his hypothesis of target approximation, since the goal is to reach the target, and the time scope is the syllable. This idea is developed further by Xu and Wang (2001), who argue that this consistency of alignment at the ends of syllables is evidence for the syllable boundary being the place where the changeover from one tone to another occurs.

While this hypothesis does describe the Mandarin data that Xu and his collaborators present, it is more difficult to expand it to an intonation language such as English, where the pitch movements extend over larger domains. Xu and Xu (2005) propose that the same syllable-based specification can be adopted into English pitch accents without much modification, with variation such as peak delay caused by articulatory limits (Xu and Sun 2002) rather than by a phonological specification. However, there are several difficulties with this interpretation. First, Xu’s target approximation theory seems to depend on a certain amount of articulatory pressure. In Mandarin (among other tone languages), each syllable has a lexical tone specified, and therefore any attempt to reach a given target can only last for as long as the syllable does. In English and other intonation languages, however, the sparser tonal specification means that this is not necessarily the case, although there are certainly times when tones come close together both phonologically and phonetically. While a pitch accent is associated with a stressed syllable, it is more difficult to assign boundary tones a specification in this way (other than by saying that boundary tones have for their domain the last syllable of a phonological unit). A second difficulty with applying Xu’s theory to English is that the attempt to reach a tonal target is said to occur within the scope of the syllable only. This means that all syllables are specified for targets, even those which in an autosegmental theory would have their pitch determined by interpolation between tones. This accounts for differences in contour shape by the specification of each individual syllable, but seems unnecessarily complex in the case of English.  

12Autosegmental theories may also struggle to account for variation in contour shape. However, they have an
While the theoretical status of alignment is still debatable in terms of whether it belongs properly to the phonetic or the phonological system, there is no doubt that subtle differences in phonetic alignment can have large perceptual effects. D’Imperio and House (1997) found that they could shift listeners’ perception of questions and statements in Neapolitan Italian by shifting the location of the F0 peak relative to the accented vowel. Building on this, D’Imperio (2000) argues that the interpretation of sentence type in Neapolitan Italian is not based on a difference in the tonal sequence, but rather in the alignment of the tones to the text. Specifically, she demonstrates that a similar tonal sequence is used in both questions and statements, but that the tones are aligned earlier in statements. She notes that starred tones are more resistant to movement or variability, with the implication that the reason for this is the meaningfulness of such variability. She argues that the interaction between alignment and association is that the latter is “...the effect of contrastive alignment as a polarizing force acting on the observable phonetic variability” (2000: 28). The location of the nuclear F0 peak is therefore crucial to the distinction between questions and statements, though she notes that the shape of the contour can have some influence on perception of where the peak falls. As suggested for German in the KIM (Kohler 1991), D’Imperio argues that Neapolitan Italian listeners recover an “invariant” location for the pitch contour, which is then mapped to a meaningful category (2000: 188). She notes that the perception of tonal targets is not limited to the individual target itself, but that the alignment of many or all of the tones in an utterance or IP is taken into account by listeners when identifying an accent category. This conclusion is also borne out by work by Niebuhr (2009) on German, demonstrating that rhythmic context effects can change the categorical interpretation of a pitch accent; in other words, it appears that listeners adjust their expectations on the basis of prior context.

Advantage over Xu’s target-approximation theory in two ways. First, different contour shapes can be canonically assigned to different tonal representations without the requirement that targets be approximated according to a consistent mechanism. Second, even if the tonal representation appears to be the same, it is possible to imagine an optional feature specification for contour shape which could be included in the phonology. Xu’s target-approximation would have to assume that the two dynamic targets were completely different from one another. This might lose a generalization made possible by the similar tonal construction of two pitch accents in an autosegmental model, although Xu’s model is intended to be phonetic rather than phonological in nature.
CHAPTER 2. LITERATURE REVIEW

2.2.5 Intonational meaning

2.2.5.1 Form and function

Intonation bears a great deal of information that is not directly relevant to the linguistic message, although it is still communicative. For example, intonation carries information about the speaker such as gender or age. Although some of this information may be relevant to interpretation of the message at a given point in time, it is not generally necessary for comprehension. Similarly, “affective” information reflecting the speaker’s emotional state and/or reaction to the message can be carried in intonational variation in a systematic way, despite not being a part of the linguistic message. Affective information has been claimed to be borne by variation related to the Biological Codes (Ohala 1983; Gussenhoven 2004) in a reasonably transparent way. The Biological Codes were proposed as a source from which linguistic intonation has developed or been grammaticalized, but the underlying meanings are still present in language use. For example, the Frequency Code relates high pitch to small size and low pitch to large size; this can be visualized in that a small animal has a smaller vocal tract and therefore produces higher-pitched sounds, while in a large animal the reverse is true. In this way high pitch can represent submission, vulnerability, or friendliness, while low pitch can suggest authoritativeness, protectiveness, or unfriendliness. Similarly, the Effort Code relates large pitch excursions with large effort, which is in theory only expended for a specific purpose, so large pitch movements can have the affective meaning of expressing surprise or helpfulness, or in general greater arousal (Gussenhoven 2004: 95).

Although these meanings are important communicatively, they are not considered to be part of the phonological system in the sense of being encoded in a grammar. If we model language as a series of autonomous layers or tiers, as in autosegmental approaches, this tier would be separate from the one specifying phrasing and accent choice. This means that these meanings should not affect category assignment of a given accent. That is to say, an accent L*H should remain L*H regardless of the size of the pitch excursion or the slope of the rise. This is the assumption of, for example, Hirschberg and Ward (1992) or Gussenhoven and Rietveld (2000), who investigate how changes in scaling affect pitch accents within categories. However, Dilley and Brown (2007) give experimental evidence
CHAPTER 2. LITERATURE REVIEW

that relative pitch levels can change the percept of peak timing and therefore the category identification of pitch accents. Also, researchers in the British tradition distinguished some nuclear accent configurations on the basis of range (e.g. low rise versus high rise, cf. Crystal 1969, Bolinger 1970). This adds additional complexity to the question of intonational meaning, in particular, because characteristics like the size of a pitch excursion must both be taken in context and assumed to affect category perception as well as emotional perception. It is unclear to what extent they may affect both simultaneously.

In some cases, the categorical shift itself may be considered to change affective interpretation; this is the case, for example, in the KIM’s contrast between medial and late peak pitch accents (Kohler 1991, 1995). The IPO model, which does not make any explicit claims about the meanings of the different contours, would implicitly also allow affective meanings to be among those which might be expressed (’t Hart et al. 1990). Auto-segmental models, on the other hand, generally separate affective meanings from linguistic (specifically pragmatic) meanings in intonational modelling, so that the only meanings directly specified in the phonology are the linguistic meanings, although the affective meanings may influence interpretation. Pierrehumbert and Hirschberg (1990) claim that “pitch range interacts with the basic meanings of tunes to give their interpretation in context... [Affective information] may make some derived interpretations of the tune seem more plausible than others” (1990: 280). They point out experimental evidence suggesting that speaker emotion is not the best way of organizing a meaning system for intonation; Ward and Hirschberg (1985) found that a variety of affective meanings were derivable from one pitch contour (L*+H L H%), while Ladd et al. (1985) identified pitch range and voice quality as being the primary factors associated with emotion. Therefore, Pierrehumbert and Hirschberg argue that

“...speakers use tune to specify a particular relationship between the ‘propositional content’ realized in the intonational phrase over which the tune is employed and the mutual beliefs of participants in the current discourse. Although the interpretation of any token of a tune type may vary along many other dimensions—voice quality, pitch range, as well as nonintonational features—any instance of a given tune will convey the same relationship” (1990: 285).

This is similar, but not identical, to Gussenhoven’s (1983, 2004) approach to intonational
meaning. Gussenhoven also treats tonal meaning as being both compositional and affected by context. He links meaning to the way in which the pitch accented information is related to the background shared by the speaker and listener. This relationship is not determined by the speaker alone; in Gussenhoven’s (2004) tonal inventory, rises have the shared meaning that the speaker is leaving it up to the listener to decide the relationship of the information to the background (hence the use of rises in many question contexts). Unlike Pierrehumbert and Hirschberg (1990), however, Gussenhoven’s analysis does not assume that the meanings of the tonal morphemes are independent even if each contributes to a compositional meaning. Although he describes pitch contours in terms of tones, and draws generalizations across contours (particularly nuclear contours) with similar pitch accents, the contour as a whole can be seen as bearing meaning. Despite this difference, all of the relevant meaning contrasts are defined in terms of tonal activity, just as in Pierrehumbert and Hirschberg’s model.

Tonal choice in Gussenhoven’s intonation model can to an extent reflect affective meanings, namely speaker attitude, through the presence of additional tonal morphemes. Grabe et al. (1997) found that the initial pitch of an utterance (i.e. the prehead, or material preceding the first pitch accent) could affect interpretation of the speaker’s attitude by listeners. Specifically, preheads with the opposite tone to the following starred tone (L pitch preceding H* or H pitch preceding L*) resulted in the perception of a more positive speaker attitude than when the prehead’s tone matched the starred tone. Grabe et al. propose that in this case there is a “polar” tonal morpheme, which has a meaning value of something like “politeness”; it is unclear whether this should be regarded as a linguistic specification or an attitudinal one.\(^\text{13}\) Regardless of the conclusion drawn about the semantic value of the specific tone, Grabe et al.’s study is also noteworthy in that it supports a compositional model of intonational meaning: listeners identified different meanings on the basis of the presence or absence of one tonal morpheme.

\(^{13}\)One complication with this distinction is that many languages have linguistic specifications for politeness. This occurs not only in Romance languages, where the version of “you” selected varies depending on the addressee (and verbs must agree with this choice), but it is present to an even greater extent in languages such as Japanese or Korean, in which the verb form used varies depending on who is being addressed or even referred to. In these languages the “polite” form is clearly a linguistic aspect. This means that it is possible to consider “politeness” as being linguistic in Dutch as well, although the possibility of something being represented in the linguistic structure is of course not a sufficient condition for its inclusion.
2.2.5.2 Semantic value

Taking one step back from the question of compositionality, the question still remains as to the semantic value of intonation as a whole. A variety of different approaches to this question have been taken, some of which were already mentioned above: these range from the decision of ’t Hart et al. (1990) to concentrate on similar forms without explicit reference to function, to the detailed compositional analysis of Pierrehumbert and Hirschberg (1990). Within approaches that have tried to account for meaning, there appear to be two main directions of thought. On the one hand are approaches that see intonation as adding meaning to the lexical-semantic content; on the other hand are approaches that see intonation as constraining possible interpretations. For example, in both the approaches described above (Pierrehumbert & Hirschberg 1990, Gussenhoven 2004) intonational meanings can be seen as “adding” information in that they specify relationships between the utterance and speakers’ beliefs, or between the utterance and shared background. However, other approaches see intonation’s role as limiting or selecting from a variety of possible interpretations. Jill House (1990, 2006, 2007) takes the viewpoint that intonation constrains interpretation, “whether by clarifying syntactic structure or by organizing the information so as to give access to a set of background assumptions... Intonation... [allows] us to eliminate a range of possibilities in favour of one interpretation in particular.” (2006: 1545) House places intonation into a Relevance Theory framework by identifying it as something that makes communication more efficient. In Relevance Theory, the assumption is that by saying something, the speaker is communicating that it is optimally relevant. The listener also assumes relevance and therefore tries to find the interpretation of an utterance that both makes sense and requires the least processing effort to fit the context (Sperber & Wilson 1995).

According to House (2006), intonation guides this process by helping listeners perform the appropriate inferences, for example by focusing (making salient) certain information to mark it as more important for interpretative purposes. Prosodic focus can be defined as the use of prosody to make some element of an utterance more prominent than others. In English, prosodic focus is normally carried out by the placing of a nuclear pitch accent on the word or phrase which is to bear focus. This is not a syntactic function, but a discourse func-
tion: the intonational phrase’s relationship to the discourse is what is marked. Similarly, prosodic boundaries function to block certain interpretations, for example the carry-over of negation beyond its intended scope. Although this can (and generally does) coincide with a syntactic boundary, the goal of the phrasing is the meaning constraint. It is not just the addition of salience which is relevant in House’s framework, but also non-addition; that is, the speaker’s decision not to accent certain utterances gives clues to how the speaker interprets the background or context. This background is essential for inference interpretation. Decisions about accenting and deaccenting are complex and depend on the structural characteristics of the utterance and not simply the information structure, however. Terken and Hirschberg (1994) show that deaccenting only occurs when the grammatical function and surface position of the element in question persist from previous utterances, even when the element is clearly given and should be accessible from the previous discourse.

Focus may serve to indicate, not just how to choose one interpretation from many, but even the possibility that more than one interpretation exists. Rooth (1992) argues that the meaning of focus is to select one alternative from many; but the existence of focus in a given position is also what indicates that a set of alternatives was possible. In other words, it is only when focus is present that the alternatives matter to the discourse. For example, new information may be presented with intonational or syntactic focus (in English and many other languages, though not all). The lack of focus on any given information in the utterance makes sense, because the given information is already settled in the background for both speaker and hearer; no alternatives are necessary. On the other hand, the new information has not yet been added to the background, and depending on the context, comes from a potentially infinite set of propositions (in the sense that new information may not be predictable). Focus therefore logically selects the new element, which is one out of many possibilities. Rooth, like House, sees focus as primarily a device for constraining interpretation, particularly in cases where adverbs also contribute to focus. He gives the example,

(2.2) Mary only [read] The Recognitions. (Rooth 1992: 78)

In this case, although we know what Mary did (i.e. read the mentioned book), the use of focus suggests another possible set of things that she did not do; for example, she may not
have understood the book, or she may not have written it. The interpretation of what Mary did not do, of course, depends on the context of the utterance, and Rooth sees this as a separate pragmatic process.

2.3 Discourse Topic Structure

2.3.1 Definition of topic

Although the word “topic” is something that is relatively easy to intuitively grasp, defining it as a technical term is more difficult, particularly since the term already has technical uses which are not relevant to the current study. Brown and Yule (1983) claim that “topic” may be “…the most frequently used, unexplained, term in the analysis of discourse” (1983: 70); and indeed they claim that actually defining the term is impossible since different interpreters of a text may have different ideas about what the topic is.

While it is the case that different interpretations of topic are always possible, it seems likely that there will normally be a common thread to these interpretations. Therefore, as a first approximation to define “topic” we can use the concept of “aboutness”: the discourse topic is what a unit of discourse is about, whether this is a referent, an idea, or some other element. Chafe (1994) defines discourse topic as an “aggregate” of information that is semi-active at a given time during a discourse. He points out that a topic need not be overt, and claims that whether or not it is verbalized depends on how interesting it is; that is, whether the speaker thinks that the topic will conflict with the listener’s expectations. Different theories of topic structure based on different kinds of texts have often formed more specific technical definitions, which will be discussed in more detail below, but the idea of a set of active information remains relevant to these varying definitions.

It is important at this point to make one distinction. The term “topic” has sometimes been used to contrast with the the term “focus” on a sentence level in semantic literature, as an alternative terminology for “theme” and “rheme” (cf. the Prague School, e.g. Mathesius 1975). In this case, “topic” refers to information in an utterance that is given, while the “focus” is the new information. This definition of topic is not what is intended in the following, although it is not completely separate; see Büring (2005) for a longer discussion of the relationship between Sentence-Topic and Discourse-Topic. It is Discourse-Topic
which is of particular interest for this thesis, since this is the kind of topic which has scope over units longer than sentences.

### 2.3.2 Theories of topic structure

Although topic itself is difficult to define, many theories of discourse topic have made use of a “topic structure” or “topic framework” in order to help describe the organization of a text. This framework might be described as containing all the coherently-related contextual information that is necessarily active for the interpretation of the text (cf. Brown & Yule 1983; Chafe 1994). However, not all this information is necessarily verbalized; the location in which a conversation is taking place may be relevant when a speaker makes a deictic reference to “here,” but presumably the conversational participants are perfectly aware of where “here” is without needing to be told. In investigating the discourse topic structure of a text, what may be most interesting is the way in which these topic structures are built and changed.

In the investigation of the prosody of discourse structure, researchers often perform the somewhat circular task of looking for prosodic phenomena which can be correlated with the discourse structure, in order to make claims about the discourse structure itself. Research into the prosody of topic structure has tended to take one of two discourse-theoretical directions, a category-based approach or a hierarchy-based approach. These two directions will be reviewed here in greater detail. It is important to note that the type of theory of topic structure developed is not independent of the type of text used to develop it, so that although these two directions cover a large amount of ground, they may not have as their scope all types of text.

The basic difference between a category-based approach and a hierarchy-based approach is the number of different kinds of functions that are available within the topic structure. A category-based approach assumes that there are a variety of utterance functions, which can be identified and drafted into use in appropriate contexts. A hierarchical approach, on the other hand, assumes that there are too many possible functions for definition, and that it is only possible or reasonable to identify relationships between utterances in a more general way.
2.3.2.1 Hierarchical theories

One commonly-used kind of topic structure theory takes as a starting point the idea that there are too many functions in language to be adequately grouped into just a few categories, and that instead it is more useful to describe the relationships between utterances more generically. Within research on prosody and topic, the most widely-referenced work in this area is Grosz and Sidner’s (1986) paper outlining a series of relationships which can hold in discourses and the reasons for them.

The basis for Grosz and Sidner’s work is Relevance Theory (Sperber & Wilson 1995, cf. section 2.2.5.2), which has as its basis an assumption of maximum efficiency; that is, speakers attempt to convey the most new information with the least amount of effort. This is done by the consideration of relevant information; all information in the discourse is considered to be relevant to the discourse on the basis that the speaker chose to communicate it, and the speaker is assumed to use the most economical method possible to do so (cf. Grice’s 1989 Maxim of Quantity). Sperber and Wilson’s theory is functionally-oriented in the sense that they are interested in the task that the speaker is trying to accomplish (cf. Drew 2005), although instead of the term “task”, which they believe is too limited, they use the term “intention”.

Since there are so many possible intentions that a speaker could be trying to accomplish, both in the sense of interaction with an interlocutor and in the real world, they do not believe it is possible to make an exhaustive list of intentions. They argue instead that regardless of how many kinds of intentions may exist, there is a limited set of relationships that can hold between them.

According to Grosz and Sidner (1986), discourse structure has three components: the linguistic structure of the utterance, the intentional structure (i.e. the task(s) the interlocutors are trying to accomplish), and the attentional state, which is an abstraction representing which items are most available or salient at any given time in a discourse. They see discourse as being composed of discourse segments; an utterance is to a discourse segment what a word is to a phrase from a structural standpoint (1986: 177). Discourse segments are therefore like phrases in the sense that they fulfill functions with regard to the overall discourse. However, Grosz and Sidner note that there is no strict decompositionality of

\[14\] However, this term can be understood in the same way as a conversational “task” in other theoretical terms.
discourse segments; parts of discourse segments may be separated from one another (e.g. by a parenthetical utterance) and individual utterances can occur at the same hierarchical level as a discourse segment. In other words, the properties of the discourse are a function of the properties of the discourse segments and the properties of individual utterances. The meanings of the discourse and the discourse segments/utterances are related in a somewhat circular fashion; the discourse structure affects the utterance meanings, but utterance meanings also affect the discourse structure.

In terms of discourse structure itself, Grosz and Sidner focus mainly on what makes a discourse coherent. The Discourse Purpose (DP) is defined as “the intention that underlies engaging in a particular discourse,” while a Discourse Segment Purpose (DSP) comprises “how this segment contributes to achieving the overall purpose.” (1986: 178) Despite these slightly differing definitions, DPs and DSPs are considered to be essentially the same thing: a purpose for part or all of the discourse, which is (normally) intended to be recognized by the interlocutors as such. This is not the same thing as having the purpose be produced explicitly in the discourse, although this is certainly a possibility. Blakemore (2002), following Giora (1997) and Wilson (1998), believes that local hierarchical relationships such as those proposed by Grosz and Sidner (1986) are sufficient to deal with the notion of global unity of texts; she claims that the idea of topic itself can be derived from or defined on the basis of local relationships (2002: 159). In other words, in Blakemore’s approach, topic is no more than an artifact of other organizational principles. Grosz and Sidner do not make this claim, but their theory allows it to be a possibility, unlike a categorical theory, in which the topic must be known since the categories of other utterances are defined in relationship to it.

Since DPs/DSPs can be essentially any purpose (see definition of “intentions” above), the structure of the discourse is described in terms of the relationships that can hold between purposes. These relationships are of a limited number: there can be dominance relationships, where if one DSP contributes to another DSP, then the latter DSP dominates the former; and satisfaction-precedence relationships, where one DSP must be satisfied before another DSP can be satisfied; satisfaction entails a kind of completion or “case closed” in the interpretation. This latter relationship is particularly relevant in task-oriented settings, where one portion of the task must be completed before another: for example, slicing bread.
before spreading butter on it.

The role that DPs and DSPs play in discourse organization is to modulate the attentional state. This is a property of the discourse rather than of the participants, and is modelled as a set of “focus spaces” tied to each discourse segment which contain the DSP as well as any salient entities from the segment. Focus spaces are local in the sense that they contain only information relevant to the discourse segment they are linked to, rather than information relevant to the whole discourse. The organization of these focus spaces is a stack, where focus spaces that are higher in the stack are more accessible than lower ones, and items that have been “popped” from the stack (i.e. by discourse segment completion) are no longer accessible. The focus structure constantly evolves, as the intentional structure adds or removes items from the stack by introducing new DSPs and then satisfying them; it does not exist in any a priori fashion (Grosz & Sidner 1986).

The focus structure can be interpreted as representing the discourse structure from an internal perspective; that is, the structure as it stands at any given time point in the discourse. The intentional structure, on the other hand, exists as a whole, at least by the end of the discourse, and can therefore be understood as the discourse structure as seen from the outside. It is the focus structure which Grosz and Sidner consider to influence processing, while the intentional structure cannot. The intentional structure might be better understood as the source of metalinguistic judgments, however.

Grosz and Sidner (1986) do not make explicit predictions about the prosodic representation of the discourse structure. Later work by Hirschberg and Grosz (1992) does; they find evidence that discourse structure as predicted by their hierarchical structures has co-occurring intonational variation, and they show that a variety of intonational features may be used to signal the structure. It is important to note in relation to these results that the status of the focus stack rather than the status of the intentional structure influences the prosody, since it is the focus stack which has links to the on-line processing of the discourse. This means that prosodic decisions would be made on a relatively local basis, depending both on how many items were already in the focus stack, and how many were to be removed (or “popped”, to use Grosz & Sidner’s term) or added. In other words, it is essentially impossible to define one “level” of the hierarchy as the level of topic in the same way that a categorical theory would identify a topic; local relationships rather than global structure...
2.3.2.2 Categorical theories

An alternative to the hierarchical approach to topic structure is an approach in which discourse relationships are closely tied to the semantic content of the utterances. Wichmann (2000) points out that although the number of potential relationships that can hold between elements in a discourse is large, for a specific discourse or type of discourse it is likely that only a small set of relationships will be used. This means that it is possible to define discourse structure on the basis of categories into which different utterances can be grouped. She notes that although different authors have used different names for these categories, there tends to be a great deal of similarity between the categories that are identified, especially for similar types of discourses. Halliday and Hasan (1976) and van Dijk (1977) see texts as being composed on a macro level of elements with over-arching functions: “high-level propositions or macro-propositions” (Wichmann 2000: 15). These high-level functions correspond to categories within the topic structure of a discourse.

Wichmann’s (2000) choice of topic structure categories for her data follows on from a study by Nakajima and Allen (1993). Nakajima and Allen’s speech data comes from the TRAINS corpus (Allen & Schubert 1991), and consists of task-based dialogues in an imaginary world wherein one speaker tries to organize manufacturing and shipping, while the other speaker provides information and assistance about the state of the imaginary world; the speakers sat in different rooms and communicated through microphones and headphones. Nakajima and Allen identify four main categories of topic-structure boundaries, some of which may be divided into subsets (all direct quotations from Nakajima & Allen 1993: 201):

- Topic Shift
  - Topic Development\(^{15}\)
  - Interruption

\(^{15}\)This appears to be something of an odd man out in the categorization; it might be better seen as a subcategory of the Topic Continuation.
- Topic Continuation: “The linkage between the current topic and the previous one is comparatively strong.”

- Elaboration
  - Elaboration: “The current utterance adds some relevant information to the previous statement.”
  - Clarification: “The current utterance clarifies some propositions made in the previous utterances.”
  - Summary

- Speech-act Continuation: “A single speech-act\textsuperscript{16} continues over several UUs.”

Nakajima and Allen believe that the type of boundary (classified on the basis of the four overarching categories) has a primary effect on some of the prosodic parameters that they measure. On the basis of the scale of their prosodic measurements (cf. section 2.3.3 below), they order the categories as given above, with Topic Shifts being the most saliently marked, and Elaborations the least saliently marked. They make a distinction between local and global discourse structure:

“The global level is concerned with topic changes, that is, the discrimination between TS [Topic Shift] and TC [Topic Continuation]. The local level corresponds to the identification of the fine structure of UUs [Utterance Units] which are uttered for the same discourse goal (by the same speaker). This level of identification ought to include not only the relation between UUs but also the hierarchical structure of UUs.” (1993: 207)

This distinction is essential when examining other studies of topic structure, which have often focused on the global level while neglecting the local level. Nakajima and Allen also make the important point that a categorical model of topic structure does not exclude the existence of a hierarchical structure as well; however, the hierarchical structure is not the main element creating or constraining the coherence of the discourse.

\textsuperscript{16}i.e. a single communicative function
Wichmann’s (2000) study adopts similar categories to those used by Nakajima and Allen (1993), although she adapts them to her texts, which are news stories read by professional BBC newsreaders. These texts may be considered “spontaneous” to the degree that they are not practiced; however, the readers had a great deal of experience in reading, and can be expected to differ a great deal stylistically from the nonprofessional speakers and setting in Nakajima and Allen’s study. Wichmann’s categories are as follows (2000: 116):

- New Topic
- Additional information, same topic
- Elaboration/explanation of previous sentence
- Reformulation of previous sentence
- Metatextual comment (not relevant for the topic-structure prosody analysis she proposes, and will not be discussed further)

Wichmann (2000) makes explicit that she believes that these categories can also be defined at least in part by hierarchical relationships. A New Topic is always the highest level in the hierarchy; an Addition is seen as a coordinate to New Topic, while Elaborations are subordinate. Reformulations are subordinate to whatever they reformulate so are more difficult to give a definitive location in a hierarchy. However, Wichmann ends with the final order of prominence as follows: New Topic, Reformulation, Addition, Elaboration (2000: 117). This ordering is ultimately based on relative newness of information, which Wichmann takes as a general principle for this kind of organization.

The key difference between category theories and hierarchical theories is that the latter are more neutral as to the semantic content of the utterances when it comes to the structure. A hierarchical structure could be seen as taking into account only a minimum of semantic information for organizational purposes. A categorical theory, on the other hand, is intimately tied to the content of the utterances themselves: either to their relative newness (as suggested by Wichmann 2000) or to some other kind of informational relationship. This

---

17 It is impossible to be completely neutral to the semantic structure, otherwise there would be no way to determine which elements dominate or are subordinate to other elements.
means that the specific content of what is said matters more to a categorical theory; this is unsurprising, given that, as already pointed out, the categories used probably depend at least to some degree on the specific discourse, or at least the type of discourse.

The requirement of having to use different types of categories, choosing from a potentially infinite set, could be seen as a weakness of categorical theories, and indeed this is one argument made by proponents of hierarchical theories, cf. section 2.3.2.1. However, the requirement for a variety of kinds of categories could also be seen as a strength, as it reflects the variety of types of conversational interaction that may occur. One approach to dealing with the large set of possible categories in speech is that taken by Conversation Analysis (CA). The basic assumption of CA is that the study of conversation is the study of “the actions and activities through which social life is conducted” (Drew 2005: 74; cf. Schegloff 1996), and its goal is therefore to investigate and identify the ways in which people participating in conversational interactions behave and respond to one another’s behavior in these contexts. Conversational sequences are organized around different functions, similarly to what has been proposed for the topic structure organization above, in that it is the functions that determine the shape of the interaction. The functions identified in CA are things like agreeing or disagreeing with an assessment or issuing and responding to an invitation, and as suggested by these examples are often investigated in the context of adjacency pairs (cf. Sacks 1992). Adjacency pairs are pairs of conversational actions (not necessarily utterances) which often occur together (e.g. an invitation and a response to the invitation), and conversational meanings are encoded not only by the content of the utterances but the form they take, and specifically the extent to which they correspond to expectations. If the second half of an adjacency pair does not meet expectations, or even is not produced, this is meaningful in and of itself in the conversational context. Positive actions, defined as those enhancing “social solidarity” are preferred over negative ones (Heritage 1984; Schegloff 1988) and tend to have unmarked forms, while negative actions tend to be marked. This is a consistent enough pattern that it is possible to make the assumption that when something appears in a phonetically marked form, it is a dispreferred response in some way, even if its content might not immediately suggest this interpretation. Dispreferred responses can sometimes be used to change the direction of a conversation, as in the case of a “conver-

---

18Wichmann (2000) sees her work as falling within this general scheme of research.
tion restart” after talk about a problem (Sacks 1976, cited in Jefferson 1984). This restart is essentially a change of topic, though Sacks interprets it as literally a new conversation, not just a topic shift.

Topic changes in CA are relevant not in and of themselves as propositional content, but rather as a social/conversational function. They have been examined in depth by Jefferson (1984, 1993) in a variety of conversational contexts. Jefferson (1984) focuses on topic changes after “troubles-telling”, or talk about problems. She notes that while new topics are often disjunctive (as is assumed in most theories of topic structure) and may show that the topic of previous conversation is somehow problematic, this is not always the case. Instead, there is an alternative type of topic-shift that moves gradually away from one topic into another, in what she calls “stepwise transition”. In this kind of shift, an interlocutor takes advantage of a sort of middle-ground topic which is connected to both the old topic and the new one, making conversational moves to “stabilize” the intermediate topic as the main focus of the interaction, then taking advantage of a later “pivot” utterance to lead into a different topic altogether. A key feature of this kind of shift is that it is not marked for topic change; in particular, it is not marked as a topic-closing by any lexical means, such as a summary assessment. Jefferson (1984) suggests that this is because the conversational participant, having run out of things to say about the troubles-telling topic, does not want to end the conversation, and therefore uses a method of maintaining the conversation while shifting the topic to something new. She argues that some types of topic are open to topic change via an immediate introduction of a new topic, while other kinds of topic (like troubles-telling) are not, and that the type of topic therefore influences how topic change is conducted in conversation. In a more canonical or direct topic shift, signalling the closing of a topic may be as simple (from a lexical point of view) as providing an acknowledgment or assessment of the interlocutor’s previous contribution (Jefferson 1993).

The CA approach to topic is not identical to a category-based approach in that it defines “doing topic” as a social/conversational function rather than as a part of the linguistic structure in the sense of syntax, morphology or phonology. However, it is similar to a categorical approach in that the interlocutors’ purpose(s) for the interaction define how topic is used. Furthermore, the evidence that different kinds of topics require different kinds of shifts is similar to a categorical approach to topic in that the internal content of the utterances (in this
2.3.2.3 Topic in narratives

The experiment reported in chapters 3-6 uses a narrative text, which has traditionally been described in a different manner than the hierarchical or categorical theories discussed above, and often with a completely different level of organization in mind.

Labov and Waletzky (1967) argue that narratives are possibly the only kind of spoken discourse with a definable beginning, middle and end. They describe the elements of a narrative abstract, orientation, complicating action, result/resolution, evaluation, and coda. It is clear that these may be considered macro-elements of a narrative; regardless of whether the narrative is short (e.g. relating an anecdote in conversation) or long (e.g. a novel), at least some of these elements must be present in order for the narrative to be well-formed.

These categories describing a narrative do not directly involve topic structure. Chafe (1994) suggests that the “topic” of a narrative coincides with the climax of the narrative; however, as Chafe himself acknowledges, a longer narrative might be expected to have more than one topic, so the relationship between the elements of narrative described above and the topic(s) of the narrative is not immediately clear. Chatman (1978) suggests that narratives also have a hierarchical structure, with “kernel” parts of the narrative (perhaps the macro-elements described by Labov & Waletzky 1967) dominating “satellite” parts. The accomplishment of macro-elements of a narrative might therefore be reconcilable with the idea of “discourse purpose” as proposed by Grosz and Sidner (1986), with allowances made for the different text types. Similarly, the topic structure categories proposed by Nakajima & Allen (1993) and Wichmann (2000) might be applicable in the description of micro-elements within a longer text, or even relationships to one topic within a short narrative, given that the news stories investigated by Wichmann (2000) could be interpreted as being short narratives. The relationship between the idea of “topic” and the structure of a narrative will be discussed further in section 7.1.
2.3.3 Prosody and topic structure

Although topic structure has been investigated in detail from the point of view of discourse analysis, it has been studied relatively little from the point of view of prosody. Early work in this area, such as that by Lehiste (1975), did not consider topic specifically as such, but tended to look at what has been more generally termed “paragraph intonation”, without much explicit reference to the reasons why speech might be considered to be grouped into paragraphs beyond the guiding principles from a writing point of view.

The term “paratone”, first proposed by Brown (1977, cited in Wichmann 2000), has been commonly used to describe a unit of speech more or less correlated with a paragraph in written speech, and a number of intonational features, including increased pitch peak height and expansion of the width of pitch range used by a speaker are widely recognized as indicating the beginning of a new paratone (Cruttenden 1986). In a rather circular definition, one way that the paratone has been identified is as the domain of “supradeclination” (Lehiste 1975, Silverman 1987). Supradeclination is the paragraph-level equivalent of declination in utterances, the phenomenon whereby the overall pitch level of a declarative utterance decreases over the course of a sentence without any perceptual effect of decreased prominence. Just as pitch gradually lowers slightly over the course of an utterance and resets at the beginning of the next, this process occurs in a larger scale over longer units, although just as in the case of sentences, the exact amount by which (supra)declination occurs is difficult to quantify (cf. Vaissière 1983).\(^\text{19}\)

Taking a paratone as a unit of speech production and/or perception means that we can look for both indications of the locations of its edges (boundary cues), as well as unified characteristics existing throughout the paratone (internal cues), which may help to signal to a listener that it is to be understood as a unit. Supradeclination could be considered an internal cue, if we assume that it operates more or less evenly across the course of the paratone. Things like pitch reset, on the other hand, might be considered boundary cues, whereby the edges of the paratone are marked. These two kinds of cues do not necessarily operate independently of one another, particularly from a perceptual point of view; supradeclination can

\(^{19}\)This is in part because of debate as to whether declination should be considered a phonetic phenomenon or accounted for by the phonological function of downstep; see e.g. Ladd 1996 for a discussion of the issues involved.
lead to pitch ending low at the end of a paratone, which then creates the maximum salience for pitch reset at the beginning of a following unit by increasing contrast between the two.

Most prosodic topic-structure research has been done on the prosodic correlates of paragraph and/or topic unit boundaries, primarily in English and Dutch. Some of this work has grown from a discourse analysis or conversation analysis perspective, other work from a phonetic perspective, and still more from a speech technology perspective. The results of research from these perspectives has converged well in many cases to present a reasonably unified picture of at least some cues to topic structure. The following sections review a number of experimental approaches to identifying topic structure cues, both boundary and internal, in prosody. It is important to note, as do Hirschberg and Grosz (1992), that the cues identified below do not occur in isolation; in most cases there are a variety of prosodic features that occur in relation to the discourse structure.

### 2.3.3.1 Boundary cues

Most of the research done on the prosody of discourse structure has focused on boundary cues. Boundary cues can signal the beginning or the end of a paratone or topic unit. Because of this emphasis on boundaries, this research has also tended to center around global rather than local topic structure (cf. section 2.3.2.2).

**Pitch reset**  The most commonly recognized initial boundary cue for a paratone, as previously mentioned, is pitch reset: a speaker beginning a new unit will use a higher pitch (and a wider pitch range) than what was previously in use to signal that the new unit has started. The term “reset” refers to the idea of (supra)declination; that is, the declination that has been occurring across a unit has now been reset to some initial state. Lehiste (1975) was the first to show pitch reset occurring in an experimental setting; using a series of sentences that could be placed in paragraphs in different orders or produced in isolation, she found that extra high pitch was the most reliable prosodic correlate of what she calls a “conceptual paragraph”. She also found that readers started sentences on a higher pitch if the sentence was at the beginning of a paragraph, compared to both paragraph-medial and isolated sentences. In a listening test, she found that listeners could reliably identify which sentences had been produced as paragraph beginnings, compared to other positions. This result has
also been reported for spontaneous English speech by Brazil et al. (1980), Brown et al. (1980), and Yule (1980).

A similar study to Lehiste’s was conducted by Sluijter and Terken (1993) for Dutch. Again, target sentences were placed in a variety of positions in texts and two male subjects read them aloud, as well as reading the sentences in isolation. As in other studies, they found that new paragraphs in Dutch also begin with high pitch, about 25 Hz higher than the next position in the paragraph. Sluijter and Terken compared the production of the same utterance in four positions in a paragraph: the first, second, fourth, and fifth (i.e. last) position. They found a consistent lowering of the pitch topline across these four positions. Isolated sentences behaved similarly to paragraph-initial sentences in the case of the pitch topline onset, though their baseline pitch was closer to that of sentences in second position in a paragraph, suggesting that there were multiple cues: the baseline and the topline apparently varied independently. Sluijter and Terken also found that including paragraph intonation provided a small but statistically significant increase in the rated naturalness of synthesized speech.

Similar results have been reported for a number of other languages. Bruce (1982) and Swerts et al. (1996) reported that pitch reset and a following pitch fall signal the beginning of a new topic or important information in Swedish. Botinis (1989) reported pitch reset at new topic beginnings in Greek. Oliveira (2000) showed that pitch reset was present at “narrative boundaries” in Portuguese, with pitch values being about 25 Hz higher after these boundaries than in other contexts. Thorsen (1985) similarly reports pitch reset at paragraph beginnings in Danish. Pitch reset also appears to be a cue to discourse structure even in tone languages. Tseng et al. (2004, 2005) identify pitch reset at the beginnings of “paragraph groups” in Mandarin as well as pitch lowering at the ends of these groups; they consider these boundary locations to be the only places where pitch information is taken into account in relation to discourse structure (2005: 289). Following on from this study, Xie et al. (2007) show that with normalization to speaker and tone, pitch reset is a useful acoustic cue for automatic segmentation of spoken Mandarin discourses.

**Other boundary cues** Swerts and Geluykens (1993), Swerts et al. (1994) and Swerts (1997) also examined boundary characteristics of topic structure elements in Dutch. Swerts
and Geluykens used spontaneous speech data from a corpus of instructional monologues. In addition to F0 declination across utterances (see section 2.3.3.2), they investigated the pause structure of these dialogues, and found that pauses were longest at “crucial points in the information flow” (1993: 193), specifically at topic shifts. However, they found that listeners were not very sensitive to pause length as a cue to topic structure when melodic information was removed, so they concluded that the F0 variation was more important. Swerts et al. (1994) investigated cues to “finality”, which they define as the likelihood that an utterance is the last one in a discourse unit. They found that pitch register (i.e. location of baseline), pitch range (i.e. size of pitch movements), and contour shape (in this case, ending in a rise versus ending in a fall) can all affect the degree to which listeners interpret an utterance as being discourse-final. Specifically, a lower register and a smaller pitch range both contribute to the percept of finality. Contours ending in falls were far more likely to be perceived as final, and pointed hat patterns were perceived as more final than flat hat patterns in the same contexts, although this appears to be due to a timing difference between the two conditions rather than the shape of the contour.

The two studies above reported variation primarily in the global, binary condition of topic change versus topic hold. Swerts (1997), on the other hand, was interested in more degrees of variation than this simple contrast; that is, in the local topic structure. He asked two sets of participants, one reading an unpunctuated text and one reading and listening to it, to mark paragraph boundaries, without providing a definition of “paragraphs”. He then counted the number of boundaries marked at each location to come up with an index value for the strength of the boundary. Listeners were more consistent in their responses than participants who only read the text, and Swerts found that the strength of the boundary marked by rater agreement was consistent with several prosodic cues: longer pauses preceding the boundary and very high F0 in an accented syllable following the boundary were associated with stronger boundaries, and the choice of a final low boundary tone (as opposed to nonlow, either mid or high) was also more likely at stronger boundaries.

A multiple-characteristic approach to discourse structure boundaries was taken by Herman (2000), who studied prosodic characteristics of final boundaries. Her study, unlike previous investigations, controlled for a wider degree of phonological characteristics, including the nuclear pitch accent at the end of the utterances under investigation. By com-
paring only pitch accents in the same category, she was able to avoid a source of variation in her data (although she does not report on patterns in the choice of pitch accent, which might have been interesting in and of itself). Herman compared variations in RMS amplitude, length, and F0 across different tokens of the same pitch accent in discourse-medial and discourse-final position. She found that lower F0, lower RMS amplitude, and greater degrees of lengthening were present in discourse-final position than in discourse-medial position for the same accent pattern. Interestingly, her results are not as robust as those of previous studies; she attributes this to having controlled her data more stringently than those studies from a phonological point of view, though this seems to be counter to what might be expected. It is possible that her weaker results are due to a flaw in her methodology; while she compares production of the same utterance in different contexts, with the attempt to force the medial/final interpretation via orthography, a review of her materials suggests that some of her medial contexts seem to allow for a final (i.e. topic) boundary, which introduces a potential confound into her study. It is of course difficult to know to what extent this would have affect her participants, but assuming they were sensitive to it, it would obscure differences between the two positions, especially if the difference between the productions is gradient.

Herman makes the interesting point that variations between discourse-medial and discourse-final utterances begin at linguistically relevant points; that is, the variations she observed occur during (nuclear) pitch accents or at boundary locations, not at locations which might be considered to be interpolations between tones, for example. The modification of linguistically-relevant points is evidence that the variations observed are relevant for linguistic understanding; that is, they are not simply non-linguistic results of the Biological Codes (cf. Gussenhoven 2004). Instead, there appears to be evidence for a discourse-organization system which influences the (phonetic) output of the phonological system.

2.3.3.2 Internal cues

I have identified boundary cues as those which fall primarily at the edges of the utterance in question, although the “edge” can be used in more than one sense; it may be a phonological edge (i.e. an initial pitch accent) rather than the phonetic edge (i.e. something immediately following a silent pause). Internal cues, on the other hand, are cues that persist in some way
CHAPTER 2. LITERATURE REVIEW

over the whole course of the utterance.

Supradeclineation  The most obvious internal cue in topic structure organization is supradeclineation (cf. section 2.3.3), which probably interacts with the boundary cue of pitch reset. Supradeclineation, or the gradual lowering of pitch across a topic unit, is one way in which speakers show that the unit of speech has not changed. For some languages, supradeclineation is considered to be the primary cue to topic grouping, rather than pitch reset. For example, Thorsen (1986, reported in Grønnum 1998) claims that this is the case in Danish, because the lengths of the sentences involved impacted the pitch values measured; she proposes that a minimum slope requirement is present for declarative sentences, and that sentence boundaries are affected more by this than by the discourse structure. This may also be the case for tone languages, in which individual pitch points may bear too heavy a functional load to be individually relevant for topic (though see Xu & Xu 2005 for an indication that this is not the case).

Other internal cues  Hirschberg and Grosz (1992) investigated prosodic variation in English in relation to the predictions made by Grosz and Sidner’s (1986) theory of topic structure. They found a number of prosodic variations which could be considered to be internal cues, namely difference in pitch range between utterances and differences in speech rate. While the speech rate variations were only relevant between new-topic and non-new-topic utterances (cf. also Brubaker 1972), with new topics produced more slowly, the pitch range variations were present in almost all of the discourse relationships they investigated, with superordinate categories having greater pitch ranges than subordinate ones.

Wichmann’s (2000) study of a corpus of BBC newsreaders suggests a more detailed version of supradeclineation variation related to topic structure. She posits that in addition to the supradeclineation occurring over the course of a paragraph-sized unit, there are also varying pitch range rules related to different topic structure categories. In this framework, there would be a general overall pitch declination across the course of a long discourse, but different topic structure categories would also show varying degrees of pitch compression (relative to the declination) over the course of the utterance. This would provide a cue to topic structure by means of the relative pitch range in relation to what might be “expected” in
that position in the paragraph (cf. CA’s concern with the breaking of expectations, section 2.3.2.2). Wichmann does not make any statements about what might be considered the neutral point of comparison; presumably listeners are expected to be able to make some kind of prediction based on the speaker’s potential pitch range and the length of the discourse (as well as the additional information contained in the semantic content, which the prosodic cues would support).

Another internal cue to topic structure organization is also related to declination. Nakajima and Allen (1993) measure a variable they call “peak ratio”, defined as “the ratio of the current utterance’s maximal peak F0 to that of the previous one.” (1993: 205) They found that peak ratio was highest (around 1.15, i.e. the current peak was 1.15 times higher than the previous peak) in cases of topic change, whereas other topic-structure relationships had lower peaks (at or below 1), with Topic Continuation and Speech Act Continuation having ratios of around 0.95 and Elaborations having ratios of around 0.89. This is a relatively simple way of measuring whether global pitch is going up or down, though it is not without problems. The peaks measured were simply the highest peaks in the utterance, without taking into account any phonological or semantic information (e.g. nuclear versus non-nuclear peaks; contrastive focus). This means that to a certain degree the peaks may not have been directly comparable. On the other hand, it does allow the utterances to be grouped in a very global way, since regardless of the location of the highest peak, it must fit into the overall pitch context of the utterance containing it. With reference to Herman’s (2000) finding that it is linguistically relevant points that undergo modification (at least in terms of signalling discourse-finality), the variation in the peak ratio could suggest a modification that is applied to the intonational tier of a whole intonational phrase.

Internal cues to topic structure have been somewhat less studied than boundary cues, since they assume both the existence of an internal/local discourse structure and potentially that this structural information is relevant throughout the course of an utterance. Therefore there are still a number of knowledge gaps in terms of the kinds of prosodic characteristics which might be used as internal cues; it seems likely in particular that non-intonational cues might be a useful source of information for listeners.
CHAPTER 2. LITERATURE REVIEW

2.4 Implications and Research Aims

The literature reviewed in the preceding sections reveals a rich system of functions of prosodic and specifically intonational information in creating or constraining both linguistic and non-linguistic meanings in communication. It also suggests that there is structure in language above the level of the utterance and that prosody has been shown to be a cue to this structure at least in certain contexts. However, there is still a large amount of research which has not been done in these areas.

First, investigation into (particularly non-intonational) prosodic features of topic structure could be extended a great deal. Many of the studies reported have investigated boundary cues without taking into account details of discourse structure beyond topic shift and topic hold, or even in some cases without making explicit claims about topic at all and simply identifying grouping characteristics (which may be interpreted as elements of the global discourse structure). However, the more detailed claims about topic organization discussed in section 2.3.2, which have already been shown to an extent to have a good prosodic motivation as well as a structural motivation from a content point of view, indicate that investigations of the prosody of long units must take topic structure into account in a more sophisticated way.

Second, the investigation of prosodic features of topic structure has been largely, though not entirely, centered around intonation, and in particular F0. This is in part due to the fact that F0 is better understood than a number of other prosodic characteristics, and perhaps also due to the fact that it appears to be more salient than these other features, which may co-vary with it. However, a number of the studies cited above provide evidence that non-intonational prosodic features can and do vary independently of the F0 variation, and that these variations can be meaningful in terms of interpreting the structure of spoken discourses (e.g. pause structure). It is therefore necessary to expand the investigation of non-F0 features of the speech signal independently, and to examine how these features interact with F0 and each other both in terms of combining to signal topic structure variation, and in terms of affecting how the cues are perceived by the listener (e.g. if one cue may emphasize or cancel out another).

Finally, with regard to topic structure, it is important to ask the question of to what extent
listeners are truly aware of and sensitive to this kind of structure in everyday language. CA approaches focus on this question by looking at how interlocutors respond to one another’s linguistic behavior, but tend towards qualitative rather than quantitative investigations, making it difficult to generalize their results. Most studies explicitly focused on topic structure have asked participants to identify this structure in a more or less metalinguistic way, which raises the possibility that listeners only use this kind of cue when necessary for a specific task, rather than in speech perception in general (although the consistency of production cues in the studies reported does at least provide a basis for this assumption). Further research is necessary to determine to what extent the prosodic cueing of topic structure in long discourses is really considered relevant by listeners.

This thesis addresses these issues by means of a production experiment and a perception experiment dealing with prosodic variations related to topic structure. The specific research questions addressed are:

1. What role do F0 and other types of prosodic variation play in cueing topic structure in SSBE? In particular, do previous findings hold when more sophisticated models of phonetics, phonology and discourse structure are brought together?

2. What is the relationship between intonational (i.e. F0) variation and variation in other prosodic characteristics in terms of signalling the discourse structure? How do different prosodic characteristics relate to or interact with one another?

3. Are listeners sensitive to the kinds of prosodic variation observed in the production study? Do they use topic-structure cues to guide their interpretation of utterances during the normal listening process, or are they only accessed post-hoc if a metalinguistic task is required?

4. Does the observed prosodic variation motivate the existence and form of the proposed model of discourse structure? Would the patterns of variation observed be better accounted for by a different type of topic-structure model?

5. Should topic-structure-related prosodic variation be characterized as linguistic or non-linguistic? Specifically, can the discourse topic structure itself be thought of as forming part of the linguistic system, or is it better described as representing general com-
The production experiment reported in chapters 3-6 directly addresses (1) and (2), while the perception experiment in chapter 8 addresses (3). These experimental results alongside the validation study in chapter 7 provide the basis for a discussion of (4) and (5), taken up in chapter 9.
Chapter 3

F0 Cues to Discourse Topic Structure

3.1 Introduction

As we have seen, House (2006) proposes that intonation has three uses: indexical, which is paralinguistic and speaker-oriented; grammatical, which can affect the propositional content of the utterance; and discourse, which is to do with interaction with an interlocutor. However, House cautions, it is difficult to know whether all of these uses are active at once, and how they might interact. One of the continuing tasks of intonation research is to separate out these threads in speech communication.

3.1.1 Intonation and discourse structure

While grammatical uses of intonation can affect the informational content of the utterance, discourse uses are concerned with coherence: that is, how the elements of a text or discourse are related to one another, and how those relationships are identified by the listener. Discourses can and do extend well beyond the length of a single utterance, or even beyond several utterances. For this reason, it is also important to consider the ways in which a discourse is structured beyond having coherent individual utterances. Focus structure and the identification of new and given information can contribute meaning that is relevant across sentences, but it is possible to find phenomena that have larger scope as well. Falling within the realm of above-sentence-level discourse structure is topic structure.

By topic, I mean in a very general way “what the discourse (portion) is about”; see section 2.3.1 for more discussion of this definition. When dealing with topic, particularly
in spoken communication, it is unclear what an appropriate name for a topic unit is. In written texts, children are traditionally taught that a new paragraph begins when the topic changes; however, in practice, written paragraphs are as often modified to make the visual task of reading easier as they are grouped in ways to do with their “aboutness” (cf. Wichmann 2000). In intonation research, topic units have sometimes been called “paratones” by analogy from the term “paragraph”; however, this is problematic due to some of the connotations of the latter term, particularly the aforementioned. I will therefore use the term “topic group” to refer to a set of utterances that share their “aboutness”.

There are a number of different theories about topic structure, but those that have been used in intonation research fall roughly into two groups: those based on hierarchies (cf. section 2.3.2.1, Grosz & Sidner 1986), and those based on categories (cf. section 2.3.2.2, Nakajima & Allen 1993, Wichmann 2000). One of the primary sources of external evidence for the structure of discourse is the way in which speakers systematically vary their production in order to signal this kind of organization. The studies presented below and in the next chapters aim to identify variation in the prosodic realization of speech which can be systematically related to the topic structure of the spoken discourse.

3.2 Production Experiment

3.2.1 Motivation

A number of experiments have investigated various aspects of prosody in connection with topic structure. Lehiste (1979) found for English that units of discourse larger than sentences underwent temporal modification, and that the difference in the length of pre-boundary pauses was more systematic, with longer pauses correlating strongly, though not perfectly, with the identification of paragraph boundaries, which may be understood in this case to correlate with topic boundaries. Lehiste also found some effects of cue trading, with the amount of final lengthening and the presence of laryngealization interacting with pause length in her study. Grosz & Hirschberg (1992) found that variation in speaker pitch range and the timing of fundamental frequency (F0) peaks in English helped listeners to decide where to draw boundaries between groups of utterances based on their topic. Sluijter & Terken (1993) found evidence in Dutch of a supradeciliation over longer stretches of speech, which was
reset at the beginning of a new spoken paragraph. Also for Dutch, Swerts & Geluykens (1993) found that in spontaneous speech, the distribution of low and high boundary tones correlates with topic shift or continuation, and high F0 tends to occur with the introduction of a new referent into the discourse. Swerts et al. (1994) built on this study by noting the contribution of global pitch register and span in addition to the speaker’s choice of boundary tone in the perception of final boundaries, although as this study investigated the production of lists, it may not be directly relevant to topic change. House & Wichmann (1996) and Wichmann et al. (1997) found evidence that F0 peaks are delayed at the beginnings of new topic groups in English, although this delay was not necessarily statistically significant. Herman (2000) found that in American English, the relative finality of discourse units could be predicted on the basis in variations in F0 and RMS amplitude in the final pitch-accented syllable of an utterance, with discourse-final utterances having lower values for both of these measurements than discourse-medial ones. Oliveira (2003) found high pitch reset at the points of greatest discontinuity with preceding material in spoken Portuguese. A corpus study by De Looze & Rauzy (2009) found that pause length alone could be used to automatically classify topic changes in French, although the system was less robust for English. Similarly, Lin & Fon (2009) found that speakers of Taiwan Mandarin relied heavily on pause length to the exclusion of other temporal cues to identify the strength of discourse boundaries; however, their study consisted of nonsense syllables so the relationship of this finding to topic structure cannot be established.

The results of all the studies reported above primarily point towards what could be identified as boundary phenomena of larger units of speech. They are fairly unambiguous in that they are to do with grouping or chunking of speech on a larger scale, although many of these authors point out that speakers can violate most or all of these prosodic practices, and that it is furthermore extremely unusual for two listeners to make identical segmentations of the same discourse. Even if we assume that these larger units have a reasonable amount of stability, however, by identifying their boundaries we have only come partway. As noted before, Intonational Phrases (IPs) are characterized both by external (i.e. boundary) and internal characteristics, and these internal characteristics prove more difficult to pin down.

A few prosodic studies have investigated the internal structure of these units. It seems relatively clear that this structure is not a part of the phonology of a language per se; that
is, we would not expect to find topic-structure variation encoded as part of an intonational grammar. Instead, it is part of the discourse structure. Prosodic studies have therefore looked for phonetic correlates of the way the discourse is organized, particularly the way it is organized around topics. The identification and definition of topics is problematic, but for the moment, the relevant intuition is that the internal structure of these larger units of speech goes deeper than a simple linear pattern. In the same way that a pitch accent late in an IP can be made more prominent than an early one by ignoring pitch downstep or declination, an IP later in a large group can be made phonetically more prominent by ignoring supraclelination, apparently without triggering the beginning of a completely new utterance group (cf. section 2.3.3.2). Nakajima & Allen (1993) found correlation of the initial F0 peak height and final F0 level in English spontaneous speech with four categories of topic structure that they defined (see Table 2.1 above). Wichmann (2000) found a correlation with F0 peak height for all of her topic structure categories, and F0 peak delay for some, also in English. A study in Dutch by den Ouden et al. (2002) found a correlation between pause length and initial peak height and an utterance’s position in a discourse hierarchy. These studies accounted separately for factors of supraclelination, so the effects observed are not simply a result of the position of the utterances within the utterance group. These characteristics may be seen as “internal” prosodic characteristics of the topic structure, although many of them are local phenomena (i.e. they occur locally, not necessarily during the entire stretch of speech over which their meaning may have scope), because they relate to how the longer stretches of discourse interact with one another, particularly within a topic unit. These internal phenomena will be important as we come to the question of what constitutes a topic, because they will potentially have the power to distinguish between different theories depending on their variation patterns.

A variety of cues, therefore, have been shown to be manipulated consistently in relation to discourse, but few studies in a phonetic perspective have taken into account the depth or complexity of the underlying structures. As we saw above, more complex topic structures than simply “topic/non-topic” have been proposed, but it is unclear whether there may be further categorical or gradient variations in the use of prosodic cues which signal these differences. Although the lack of systematic prosodic cues for a more complex structure does not necessarily deny its existence, the presence of such cues in speech would certainly be
a strong support to such theories. The starting point to answer this question must therefore be the investigation of prosodic phenomena in relation to topic structure in speakers’ production.

The following study was primarily an opportunity to investigate how speakers produce prosody in relation to the structure of their discourse. On the basis of previous studies, it was hypothesized that speakers would adapt their prosodic production to reflect four categories of topic structure. The use of categories allows for comparison with previous studies on different kinds of texts that have used topic structure categories (Nakajima & Allen 1993, Wichmann 2000), as well as allowing for the adoption of an alternative means of analysis (e.g. a hierarchical analysis as in Grosz & Sidner 1986) should the categories not prove helpful in describing the text. In particular, it was expected that in falling pitch contours, (1) the timing of F0 peaks relative to the segmental structure, (2) the height of F0 peaks relative to the speaker’s baseline, and (3) the span and/or slope of F0 falls would all be sensitive to the topic structure, with utterances in a higher-level category (e.g. new topic) showing later F0 peak timing, higher F0 peaks, and larger spans/slopes.

3.2.2 Materials

In order to investigate speakers’ production of intonation in discourse, a body of data with a number of essential characteristics was required. First of all, it must be possible to identify the topic structure of the discourse(s). If the structure is very ambiguous or otherwise difficult to define, it will not be possible to draw conclusions about the prosodic features co-occurring with and signaling the structure. Second, in order to allow for valid acoustic measurements, for example of peak timing, the same or similar lexical items must be compared. Otherwise, segmental features such as syllable length or even segment identity can be a source of confounds. This makes the use of spontaneous speech very difficult, as it is unlikely that speakers will use the same words and sentence structures. Even if a highly constrained task is successful in eliciting spontaneous speech that is segmentally comparable, it is unlikely that all speakers would spontaneously produce the exact same discourse structures. Third, despite this difficulty in using spontaneous speech, it is highly desirable for there to be enough flexibility in the language task for speakers to make their own decisions about how to produce utterances, and for there to be variation between speakers.
After all, we are most interested in what speakers naturally do, not what they may have been somehow prescribed to do. Finally, a number of technical requirements must be met, most notably the availability of sound recordings in an appropriate format and of high enough quality to be able to take the desired measurements with accuracy.

A number of available corpora were investigated with regard to their suitability on these criteria. Previous studies of intonation and discourse structure (e.g. Wichmann 2000) have often made use of recordings of newsreaders. However, such recordings would generally not meet the third criterion for flexibility and variation, since normally two newsreaders would not read the same news item, and the second criterion, repetition of lexical items, would vary depending on the content of the items themselves. This latter turned out to be a consistent difficulty in many of the corpora, as did the difficulty in accurately describing the topic structure. On the basis of this investigation, it seemed most suitable to design a task to collect a corpus of data specifically for purposes of exploring the ways speakers use prosody to indicate discourse structure.

Due to the difficulties noted above with the use of spontaneous speech, it seemed preferable to use a read text, written specifically for this experiment. This made controlling the topic structure and segmental characteristics considerably easier. Furthermore, the purpose of the production task was to provide a baseline of data about prosodic features of discourse from which perception experiments could proceed. These later experiments would be able to shed more light on the naturalness of any phenomena found in the read-text experiment.

The text used was a single long narrative, the “Emory story” (see Appendix I). This contrasts with earlier studies, which have sometimes used paragraph-length discourses, either monologues or dialogues, repeating the same utterance in different contexts. I considered this paradigm but ultimately rejected it for several reasons. Firstly, the use of this paradigm has had only varying success in previous studies. Braun (2006) used it to elicit contrastive versus noncontrastive pitch accents in German sentences. She placed her target sentences in the same (paragraph-medial) position in different context paragraphs. One context paragraph was intended to elicit the production of a contrastive accent on her target sentences, while the other was intended to elicit a neutral production. In the utterances that she was able to maintain, the production of different accents was successful; however, Braun reports that she had to discard a number of her recordings. Some speakers were discarded due to
being either too monotonous or too emphatic, and some utterances had to be discarded due to either too many hesitations or speech errors, or because the accent structure of the entire utterance was different in the different conditions.

Herman (2000) also used a similar paragraph-based methodology, with the specific intention of investigating topic groups, and found that only about one-third of her pairs of utterances could be distinguished by listeners as playing different roles in the topic structure. Upon investigation of her materials, it becomes apparent that a number of her dialogue contexts are ambiguous as to their finality in the discourse. The same sentence, although moved to a different location in the overall course of the dialogue, often gives rise to a similarly new or unrelated discourse element. Furthermore, even supposing that her contexts created a topic change (or the introduction of a new discourse segment purpose, since she follows Grosz & Sidner’s (1986) analysis of discourse structure), it is impossible to compare phenomena related to the position of an utterance in one speaker’s overall production versus phenomena related to topic organization, since her sentences often appeared spoken in isolation by one speaker. Since my goal was to investigate the internal structure of longer discourses, phenomena occurring in isolated sentences would be a confound in my data, particularly as Sluijter & Terken (1993) found that isolated sentences show a similar pattern to new-topic utterances in terms of their prosodic features.

Besides the consideration of this flaw in the design of the discourse texts, I also determined that it was likely to be unnecessary to repeat full sentences, since the pitch accents in question would be unlikely to stretch over the course of an entire sentence. Using target lexical items which were long enough to carry a full pitch accent meant that not only could the items be compared across position and topic structure within a paragraph, but it was also possible to compare realizations of the same word in multiple positions within an utterance, for example at the beginning versus at the end. This allows for more flexibility in the analysis, as well as providing another potential way of accounting for variation in the data.

The text was a fictional narrative about an explorer. It contained sentences or utterances falling under one of four topic structure categories, modified from Wichmann (2000) and Nakajima and Allen (1993). The category modifications were made on the basis that the text used was of a different format than either Nakajima and Allen’s study (a spontaneous
discourse giving instructions for a task) or Wichmann’s (newsreaders reading news stories), although the structure of news stories seems likely to have more in common with a narrative than the structure of a spontaneous task-based discourse. The categories used for this study reflect those proposed by both of these previous investigations, but are intended to take into account the fact that the text in question had a different form. They are appealing from an analytical point of view since, if they are useful for describing the structure of a variety of text types, it becomes possible to coordinate analyses of different types of text. This is not true of, for example, the categories proposed for the structure of a narrative by Labov & Waletzky (1967), which are specific to only one type of text and do not lend themselves so easily to generalization. In fact, Labov & Waletzky (1967) argue that a narrative is a unique type of text, being the only kind of spoken text to have an obvious beginning, middle and end, so for them the lack of generalizability would probably not be problematic. However, taking the point of view that different types of text all constitute examples of language in use, the possibility of having a set of widely-applicable categories is appealing.

The categories were originally assigned on the basis of the experimenter’s interpretation of the text; a post-hoc validation experiment reported in chapter 7 explores the extent to which the categories originally assigned match up with readers’ intuitions about the topic structure of the text used in this experiment. This is particularly relevant in light of the fact that the categories used for this experiment do not reflect categories that have typically been used to describe narrative structure (cf. Labov & Waletzky 1967; Chatman 1978; Chafe 1994), but instead have been adopted from investigations on other types of text.

The categories used in this analysis are defined as follows:

**Topic**  the beginning of a new topic

**Addition**  new information on the same topic

**Elaboration**  more detail or clarification of a previous utterance

**Continuation**  completing an idea or purpose begun in the previous utterance

Each utterance was controlled for length insofar as was possible, with a preference for sentences that were likely to be produced with three pitch accents. An utterance was not necessarily a complete orthographic sentence, and Continuations in particular normally ap-
Table 3.1: Most common target words for each segmental condition

<table>
<thead>
<tr>
<th>/e/</th>
<th>/æ/</th>
<th>/I/</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emory (character name) (86.7%)</td>
<td>Rinnering (character name) (50%)</td>
<td>Amazon (40%)</td>
</tr>
<tr>
<td>memory (6.7%)</td>
<td>minimum/minimal (30%)</td>
<td>animal(s) (30%)</td>
</tr>
<tr>
<td>enemy (6.7%)</td>
<td>mineral (13.3%)</td>
<td>annual (23%)</td>
</tr>
</tbody>
</table>

peared orthographically as the second half of a sentence. A “topic group” consisted of 5 utterances, although measurements were only taken for target items appearing in the first, second, fourth or fifth sentence of the topic group (cf. table 3.2). This follows the results of previous studies (e.g. Sluijter & Terken 1993), in which it was found that while initial and final positions differ significantly from one another and from medial positions, the medial positions do not tend to show differences between themselves, at least not in ways that are perceptible to listeners.

The target items were 3-syllable words with the lexical stress falling on the first syllable, or in a few cases, combinations of words with the same stress pattern. All consisted of voiced continuant segments; most words consisted completely of sonorant segments. In order to account for possible effects of segment identity, three sets of target items were created, each with a different stressed vowel: /e/, /æ/, or /I/. Within a set, the words were as similar as possible and were repeated as much as was practicable while keeping the text coherent (see table 3.1); the full text is available in Appendix I.

Within an utterance, the target item could fall in three positions: Initial, as the first word; Anacrusis, at the beginning of the utterance but following two unstressed syllables; and Final, as the last word in the utterance. An item from each set of targets appeared in every sentence position, in every topic condition, and in every paragraph position, with the exception that Topics could only fall in the first position in the paragraph, and no other topic condition could fall in the first position. This meant that there were a total of 90 possible combinations of conditions, 30 position and topic structure conditions for each stressed vowel.
Table 3.2: Experimental design: condition combinations, with topic structure conditions in columns, and sentence position conditions in rows

<table>
<thead>
<tr>
<th>Position 1</th>
<th>Topic</th>
<th>Addition</th>
<th>Elaboration</th>
<th>Continuation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Initial</td>
<td>xx</td>
<td>xx</td>
<td>xx</td>
</tr>
<tr>
<td></td>
<td>Anacrusis Final</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Position 2</td>
<td>xx</td>
<td>Initial</td>
<td>Initial</td>
<td>Initial</td>
</tr>
<tr>
<td></td>
<td>Anacrusis Final</td>
<td>Anacrusis Final</td>
<td>Anacrusis Final</td>
<td></td>
</tr>
<tr>
<td>Position 4</td>
<td>xx</td>
<td>Initial</td>
<td>Initial</td>
<td>Initial</td>
</tr>
<tr>
<td></td>
<td>Anacrusis Final</td>
<td>Anacrusis Final</td>
<td>Anacrusis Final</td>
<td></td>
</tr>
<tr>
<td>Position 5</td>
<td>xx</td>
<td>Initial</td>
<td>Initial</td>
<td>Initial</td>
</tr>
<tr>
<td></td>
<td>Anacrusis Final</td>
<td>Anacrusis Final</td>
<td>Anacrusis Final</td>
<td></td>
</tr>
</tbody>
</table>

3.2.3 Hypotheses

For all conditions, it was hypothesized that the form of the intonation contour would be influenced by the topic structure condition of the utterance, in a descending order of magnitude from Topic > Addition > Elaboration > Continuation (hypotheses 1-2). The first two hypotheses are based on the results of Wichmann (2000), who found that F0 peak height varied consistently with her four topic structure categories, and that new-topic F0 peaks were timed later than other F0 peaks. In addition to these two hypotheses aimed at replicating Wichmann’s results, two further predictions address the behavior of F0 in relation to topic structure (hypotheses 3-4).

1. Timing of F0 peaks was expected to be latest relative to the segmental structure in Topics, and earliest in Continuations.

2. Height of F0 peaks relative to the speaker baseline was expected to be highest in Topics, and lowest in Continuations.

3. Span of F0 falls (distance between H* and following L) was expected to be greatest in Topics, and least in Continuations.

4. Slope of F0 falls was expected to be greatest in Topics, and least in Continuations.
3.2.4 Recordings

Subjects were 18 native speakers of Standard Southern British English who were paid £4 for their time. Each subject read the text aloud in a sound-attenuated booth at the Phonetics Laboratory, University of Cambridge. The text was just under seven pages long, and subjects had the opportunity to pause for a break at the end of every page, at which point the recording was stopped, so seven separate sound files were recorded for each speaker. Subjects were encouraged to read in any way that felt natural to them, and to correct speech errors, if they so desired, by returning to the beginning of the sentence. All subjects were given time to read through the entire text before beginning the recording process.

The recordings were made using a Sennheiser MKH 40-P48 microphone and a Marantz PMD670 solid state recorder, with files written to an SD card. The sound files were in 16-bit PCM WAV format with a sampling rate of 44.1 kHz.

3.2.5 Analysis

The recordings were analyzed using Praat (Boersma & Weenink 2008-2010). All instances of the token lexical items were labelled and annotated for syllables and segments, as well as for the intonation contour. Segmentation followed the conventions described by Peter-son and Lehiste (1960), and depended on consonant characteristics such as location of the consonant burst, onset or offset of frication noise, or sudden changes in harmonics or formants. If it was impossible to determine one boundary point between segments on the basis of these criteria, then the midpoint between two possible boundary cues was used. F0 peaks and valleys were labelled as H and L respectively, following the autosegmental tradition. The peaks and valleys were normally marked on the basis of Praat’s calculation of the highest or lowest F0 value in a certain time span\(^1\). However, in a few cases, the peak or valley was placed manually in order to account for the presence of a plateau. This was particularly important in the peak timing condition as F0 plateaus in falling contours have been shown

\(^1\)An alternative labelling process was carried out using an automatic two-line regression method to calculate the locations of L turning points (cf. D’Imperio 2000), which has been proposed to be a more accurate way of identifying these points. However, the calculated points turned out to be not very consistent with regard to the measured F0 (although they may reflect other characteristics of the contour shape, cf. Cangemi 2010). In particular, they did not consistently reflect low pitch measurements which were relevant for the calculation of the range characteristics. For this reason, the hand-labelled rather than the automatically-calculated L turning points were used in the analysis.
to pattern with F0 peaks in terms of timing (Knight 2003). The end of the plateau was identified visually as an elbow where the more or less level plateau began to fall steeply. In addition to the peaks and valleys, the first file for each speaker had final low boundary tones (L%) marked in order to obtain a mean value for the speaker’s baseline. The pitch values for these boundary tones were all taken from the lowest point in a stretch of modal phonation; stretches of creaky phonation, which would be expected to produce a lower pitch value, were discarded from the baseline measure.

Although the initial labeling scheme was not intended to be theoretical, in order to make sure the comparisons between items were accurate, it was necessary to apply a theoretical phonological framework to group like contours with like. Due to the large amount of variation in the contour shapes, particularly in prenuclear contours, the analysis for British English proposed by Gussenhoven (1983, 2004; cf. also Grabe 2001 on the labelling conventions for the IViE corpus) was selected as an appropriate scheme. This analysis was preferred to the ToBI analysis for several reasons. First, whereas ToBI has been developed primarily on the basis of American English, Gussenhoven’s analysis is specific to British English and is directly linked to the British tradition of intonational analysis, applying autosegmental-style labels to the British tradition’s intonational categories. Second, Gussenhoven’s analysis allows for a variety of feature specifications modifying pitch accents, while the ToBI analysis prefers an analysis on the basis of different tonal structure; as will become clear in chapter 4, the importance of this difference is particularly relevant for the analysis of the peak timing data. Gussenhoven’s analysis, like Pierrehumbert’s (1980) initial proposal, suggests that all contours that are possible in nuclear position are also possible in prenuclear position, allowing a large amount of variation in the form of prenuclear accents to be accounted for in the intonational phonology. Current interpretations of ToBI do not explicitly deny this, but there appears to be a general neglect of discussion of contour variation in prenuclear position in the ToBI literature.

Only target items in prenuclear position with a falling pitch contour (that is, H*L in Gussenhoven’s analysis; see Gussenhoven 2004: 296ff for a description of the phonetic characteristics of a fall) were used in the initial analysis. By using only items in prenuclear position, items bearing narrow or contrastive focus were also automatically excluded from the analysis, removing another possible source of F0 variation (cf. Ladd 1996 inter alia).
Although F0 peaks almost always had measurable pitch values, some following low valleys had undefinable pitch values (e.g. due to creaky phonation), so the subset of data for the F0 fall range and slope analyses is further limited to those items for which measurement was possible at both of these points.

Many previous studies have used listening paradigms in order to independently identify boundaries between topic units, or topic changes (e.g. Swerts & Geluykens 1993, Swerts 1997). In this experiment, because the topic structure of the written text was specifically controlled for, I did not obtain listener judgments for the relative strength of boundaries; however, see chapter 7 for a validation experiment discussing the interpretation of the topic structure of the text by readers. The results of the phonetic analyses as well as the validation experiment will provide insight into how topic structure may best be described (cf. section 9.2.2).

### 3.2.5.1 F0 peak height

The speaker’s baseline was calculated as the mean of approximately 20 values per speaker taken from L% tones in the first recording; these values were taken from IP-final locations, and the exact number taken per speaker varied slightly due to different choices in phrasing as well as the necessity of discarding tokens which did not have a reliable pitch value, due to phrase-final glottalization (cf. section 2.1.1.2). The peak height was measured as the difference between the value at the F0 peak and the speaker’s baseline. All pitch values were normalized to semitones compared to 100 Hz in order to be able to compare across different speakers. A logarithmic or semitone scale has been shown to be a highly accurate model of perceptual equivalence in intonation, particularly for measures of pitch span (Nolan 2003). Therefore, by taking the measures in semitones, it should be possible to capture the perceptual similarity of intonational variation produced by a variety of speakers, despite their different pitch ranges.

### 3.2.5.2 F0 fall span, length, and slope

The fall span was calculated as the difference between the pitch at the F0 peak and the turning point of the first low valley following it, seen as the end of the fall. Again, pitch values were normalized to semitones. The fall length was the length of time between the
H* and the following L turning point, and the slope was the ratio of the span to the length.

### 3.2.5.3 F0 peak alignment

The F0 peak in falling contours was measured as a ratio of the time from the onset of the stressed vowel to the peak compared to the length of the stressed vowel, in order to normalize for effects of vowel length to compare vowels and also to be able to compare between speakers. Therefore an F0 timing value between 0 and 1 would indicate that the F0 peak fell during the stressed vowel, while a value greater than 1 would mean that the F0 peak occurred after the offset of the vowel. There were some cases in which the F0 peak fell before the onset of the stressed vowel (i.e., during the anacrusis preceding the target item), but these items were ultimately discarded from the analysis due to difficulty in interpretation. The correct phonological analysis for these items may be H+!H* (an early peak, cf. Gussenhoven 2004); if this is the case, then as downstepped tones they are not comparable with the other H* tones included in this study.

### 3.2.6 Results

For the measurements indicated above, items in Final sentence position were not found to vary systematically in any of the statistical analyses carried out. Therefore, I present the results for Initial and Anacrusis positions only. Unless otherwise stated, these two conditions are combined into a single analysis of the F0 characteristics. This is because the patterns observed in these two conditions were consistent and can presumably be accounted for by the same mechanism. It is important to note that in terms of the phonological structure, items in the Initial and Anacrusis conditions can be considered comparable in that they both comprise the first (prenuclear) pitch accent in their IP.

#### 3.2.6.1 F0 peak height

The hypothesis that new Topics would have the highest F0 peaks was confirmed (figure 3.1), but apparently only because Topics always fell in first position in the group of utterances (that is, it was not possible to cross the Topic condition with any of the utterance-group position conditions). F0 peak height from the speaker baseline correlated significantly only with
3.2.6.2 F0 fall range

Although the first two measures proved disappointing in terms of identifying possible F0 correlates to topic structure, the span of F0 falls (figure 3.2) produced on the target words shows a more promising pattern. A simple linear regression shows a significant main effect for topic condition (ANOVA: F(3, 483)=16.87, p<0.05 with the Bonferroni correction applied). The linear regression suggests a clear trend in which T>A=E>C. Although the model shows a small difference between the means for the A and E groups in the direction expected (that is, A>E), this difference does not attain statistical significance in the model.

Given that the difference between A and E was not significant in this model, though the other comparisons were, one possibility is that the theoretical model of topic structure as proposed in the experimental design is not accurate. Therefore, the analysis of the F0 fall range was repeated, this time with the items labelled in a simple hierarchy, following Grosz & Sidner’s (1986) model. The utterances in the text were labeled according to their level of embeddedness, with 1 referring to the highest level (i.e. unembedded) utterances, 2 referring
CHAPTER 3. F0 CUES TO DISCOURSE TOPIC STRUCTURE

Figure 3.2: Size of F0 falls by topic condition

Figure 3.3: F0 fall range by level of embedding
to utterances which were subordinate to one other utterance, and so on. With this new labeling scheme (figure 3.3), although the trend is similar to that of the category analysis with new topics (in the highest hierarchical position) bearing the largest pitch falls, and the most-embedded items bearing the smallest pitch falls, only the difference between the highest level (unembedded) and all other levels (embedded) attained statistical significance (ANOVA: F(3, 429)=9.79, p<0.01).

3.2.6.3 F0 fall slope

Length of the F0 falls did not vary significantly for either the utterance group position or the topic structure category. Therefore, variation in the slope of the F0 falls can be assumed to be a single manipulation rather than the result of a combination of other factors. In this case, Topics were found to have larger (which is to say steeper) slopes than the other topic structure conditions (ANOVA: F(9,255)=3.358, p≈0.05 with the Bonferroni correction applied). However, due to the lack of variation among the other conditions, it is unclear whether this variation is due to the category Topic, or due to the positional factor.
3.2.6.4 F0 peak alignment

The results as they related to the alignment of F0 peaks (figure 7.3) were inconclusive. Although a visual inspection of all the data together suggested that F0 peaks were timed later in new Topics than elsewhere, an ANOVA of the model showed a significant main effect on F0 peak alignment only for the presence or absence of an anacrusis, with an interaction of the anacrusis condition with the topic condition ($F(7, 540)=33.79, p<0.01$ with the Bonferroni correction applied). In Initial sentence position (I), the mean peak timing for new Topics was slightly later than for the other conditions, as predicted; however, there was no significant difference between any of the conditions. In the Anacrusis condition (A), the results are even more surprising, given the hypothesis: here the Topic peaks are timed earliest of any of the topic conditions, contrary to the hypothesis.

3.3 Discussion

The results presented here give a rather mixed characterization of the role of intonation, and specifically F0, as a cue to topic structure.

3.3.1 F0 peak height

The height of F0 peaks above the speaker’s baseline showed a clear pattern related to the position of the utterance in the group: utterances at the beginnings of groups had the highest F0 peaks, utterances at the ends of groups had the lowest F0 peaks, and utterances in the middle had peaks falling somewhere in the middle (and were not differentiable from one another). This is consistent with the findings of Sluijter & Terken (1993), who found that beginnings and ends were highly distinguishable, but that middle levels were more difficult to separate out from one another. Furthermore, these results match what we would predict on the basis of the Production Code (Gussenhoven 1999, 2004): i.e. that beginnings are higher and ends are lower, on the basis of (historical) physiological factors. It would seem, therefore, that F0 peak height above the speaker’s baseline pitch is a good candidate for a universal communicative device, rather than being specified by individual languages.

It is important to keep in mind when interpreting these data that readers knew at all times where the beginnings and ends of utterance groups were, since the written text they
read from was divided into paragraphs. It is possible that this accounts for the contrast between my data and that of Wichmann (2000): depending on the format in which the newsreaders in Wichmann’s study were reading, it may or may not have been clear to them where final boundaries at least were going to fall. It may be the case that peak height is able to vary on the basis of topic structure (perhaps in more skilled readers) but that the nature of my task forced a positional usage instead.

The nature of the text used in the current experiment was intended to force a division between factors that were positional and factors that were related to the topic structure, to determine whether or not a difference actually exists between the two. By comparing the F0 peak height data to the F0 fall span data (discussed below), it would seem to be the case that this distinction does in fact exist. However, the fact that it is apparently possible to separate topic organization from sequential organization, even in a language like English whose organizational conventions are primarily linear, does not mean that this occurs in all circumstances. Different text genres may exist in different languages, and even when the same genre (e.g. a narrative, as in this study) exists, it may be established stylistically in different ways across languages (Chafe 1980; Martin 2001). Although different languages appear to share common stylistic categories (cf. Propp 1968), there is still a great deal of room for flexibility.

It is possible that the peak height above the speaker’s baseline is not a reflection of a tendency to declination, but rather a local cue to group beginnings (and ends, in the case of final lowering). This may be a more plausible explanation for spontaneous speech, in which the speaker may not have fully planned the length of a unit and therefore is not able to apply an “equal” amount of declination to each element in that discourse unit to signal for example the amount of distance remaining before the end of the unit. This would explain why mid-group peaks have similar heights to one another; they could be seen as an unmarked pitch level denoting continuation of a group, while extra high or extra low peaks (in terms of the speaker’s own range) would be marked to indicate beginnings and endings, as relevant. The fact that this phenomenon of mid-level peaks is observable even in the case of a read text in which speakers were fully aware of the location of the end of the group lends support to this hypothesis. In addition, the identification of the extra high or extra low peaks as local events rather than an overall declination effect might allow for more flexibility in other cues.
3.3.2 F0 fall span and slope

The F0 fall span showed a clear differentiation between Topics, Additions/Elaborations, and Continuations, in essentially the order expected (i.e., $T>A=E>C$). This could support the hypothesis that there are three categories of topic structure which are signaled by variations in the fall span. However, it could also support a simple hierarchy theory: Topics were always unembedded, and Continuations were always embedded by at least one level. A piece of evidence against the simple hierarchy theory, however, is that an analysis based on simple embedding does not give a pattern that would match the predictions of the hierarchy theory. Although the observed pattern follows the trend that would be expected, the only statistically significant difference between the groups is between the highest (that is, unembedded) level and all the others; in other words, greater levels of embedding do not appear to differ significantly from one another. In other words, using the category analysis, we are able to distinguish three groups, while using a simple hierarchical analysis, we can only distinguish two. In addition, the statistical model for the category analysis is somewhat more powerful than that of the hierarchical analysis; for the category model, the adjusted $R^2=0.08899$, while for the hierarchy model, adjusted $R^2=0.0584$.²

The F0 fall slope data only show a significant difference between Topics and the other categories, which results in the obvious question of whether the pattern is based on the introduction of a new topic, or whether it is simply positional, a question that is impossible to answer on the basis of these data. However, we must note that in the case of the data which followed a strictly positional pattern, the peak height above speaker baseline, there was both an initial and a final phenomenon. This is not the case here; only Topics/first-position utterances show different behavior to the other groups. Furthermore, the predictions made by the Biological Codes, and particularly the Effort Code, are against a strictly positional interpretation: the Effort Code predicts that more articulatory effort (required to produce a larger slope) correlates with, for example, informational meanings such as urgency; this could be extended in application to providing more salience to new information. However, it seems unlikely that signalling a new unit would be somehow more urgent if it was not for the sake of the content.

²While these values for adjusted $R^2$ are relatively low, the explanatory power of both models is greatly increased if individual differences are taken into account.
3.3.3 F0 peak timing

The peak timing results did not conform to the hypotheses or to previous findings in this area. In fact, the only finding they confirmed was that of Nolan and Farrar (1999), who showed that intonational peaks are timed earlier if there is an anacrusis before the target syllable with which the peak is associated. This at least is valuable as it shows that the observed variations in peak timing were not entirely random. Given this observation, the question of how to analyze the F0 peak timing data is addressed in more detail in chapter 4.

3.3.4 Individual differences

There is a large amount of variation between individual subjects (cf. Hirschberg 2006, section 3.2.1 above). While some subjects very clearly demonstrate the patterns in F0 fall range presented above, others show only parts of them at best. (In contrast, all speakers show the same pattern based on group position for the F0 peak height, lending strength to the hypothesis that this is a universal feature of discourse organization in speech production.) Further analysis presented in chapter 5 provides evidence that speakers who did not make such clear use of the F0 fall range to give cues to discourse organization showed variation in other phonetic cues, such as speech rate, which corresponded similarly with the categories and/or hierarchical analysis as above. Previous studies which identified topic structure categories (Nakajima & Allen 1993; Wichmann 2000) used more than one prosodic feature in dividing these categories; for example, Nakajima and Allen (1993) identified the prosodic characteristics of onset pitch, final pitch, and the ratio between successive F0 peaks as features which were combined by speakers in order to identify their four boundary categories. None of these three characteristics individually appears to have been sufficient to separate the four types of boundaries, but when used together they combined to produce four distinctive prosodic configurations. Given both the wide variety of prosodic phenomena available, as well as the varying behavior of individuals in terms of production, it seems reasonable to assert that multiple prosodic cues are in regular use in speech production, and that the absence of one cue in one speaker’s productions may be accounted for by the use of a different one.

The flexible nature of these cues, in the sense that different combinations of cues rather
than different individual cues may be the key to understanding topic organization, is considered further in the next chapter.

3.3.5 General discussion

As we have seen, fundamental frequency signals to topic structure organization are a mixed bag. While the span and possible slope of F0 falls is a useful cue for many speakers in signalling the topic structure, the height of F0 peaks seems to be more related to positional factors. Furthermore, there may be an additional layer of complexity with topic structure organization in that F0 peak timing appears to vary consistently, not directly with the topic structure, but with a discourse factor “significance” which may be related to topicality, although indirectly.

I have not yet addressed whether the F0 fall range and slope phenomena presented here are phonologically specified or communicative phonetic universals. The short answer to this question is of course that more cross-linguistic research is necessary. However, I would like to make two arguments in favor of these factors being specified by the phonology of individual languages.

First, although these two cues could be physically dependent upon one another, in this case they are not. A pitch fall with greater span would also have a greater slope if the length of the falls were fixed (either in relation to the topic structure or in general), for example. However, this is not what we observe: while the fall span varies for three of the four proposed categories, the slope only seems to vary consistently for topic versus non-topic. In other words, speakers are able to control these two factors independently of one another, and could potentially use them to signal different things. Furthermore, we do see this independent variation in the use of other cues (see chapter 5). Some speakers make relatively little use of variations in the size of the F0 fall span, while other speakers appear to be using it as a primary signal.

The possibility of independent variation does not require that it be used; but it is a prerequisite for allowing these cues to be separately specified in different languages. One reason for speakers to use different cues or cue combinations may be that the language environment to which they are primarily exposed privileges one of these cues above another; this could be expected to vary between geographical regions, meaning that speakers of the
“same” dialect (e.g. SSBE) would be accustomed to different cues. If this kind of variation is observable even within language communities, it seems logical that it should also exist between languages. Of course, the amount of flexibility in these cues has thus far only been observed in the context of production. While different speakers may produce their speech differently, all of these cues may not have equal perceptual status among listeners.

A second argument in favor of the hypothesis of language-specific identification of these cues or cue combinations is that different languages use prosody differently in other areas, and they structure discourse differently. While in English initial position in an utterance group and topic nearly always co-occur, this need not be the case in other languages (see section 3.3.1). If the canonical form of a topic unit is different from language to language, then it stands to reason that the forms for signalling topic structure should also vary. While the topic structure signals in English seem to be tied to initiality, this would not have to be the case in a language that canonically presented topics in medial or final positions. It therefore seems likely that such a language would develop its own set of cues that had nothing to do with initiality; and indeed, the cues might not be intonational in nature at all. This is particularly relevant in the case of tone languages, which are already making extensive use of pitch features and might need recourse to an entirely different set of prosodic (or other) cues to signal topic structure if the communicative aspects of intonation (particularly F0) are already “saturated.”

Although it seems unlikely that we will discover varying pitch accent categories associated with topic structure, it is entirely possible that we will find variations in F0 use across languages to signal similar functions. Discourse intonation, it would appear, falls at the borderline of the universal communicative device and the specifically-specified language feature. The overall goal of greater prominence may be achieved by a variety of different means, and different languages may privilege certain means over others; not simply out of different patterns of language change, but also out of other uses for prosodic cues that have developed. That is, even if a language does not use a particular prosodic cue in its phonology, listeners might still use it as a perceptual cue; one example of this is the role of vowel length in identifying stop voicing in English. Although vowel length on its own is not distinctive in English, it is used as a cue to a characteristic of a nearby sound. A system like this might potentially avoid using vowel length as a meaningful cue, even though it
is not necessarily already a part of the phonology; instead, vowel length could be left as a perceptual cue without adding the potentially confusing need for it as a cue to, say, prosodic prominence.

The prosodic features observed here, although acting on the intonation in ways that are to an extent constrained by the phonology, appear to be phonetic correlates of the discourse structure rather than elements of the phonology itself. Therefore, in a less controlled speech environment we would also expect a large amount of interaction with other discourse markers, particularly lexical ones. In many languages, (cf. Gussenhoven 2007), transitions that have been here shown to be marked by prosody may be marked by syntax or morphology, instead of or in addition to prosodic cues. Focus, for example, can be indicated by word order; however, in languages such as English, word order may be manipulated in order to put a focused element in a location where it can easily receive highly salient prosodic markings, e.g. at the end of an intonational phrase, where it can carry the nuclear pitch accent. It is not yet clear to what extent prosodic and lexicosyntactic cues for topic structure may interact with one another. It is possible that strong explicit marking of the topic structure, using words such as “and” or “so” to indicate relationships between the propositional content of individual utterances, may do away with the need for prosodic topic marking entirely. It is after all possible to indicate topic structure in written texts, where the prosody is entirely absent (with the exception of punctuation, which is a much sparser system than the prosodic system available in speech production). On the other hand, written texts tend to involve more complex syntactic and other structures than (particularly casual) spoken texts. The question may not therefore be a matter of one type of cue doing away with another, but simply with different methods for expressing the same thing, which have become commonly used in different contexts. What we may be seeing is a process of grammaticalization happening before our eyes, wherein cues that originate in universal communicative means are becoming embedded in the language system by the consistent selection of particular cues as more probable indicators of topic structure variation.
3.4 Conclusion

The results of the experiment presented here indicate that in production of SSBE, it is possible to identify a set of cues in terms of F0 modulation that speakers are consistently using to signal discourse functions. In particular, modifications of the design of falling contours have been shown to correlate significantly with the topic structure of the discourse. It remains to be seen to what extent listeners are able to make use of these cues when decoding the message, since they are being simultaneously bombarded by indexical and grammatical intonation. However, the ability to identify these different categories will be of assistance in understanding how we make use of F0 and other prosodic information in perception.
Chapter 4

F0 Peak Timing and Topic

The analysis in this chapter was previously reported in Zellers, Post & D’Imperio (2009).

4.1 Introduction

In chapter 3 a set of data on F0 peak alignment was reported that did not show the same pattern as previously reported research, notably that of Wichmann (2000). Specifically, while Wichmann (2000) found that F0 peaks in new topics showed later alignment than peaks in non-new-topic utterances, no such pattern was observed in the current data set.

There are at least three possible explanations for the peak timing data presented in chapter 3. It is possible that the theory of topic structure proposed (see 3.2.2) was entirely incorrect. Although the validation experiment reported in chapter 7 addresses this question more directly, there are other reasons to defend against this claim. To begin with, despite debate about the internal structure of topics, it seems relatively uncontroversial that topic shifts occur and are phonetically marked. Therefore, new Topics should still have stood out from the other three groups, even if a less clear pattern of behavior might be expected in the Additions, Elaborations, and Continuations.1 Furthermore, if this was the case, it seems likely that the overall pattern would be consistent even taking into account different phonetic conditions such as the presence or absence of an anacrusis. Instead, the different phonetic conditions result in widely different prosodic patterns. Given that in other mea-

1Readers’ ratings of which locations constitute topic boundaries, and the differences between their ratings and the original labelling scheme reported in this study, are discussed in chapter 7; however, that chapter also demonstrates that despite some disagreement of the readers’ labels with the labels used, the overall pattern of the data does not appear to be affected.
surements the data behaved in a way consistent with the categories given, it is more difficult to argue on the basis of the peak alignment data that they are incorrect. It is worth considering the possibility that a strict hierarchical characterization would be more accurate than the category analysis; however, re-testing with the hierarchical labels rather than the topic structure category labels did not prove any more fruitful, just as in the case of the F0 fall ranges (cf. section 3.3.2).

A second possible way of explaining the F0 peak alignment data reported in the previous chapter would be to argue that the F0 peaks were not anchored to the onset of the stressed vowel, but to some other part of the word (for example, the syllable or the mora). While this is a possibility, other studies have found results measuring from the vowel onset, so it is surprising that the same should not have been found here, although a reanalysis along these lines could be a potential route of investigation.

A likely alternative to these two possibilities is that there were too many other confounding factors for any pattern in peak timing to be observable. That is, phonetic or sentence-level phonological prosodic organization may obscure any discourse-related pattern even if it is present. This also raises the possibility that Wichmann’s (2000) results regarding peak timing were confounded by factors other than topic structure, particularly since her study did not specifically investigate non-discourse-related influences on prosody. As we have seen in section 2.2.4, F0 peak alignment in many languages can be a cue to the identity of pitch accents, and is sensitive to factors such as focus, which occur at the level of the sentence. There are also structural factors which can influence peak timing, both on the segmental level (as in Nolan and Farrar 1999, see section 3.3.3), and on the level of intonational structure (cf. e.g. Arvaniti & Ladd 1995, who show that the presence of additional unaccented syllables in an IP can influence the timing of pitch accents’ F0 peaks).

I excluded tokens which carried focus from my analysis, and the text was designed in such a way as to try to keep consistent the lengths of the IPs produced by the speakers; however, these are not perfect controls, since speakers were free to read in any way they chose. It is possible that an underlying pattern like that proposed by Wichmann (2000) is present in this data if other confounds are removed. However, it is equally possible that Wichmann’s data demonstrated the patterns it did not on the basis of topic structure, but due to some other as-yet unidentified cause. If this is the case, then it is necessary to seek
Figure 4.1: Schematics of two (prenuclear) pitch-accent models. (a) Two-accent model, in which early and late peaks form two categories. (b) One-accent model, in which all peaks are considered one category.

4.2 Phonological modelling

4.2.1 Motivation

One factor that Wichmann’s (2000) analysis does not take into account is the phonological identity of the pitch movements she measures. She takes the first pitch peak in each utterance and compares its timing in relation to the stressed vowel as a direct correlate of topic structure. This neglects the possibility that the F0 peak alignment is affected by other, sentence-level phonological factors. However, as was discussed in section 2.2.4, F0 peak alignment is widely considered to be an important cue to phonological pitch accent identification. Since the study in chapter 3 included only phonologically falling pitch accents, this provides a possible means of accounting for the discrepancy between Wichmann’s results and the current study.

4.2.2 Reanalysis

The new analysis compared F0 peak timing under two phonological models of the prenuclear pitch accents involved in the study. Two labeling schemes were applied to the data
to represent the two models under examination. All relevant sentence-initial high turning points (or plateau ends, which pattern with high peaks in terms of timing, cf. Knight 2003) were labeled as H; the definition of all prenuclear accents as rises is consistent with the current ToBI literature (see below). Then these items were subdivided into groups of H*L or L*H(L) on the basis of the analysis given by Gussenhoven (2004) for SSBE, which renders the British tradition’s categories in autosegmental terms. Figure 4.1 gives a schematic of the distributions of pitch movements in time under the two different analyses.

The distinction between these two models is based on a set of somewhat problematic criteria. First, although the original model proposed by Pierrehumbert (1980) includes both (L)H* and L*H in prenuclear position, prenuclear L*H has been largely ignored in ToBI research. The reasons for this neglect are unclear, since to my knowledge no proposal to discard prenuclear L*H has been made, but it may stem from a difficulty in defining separate criteria for identifying LH* and L*H in this model. In the case of Gussenhoven’s model, the distinction between H*L and L*H is made on the basis of both the auditory impression of the accent in context, and the visual representation of the F0 trace. In terms of visual cues, rising pitch accents typically start at a low pitch and rise throughout most of the course of the target word to peak near the end of the word (or even after, though such instances were not included in this study), while falling pitch accents start at a mid-to-high pitch and often rise slightly to a peak before the middle of the target word, then

Figure 4.2: Schematic of (a) rising pitch accent and (b) falling pitch accent in relation to 3 syllables of target word with initial stress.
fall through the remainder (cf. Gussenhoven 2004). Figure 4.2 gives a schematic of the characteristic visual pitch traces, and figure 4.3 gives two examples of typical items in each category. However, for this analysis, the auditory impression was primary; if a small rise was visible in the F0 trace but the auditory impression was of a falling pitch, then the label H*L would normally be applied. Some cases were more problematic, such as that in figure 4.4. This example demonstrates several features which problematic cases exhibited. The pitch peak is located in the middle syllable, with the pitch movements on both sides fairly symmetrical. The auditory impression was ambiguous, the more so because another upstepped accent followed immediately after the accent in question. There were not very many ambiguous cases, so for purposes of the analysis reported in the rest of this chapter they were discarded; however, an automatic comparison of the pitch contours using Functional Principal Component Analysis (Zellers et al. 2010, discussed below) addresses one potential way of dealing with such ambiguous contours.

A small subset of the two-accent analysis labels were checked and confirmed by another researcher experienced with this labelling scheme. The distribution of the accents into the two categories also received post-hoc confirmation from the analysis conducted by Zellers et al. (2010) using Functional Data Analysis (FDA), and specifically Functional Principal Component Analysis. FDA adapts linear data (in this case, F0 contours) into functions, and then uses traditional statistical methods to compare these functions to one another instead of relying on individual data points. This means that whole F0 contours may be compared to each other to identify their overall similarity, rather than depending on data points selected on the basis of intonational theory, which may be prone to human bias. The Functional Principal Component Analysis (FPCA) compared 126 tokens of the target word “Emory”; the set of tokens included ambiguous cases that had not been used in the pitch-accent distribution analysis. In this case, the FPCA results in a first principal component (PC1) which accounts for 83.5% of the variance in the data, and may be roughly interpreted as whether the contour has an early or a late peak (figure 4.5). A histogram of the PC1 values for this data set shows two peaks, suggesting a bimodal distribution. It is important to remember that the PC scores and distributions are completely blind to the labels assigned by hand. If the labels are reintroduced to the PC data at this stage, strong evidence for the cause of the bimodal distribution of the PC1 data becomes apparent (figure 4.6) The means of the PC1
Figure 4.3: Typical instances of a prenuclear rise (top) and a fall (bottom).
Figure 4.4: A problematic case which was ambiguous between a rise and a fall (this item was ultimately discarded from the pitch accent analysis).

Figure 4.5: PC1 for the Emory data, showing typical shapes for contours with a positive PC1 value (+) or a negative PC1 value (−). The mean of all 126 contours is shown with a solid line.
Figure 4.6: Distribution of human labels along PC1 axis. Red = labelled as falls by human labeller. Blue = labelled as rises by human labeller.

scores for the items labelled as rises versus falls differ significantly ($t = -14.96$, $df=117.323$, $p<0.01$), despite the fact that the automatic analysis was blind to these labels, and furthermore addressed only the speech signal itself without reference to perceptual factors. FDA thus provides additional support for the two-label scheme as applied in this investigation.

### 4.2.3 Results and Discussion

#### 4.2.3.1 Single pitch accent analysis

The first analysis disregarded the evidence from the FDA analysis and assumed that all tokens included in the dataset belonged to the same category, in line with ToBI’s prediction. Investigating F0 peak alignment in this case, a clear pattern emerges (cf. figure 4.7). New Topics are timed latest relative to the stressed vowel, with the other groups following in order: $T>A=E>C$ (ANOVA: $F(3, 285)=14.7$; $p<0.01$). This result is comparable to Wichmann’s (2000) findings, with the exception that it only clearly distinguishes three groups;
Additions and Elaborations have slightly different means but the difference is not statistically significant in the current data.

A prediction of the one-pitch-accent analysis is that a large amount of variation in F0 peak timing will be accounted for by the segmental composition of an utterance. The design of the text controls for the identity of the stressed vowel, the specific target word being tested, and the existence of an anacrusis among other segmental factors. Therefore it is possible to strip away some of the extraneous variation to look more closely at peak timing variation with relation to topic structure. The first segmental influence to be accounted for is the presence or absence of an anacrusis, since Nolan & Farrar (1999) have shown that F0 peaks are timed earlier following an anacrusis than if the stressed syllable with which they associate is the first syllable in an utterance. In accordance with the findings reported in section 7.4.2, this pattern holds in this data, as seen in Figure 4.8. Peaks are aligned later in the utterance-initial (I) position than in the anacrusis (A) condition in all cases. However, by accounting for some timing variation by including the anacrusis factor, the topic-related variation becomes less clear. Although new Topics still show later-timed peaks than other topic conditions in Initial position, this pattern does not hold when there is an Anacrusis. Furthermore, within the Initial condition, there are no statistically significant differences
between A, E and C, nor are there differences in the Anacrusis group (F(3, 118)=0.21, ns).

An additional factor potentially affecting peak alignment is the segmental construction of the syllable in question, which can be controlled for by comparing different occurrences of the same word (and thus the same target syllable) in the data. A segmentally-conditioned pattern similar to that reported above emerges in this case. Peaks are timed later in the two vowel-initial conditions than in the consonant-initial condition. This is unsurprising given the extra time available for the rise before the onset of the vowel when a consonant is present, just as in the Anacrusis condition. However, although new Topics once again show later peak alignment in the case of one vowel-initial target word, “Emory” (F(3, 160)=13.51, p<0.01), this is not the case for the other two words tested, and the other topic group distinctions do not attain statistical significance (F(11, 277)=9.41, ns). Although the patterning of topic structure with F0 peak timing appears clear when all items are combined, when segmental effects are accounted for, the category distinction originally observed becomes more and more difficult to identify. If segmental effects (such as the presence or absence of a syllable onset or an anacrusis) are included in the analysis, the correlation between topic structure and F0 peak alignment all but disappears. At best, there remains a delaying ef-
fect of new Topic on the peak alignment, which is mitigated if the stressed syllable has a consonant onset or if there is an anacrusis.

4.2.3.2 Two pitch accent analysis

The alternative analysis of the pitch accent structure of the data groups the items into two sets, where one is characterized by a rise from the first accented syllable, and the other by a fall. These two groups are differentiated phonologically in that the rising category has an additional low tone associated with the stressed syllable; the F0 peak represents a trailing high tone following the low starred tone, and is therefore automatically realized later in the word. These groups may also be described tonally as L*H(L) (rising or rising-falling) and H*L (falling). In this analysis, comparing F0 peak alignment within the pitch accent groups, there are no statistically significant differences across topic structure categories.

The distribution of the two pitch accents across the topic categories, however, is striking. In Figure 4.9, we see that the ratio of L*H(L) to H*L decreases moving down the topic structure categories: T>>A>E>C ($\chi^2(3, N=289)=60.25, p<0.05$). In fact, new Topics are more than twice as likely to have an L*H(L) accent than the other categories. It is important to note that the distinction between these pitch accent groups is unaffected by the segmental
factors, both the segmental structure of the target item and the presence or absence of an anacrusis, that we observed before in the strict timing case. In these conditions, the pattern of the ratio L*H(L) to H*L remains nearly the same, with T>A>E>C. The only slight exception to this pattern is the anacrusis case, where the A and E groups appear to be reversed in position (see Figure 4.10). This is also the only analysis in which L*H(L) does not outnumber H*L in the new Topics; however, the ratio of L*H(L) to H*L is still the highest in Topics.

The distribution of L*H(L) and H*L can apparently not be directly tied to topic structure, at least in a categorical sense. If this was the case, we would expect to see (for example)
only L*H(L) in new Topics, and only H*L elsewhere. In the data, though, L*H(L) is simply more likely in new Topics than elsewhere.

Gussenhoven (2004) gives an analysis of L*H(L) which could explain the observed distribution. He proposes a tonal morpheme in English denoting ‘significant information’, realized as the prefixation of an L tone before the pitch accent in question. The new L tone becomes the tone associated with the stressed syllable, altering the pitch accent (by adding an additional meaning element) in the process, as below:

\[(4.1) \quad L + H*L \rightarrow L*HL\]

‘sensitive’ + ‘addition’ → ‘significant addition’

Affixation of the delay morpheme (following Gussenhoven 2004)

It seems likely that information higher in the topic hierarchy, by virtue of being “newer”, could be considered more “significant” than surrounding items, and therefore carry the additional tonal morpheme. Note that this tonal-morpheme analysis is different from other descriptions of peak delay (e.g. Ladd 1996), which consider delay a phonetic characteristic of a given accent category.

### 4.3 Discussion

The two intonational models presented here provide very different bases for the interpretation of F0 peak timing in relation to topic structure. The one-accent-category model, which would permit peak timing to vary in direct relationship with the topic structure, does indeed show delayed peaks in new Topics. However, once segmental influences on peak timing are accounted for, there is no way to distinguish the other proposed topic structure categories on the basis of peak timing. This leaves the categories to be primarily signaled by other phonetic factors (for example, the range of the F0 pitch movements, as in chapter 3), assuming of course that the categories do exist and are relevant for communication.

With regard to the two-accent-category model, it might be easier to propose a multiple-level analysis of the topic structure. However, the two-accent-category data cannot make a definitive statement about the topic structure theory either. This is because the accent categories themselves do not relate directly to topic, but rather to some other factor that
shows a correlation with topic (potentially “significance”, but possibly something else that is as yet unidentified). The distribution of this factor in relation to the topic structure may be categorical or gradient; further research will be required on this question. The additional evidence for the existence of topic structure categories as reported in other chapters of this thesis makes the two-accent-category analysis attractive, because it allows the different topic structure categories to be modeled in intonation. However, it is important to remember that there is no a priori reason to assume that F0 peak timing is actually sensitive to topic structure, especially when we know that it is used to make distinctions on the level of pitch accent identity in nuclear accents. As a consequence, it may be completely irrelevant that there is no consistent variation of F0 peak timing with the topic structure categories in the one-accent-category model; the relevant phonetic cues may be elsewhere.

4.4 Conclusion

By comparing two models of peak timing variation in prenuclear accents, it is possible to evaluate the contribution of a number of factors that may influence the prosody of long units of discourse in addition to the topic structure. In particular, with regard to topic structure, anything from the segmental structure of the utterance up to the identity of the pitch accent in question may be a confounding factor in a peak timing analysis, and the intonational model we choose will affect how we deal with these additional factors. Models of the prosody of topic structure cannot be made in isolation from models of intonational phonology, but must take into account their predictions. At the same time, modeling topic-structure-related prosodic phenomena can help us to compare, evaluate, and improve current models of intonation.
Chapter 5

Other Prosodic Cues to Topic:
Speech Rate

5.1 Introduction

As we saw in the previous chapters, variation in the F0 characteristics of utterances, and particularly in the size of F0 falls in prenuclear accents, correlates with variations in the topic structure of longer discourses. However, the picture of topic structure that it is possible to deduce from the F0 variation is less complex than what has been proposed to be the case in spoken discourses (e.g. by Wichmann 2000). While the introduction of new topics is easily identifiable using only F0 characteristics, the distinctions between other categories (or hierarchical levels) are less clear. Studies of prosodic variation in other contexts make it clear that prosodic meaning is carried by factors other than simply F0 even in cases where the main characteristic under study is intonational variation (e.g. Hirschberg & Grosz 1992, Swerts & Geluykens 1993; cf. section 2.3.3), which means that this kind of variation may be helpful in clarifying some of these distinctions.

Descriptive prosodic studies of topic change in speech have measured a variety of types of prosodic variation, trying to identify anything that can be seen to occur at topic structure boundaries. The features that have been noted include both intonational features such as pitch reset or boundary tone (e.g. Nakajima & Allen 1993), as well as non-intonational features such as pause length or pre-boundary lengthening (Lehiste 1970, Swerts 1997 inter alia). Most of these studies looked at boundary strength without making explicit claims
about topic structure (with the exception that very strong boundaries, ending a group of speech, would normally be considered to coincide with a topic change). Despite this lack of an explicit link, however, it is not unreasonable to consider boundary strength as being relevant to discourse topic, since less closely related utterances could easily be expected to have a stronger boundary between them than more related utterances.

Studies that have specifically looked at the phonetic correlates of topic structure have often relied on variation in multiple prosodic characteristics, especially F0/intonational characteristics, in order to classify utterances into topic structure categories. Nakajima & Allen (1993) used three F0 characteristics: the F0 maximum in an utterance, the F0 minimum, and the peak ratio (i.e. the comparative size of F0 peaks within an utterance). None of these characteristics on its own was sufficient to categorize utterances into the topic structure categories they identified. However, they propose that by combining all three of these measurements, it would be possible to create a more sophisticated algorithm which would be able to automatically classify utterances according to their topic structure. They provide a schematic of how the features could combine, with different “level” settings on different features combining in multiple ways with one another, although they do not actually implement this system. Similarly, Wichmann (2000) suggests that a combination of pitch scaling characteristics, both local and global, vary depending on the topic structure of discourse, showing schematically how pitch range expansion or compression of contours across sentences could provide the variation in F0 peak height that she observed in her data.

Although the idea of multiple cues to boundary strength has been well-investigated, much of this research has concentrated on F0 cues. While prosodic characteristics such as speech rate/final lengthening, pause duration, and glottalization have regularly been associated with “boundary strength”, and are often identified as occurring at boundaries between topic groups, it is yet to be seen whether these features may also be used to distinguish topic structure categories within discourses, as Nakajima & Allen (1993) did with their intonational cues. However, research in a Conversation Analysis framework shows the importance of taking into account a variety of co-occurring cues to different discourse functions. For example, Local and Walker (2005) report that instances of stand-alone “so” which hold topics differ in their loudness, pitch, and voice quality from instances which license topic closing. The purpose of the analyses reported in this chapter is therefore to identify the extent
to which non-F0 prosodic cues might contribute to the internal organization of discourses around topics, particularly in combination or interaction with the F0 variation reported in Chapters 3-4.

5.2 Initial Observations

If multiple prosodic cues to topic structure are available to speakers, we might expect to see individual variation in terms of the use of these different cues, even though there are overall trends in a given population. Although speakers of the same language must by definition be mutually intelligible, this does not translate to a requirement that their speech be identical. Therefore, although the distribution of a prosodic cue, in this case F0 fall range, may appear as in chapter 3 for a whole group of speakers, it is not surprising to find within this distribution the two very different patterns found in Figure 5.1a and b for individual speakers.

In Figure 5.1a, we see F0 fall range data for speaker F05, whose speech production pattern basically matches the overall pattern for this data (cf. figure 3.2). New Topics have the largest F0 falls, and Continuations the smallest, while Additions and Elaborations fall somewhere in the middle and are difficult to distinguish from one another (ANOVA: \( F(3, 41) = 2.602, p=0.05 \)). The lack of statistical significance is almost certainly due to the
relatively few data points available, since the variation produced by this speaker is similar in size to the effects observed for the whole group in chapter 3). In Figure 5.1b the F0 fall range data for speaker F04 is presented. In this case, instead of what we might describe as a stepping-down pattern in Figure 5.1a, the Topic, Addition and Elaboration categories appear to all have more or less the same size fall range. Continuations are the only category to vary noticeably from the mean of the other groups, having a more compressed fall range ($F(3, 31) = 6.988, p<0.01$).

If we looked only at the data for speaker F04 (and for the other four speakers in the data set who pattern with this speaker), we might be led to expect that there are only two topic structure categories, perhaps representing categories involving any kind of new or unpredictable information as opposed to categories with very high predictability. To an extent, this is consistent with Wichmann’s (2000) expectation of the content of topic structure categories, but it is much less specific and less detailed than her predictions. Even in a very simple topic structure theory, the most important distinction is between utterances beginning new topics and those not beginning new topics, yet this speaker appears not to make such a distinction. However, the two speakers are speakers of the same dialect of SSBE and read the same text; on an impressionistic listening, they sound similarly expressive. Even if, as seems likely, there were some small variations in how the individual speakers interpreted the topic structure of the text, these drastic differences in the acoustic patterns are striking, because they appear to represent two completely different interpretations of the text.\footnote{This possibility will be discussed in greater detail in section 7.5.}

These two speakers are generally representative of patterns in the data, and it is possible to divide the other speakers into two groups (i.e. speakers who pattern with either F05 or F04), although not all speakers showed the exact patterns or had them attain statistical significance; again, this is likely to be due to the relatively low number of data points available for some speakers. The data presented in chapter 3 show that across all speakers it was possible to distinguish at least 3 topic structure categories by means of F0 characteristics alone; it seems likely on the basis of the studies reviewed above that by adding more prosodic variables into the mix it should be possible to strengthen these observations, and perhaps to account for topic organization in discourse by speakers who make less use of F0 variation related to topic structure. This chapter will discuss variations in local and
global speech rate, following on from these initial observations. The following chapter will address another prosodic characteristic, phonation quality.

5.3 Speech Rate: Word Length

It has occasionally been suggested that variations in speech rate could signal changes in topic. Speakers certainly vary their average speech rate during discourses by more than the just-noticeable difference of 4.5-5% which has been reported in the perception literature (cf. Eefting & Rietveld 1989; Quéné 2007), making speech rate a potentially communicatively valuable cue.

In a study on spontaneous Dutch speech, Koopmans-van Beinum & van Donzel (1996) found that average syllable duration was longer at initial “paragraph” or topic boundaries, although they attribute this variation in part to the fact that new paragraphs were often introduced with discourse markers which consisted of a single syllable but which were quite long. These long discourse markers might be representative of hesitation phenomena rather than topic change marking as such, although the presence or absence of hesitations at topic boundaries compared to other boundaries might itself be an interesting avenue of investigation. Regardless of the status of Koopmans-van Beinum & van Donzel’s findings, however, there is certainly a strong association with the idea of important or unpredictable information being produced at a slower speech rate (cf. Nooteboom & Cohen 1984). This principle is commonly present in the advice given to people speaking in public to speak more slowly when they have something important to highlight. Furthermore, “incidental” or less important items are often produced at an increased speech rate (Crystal 1969, Quirk et al. 1985 for English; cf. Asuaje et al. 2005 for more detailed experimental results in Venezuelan Spanish). The idea that listeners attribute importance to the content of a message spoken at a slower rate appears to be reinforced by an advertising study by Megehee et al. (2003), which found that when advertisements were presented with a slower speech rate, listeners were more likely to respond to the product advertised rather than to the speaker or to the form of the message. This study tested effects of global speech rate rather than local variation on specific lexical items, however.

In contrast to these findings, Li & Zu (2008) found that in Mandarin, the beginnings
of larger prosodic groups were often characterized by an increased speaking rate compared to the beginnings of smaller prosodic groups (presumably those which would be medial in the large prosodic groups, although this is not made explicit in their study). They do not offer any immediate explanations for why this may be the case, but it is possible that a local increase in speaking rate could increase contrast with the final lengthening that would characterize speech at the end of a preceding unit. The effects measured by Li & Xu (2008) appear to be boundary effects; that is, they do not have scope over the whole prosodic unit in question, but rather signal relative boundary strength locally.

Also contrasting with the aforementioned findings of slow speech rate is the phenomenon of “rush-through” reported in the CA literature, in which a speaker who is completing a turn-constructional unit but who wishes to hold the floor will increase speech rate, avoid pausing, and/or continue an intonational contour into the beginning of a new turn in order to extend his or her own turn (Schegloff 1982; 1987). Rush-through occurs at locations in talk which are possible transitional relevance places; the completion of a topic would be consistent with the potential closing of a turn, which suggests that a speaker who is already speaking and who wishes to begin a new topic might therefore increase speech rate in order to ensure success in introducing a new topic.

On the basis of these studies, it seems likely that it should be possible to observe variation in speech rate on the basis of the topic structure of a spoken discourse. On the basis of the studies reported above, it seems most likely that new Topics would be spoken at the slowest speech rate, with speech rate increasing across the categories Addition<Elaboration<Continuation.

5.3.1 Methodology

The data used for this analysis were the same as those collected for the F0 analyses in Chapter 3. All target items were labelled with syllables and segments, as well as the location of the end of the intonational phrases in which they were contained, and the presence and identity of any pitch accent on the target word. As in the previous F0 analyses, only items which occurred at the beginning of an utterance (with or without an anacrusis) were considered, since the effects of final lengthening would interfere with any speech rate data in utterance-final position. As in the previous portions of the study, all measurements were
taken in Praat (Boersma & Weenink 2008).

Since matched target items in the text were available, the estimate of speech rate in this analysis was simply of word length: for all tokens of the same target item which were judged to be produced with the same phonetic segments, the length of the word was taken and compared to other tokens produced by the same speaker. Since this measure could be sensitive to changes in the speaker’s speech rate over the course of the story, the consistency of rate was tested by examining whether there was any correlation between syllable length and location of the syllable in the recording. The analysis confirmed that no such correlation existed in the data ($R^2 = 0.00611; p = 0.785$, n.s.; see figure 5.2).

### 5.3.2 Results: Word Length/Local Speech Rate

Figure 5.3 shows the length of the matched target words for the same two speakers whose F0 data was presented above (Figure 5.1). For speaker F05 (Figure 5.3a), who used F0 fall range in the same way as the overall trend in the data (cf. section 3.2.6.2), there is very
little meaningful variation in the speech rate, although new Topics are spoken marginally more quickly than the other categories (ANOVA: F(3, 43) = 2.66, p=0.06). Interestingly, for this speaker, Additions and Elaborations may also be differentiable on the basis of speech rate, although this pattern did not hold for all speakers, and indeed did not attain statistical significance here. For speaker F04 (Figure 5.3b), however, who made relatively little use of the F0 fall range, the speech rate data are striking. Target words in new Topics are spoken at a faster rate (i.e. the length of a given target word in that context is shorter) than in Additions, which in turn are spoken more quickly than in Elaborations and Continuations (ANOVA F(3, 37)=2.84, p=0.05). The differences between Topics, Additions and Elaborations/Continuations in the speech rate combined with the significant differences between Topics/Additions/Elaborations and Continuations in the F0 fall range mean that by using these two cues in combination it may be possible to distinguish between all four categories of topic structure which were posited in the study (cf. Table 5.1). New Topics might have a special status in that they also form a new group, accounting for the doubly-unique cue levels. Note that the length of the F0 fall is not necessarily tied to the length of the word, so collapsing these two cues into the single measurement of F0 slope is not possible.

For both speakers, it appears that the utterance’s sequential position in the group does not affect the length of the target word (other than the unavoidable confound of new Topics...
Table 5.1: Combination of F0 fall and local speech rate cues to differentiate four topic structure categories

<table>
<thead>
<tr>
<th>Topic structure category</th>
<th>F0 fall size</th>
<th>Initial word length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Topic</td>
<td>Large (≈5.5 st)</td>
<td>Short (≈0.35 sec)</td>
</tr>
<tr>
<td>Addition</td>
<td>Medium (≈4.5 st)</td>
<td>Medium (≈0.38 sec)</td>
</tr>
<tr>
<td>Elaboration</td>
<td>Medium (≈4.5 st)</td>
<td>Long (≈0.42 sec)</td>
</tr>
<tr>
<td>Continuation</td>
<td>Small (≈3.5 st)</td>
<td>Long (≈0.42 sec)</td>
</tr>
</tbody>
</table>

with first position in the group; see section 3.2.2).

Figures 5.4 and 5.5 plot the relationship between F0 fall range and length of the target word for all speakers in the study. A strong negative correlation between word length and fall range is immediately apparent, and indeed reached statistical significance (linear regression, F(7, 56)=6.818, p<0.05). As word length decreases, F0 fall range increases, which is the opposite of what we might expect from an articulatory point of view, where a longer word length would allow more time for a pitch movement to occur and would therefore license a larger range (although it is important to note that not all of the F0 falls were completed within the target word); this suggests that the two cues are independently variable, even though they appear to be signalling the same thing. This may be evidence of the two cues having been independently grammaticalized into the system from prominence variation originating in the Biological Codes; see section 9.3.2 for further discussion of this.

Figure 5.4 shows the distributions of the mean fall range and mean word length for each topic structure category, aggregated by speaker. The topic structure categories are not labelled in this figure, but it is possible to see how individual speakers vary in their productions of the target items. Some speakers very clearly vary primarily along the horizontal axis (fall range), such as speakers M02 and M04, represented by green and blue asterisks, respectively. Other speakers, such as F04 or F11, represented by orange x and blue x respectively, vary much more along the vertical axis (word length). Most speakers, though, do not appear to privilege variation on one of these factors over variation on the other.

This variation between speakers is interesting to observe, but it tells us little about the usefulness or reliability of the variation of these two factors as signals for the topic structure

---

2Although the examples pointed out here are male for F0 and female for speech rate, there was no overall effect of speaker gender on the prioritization of one cue over the other.
Figure 5.4: Mean fall range and word length for the four topic structure categories, aggregated by speaker. Each different symbol/color combination represents one speaker.
Figure 5.5: F0 fall range and word length, aggregated across speakers, marked with topic structure category of discourses. Figure 5.5 addresses this question by presenting the same plot, this time labelled according to the topic structure categories of the data points. In this figure, we can see a trend towards new Topics having a large fall range and a short word length, with a general decrease in fall range and an increase in word length moving down the topic structure categories, although there is a large amount of overlap, and a particularly large amount of variation in the new Topic items. It is important to recognize, however, that these are means for individual speakers, rather than direct observations from the data. For any given speaker, the distribution would likely be much clearer, since one individual’s data should be more consistent with itself than with data from other speakers in a population.
Even given these speaker-based generalizations, however, a trend is apparent, with points on the extremes of the distribution tending to be either new Topics or else Continuations, with Additions and Elaborations clustering in the middle.

The general trends suggest that the variation may be systematic, even though it does not have a large effect, and it is impossible to draw any conclusions about a potentially categorical effect. When we combine the data for all speakers into two groups, those who appeared on visual inspection to use fall range versus those who used speech rate variations to signal topic structure, some of the differences do attain statistical significance (see Table 5.2). This suggests that even within one dialect group there are still different possibilities for prosodic variation related to topic structure, and that closer inspection of individual differences is necessary in this regard. Comparisons were carried out by ANOVA, applying the Tukey HSD correction in post-hoc comparisons (fall range: $F(3, 492)=16.8$; word length: $F(3, 544)=9.74$). The variations in speech rate which do not attain statistical significance are less than 4.5% different from their neighbors; in other words, they do not meet the criteria for the just-noticeable difference established by Quené (2007), rendering the statistical result unsurprising.

One question raised by these data is the correlation of a faster speaking rate with the introduction of new Topics. As mentioned above, most of the literature would suggest that new Topics, as “more important” information, should be spoken more slowly, if there is any
meaningful variation in the speaking rate. There are a few exceptions to this: Schegloff (1982, 1987) identified “rush-through” in which a speaker who wishes to hold the floor across the end of a turn-constructional unit will increase speech rate; and Li & Zu (2008) found increased speaking rates at the beginnings of utterances, which they attributed to the positional factor of being at the beginning of a prosodic group, rather than to the topical content. Since the target words were initial or extremely early in the utterances studied, it is possible that the increased speech rate observed could be an artifact of having observed only individual words. However, given Li & Zu’s finding, the (fairly uncontroversial) hypothesis that new Topics are preceded by strong boundaries receives further support from the data presented here, in that it was more likely for the target item to be produced with increased speech rate if it came from an utterance with a higher topic-structure “rank” (or perhaps a higher hierarchical position). This in turn adds support for the idea of treating topic structure boundaries as not unlike other boundaries marked by prosody (cf. section 5.1), although the rules governing them are likely to be different.

5.4 Speech Rate: Syllable Length in Sentences

Given that most studies of speech rate have looked at changes of rate over units longer than a single word, it is possible that the rather surprising result of increased speech rate in new Topics could be reversed over the stretch of whole utterances. The following study explores this alternative by investigating syllable duration in all syllables in the utterances, rather than only in target words, in a subset of the data.

5.4.1 Methodology

A subset of 3 speakers’ data (2 female and 1 male) was used for this analysis; all three were members of the “speech rate” group (cf. table 5.2). As such they might be expected to have more control over variations in their speech rate; but on the other hand, the variations we have seen so far were in the opposite direction as what would be predicted for the current analysis. All of these speakers’ utterances were segmented into syllables by hand, as an earlier attempt with an automatic method proved unreliable; the time requirement for hand segmentation was the reason for the decision to use only a small subset of the speakers.
Table 5.3: Fixed effects from linear mixed model (ASD)

<table>
<thead>
<tr>
<th>Category</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>t-value (p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T (Intercept)</td>
<td>0.16124</td>
<td>0.003618</td>
<td>44.56 (p=0.000)</td>
</tr>
<tr>
<td>A</td>
<td>-0.00913</td>
<td>0.00454</td>
<td>-2.01 (p=0.0445)</td>
</tr>
<tr>
<td>E</td>
<td>-0.01360</td>
<td>0.00459</td>
<td>-2.97 (p=0.0031)</td>
</tr>
<tr>
<td>C</td>
<td>-0.01214</td>
<td>0.00469</td>
<td>-2.59 (p=0.0097)</td>
</tr>
</tbody>
</table>

All syllables in all phrasal positions were measured for their duration in Praat. The analysis was carried out both including and excluding phrase-final syllable tokens, which could be expected to undergo significant final lengthening; however, the outcomes of these two analyses were not significantly different and so the results from the full data set are reported here. The syllable structure of the utterances was not controlled; however, the read text consisted of 230 utterances in total, so the wide variety of syllable structures present should prevent there being any inherent imbalance between syllable structure types across the different topic structure categories or paragraph positions. The analysis included but was not confined to the utterances containing the target lexical items analyzed previously.

5.4.2 Results: Average Syllable Duration/Global Speech Rate

The syllable durations were not normally distributed, so the statistical tests were performed using the log syllable duration instead. A linear mixed-effects model including the syllable identity as a random effect\(^3\) showed that average syllable duration (ASD) was longer in new Topics than in other topic structure categories (see table 5.3). Syllables in new Topics were significantly longer than all other categories, with a mean length of 0.1612 seconds. There was some variation among the other categories, with Elaborations showing the shortest ASD; however, this variation was not statistically significant.

5.5 Discussion

The speech rate data from the two analyses above show the value of considering differences between internal and external characteristics of prosodic topic structure marking. While the

\(^3\) A similar analysis was conducted taking both the syllable and the speaker into account, but the random effect for the speaker was shown not to be a useful contribution to the model and was therefore discarded.
speech rate in the target lexical items (that is, locally/at boundaries) appears to increase in proportion to stronger boundaries, the overall speech rate in new Topics (that is, globally after the strong boundary) is slower than in other utterances. Thus we can say that the speech rate at edges of topic structure units may be more indicative of boundary strength, while speech rate within these units is more tied to the internal content of the utterances.

Overall the speech rate cues seem to be less robust than the F0 cues reported in chapter 3, which is consistent with other studies of non-F0 cues to topic structure, e.g. Shriberg et al. (2000). The local speech rate variation shows statistically significant variation across multiple categories only for some speakers, while the global variation is significantly different only for new Topics compared to the other categories, regardless of speaker. This suggests that although the speech rate cues provide support to other prosodic marking for topic structure, they are not the primary cues used to communicate this contrast.

The phenomenon of final lengthening is well known as a marker of prosodic unit boundaries, although the extent to which it varies in relation to boundary strength is still under debate. Beckman & Edwards (1990) and Wightman et al. (1992) found strong effects of prosodic boundary strength, while work by Gussenhoven & Rietveld (1992) suggested that “association domains” might be more important than prosodic phrase boundaries as such. White (2002) called into question Wightman et al.’s (1992) results in particular, due to their study being based on a piece of circular reasoning: the strength of prosodic boundaries they investigated was in part determined by the labellers on the basis of final lengthening, which was then measured in the study. However, there does seem to be a certain amount of consensus that the strength of prosodic boundaries is relevant to the amount of observed final lengthening.

Fougeron & Keating (1997) found that segments in initial and final position in a variety of prosodic units undergo “articulatory strengthening” such that consonants have greater linguopalatal contact in phrase-initial position and vowels have less contact in phrase-final position. They argue that this articulatory variation serves to heighten the contrast between the vowel at the end of one phrase and the consonant at the beginning of the new one, so as to make the boundary stronger. In addition to segmental initial strengthening phenomena such as this, a short-term increase in speech rate might contribute to maximally differentiating the end of the preceding unit from the beginning of the new one. Furthermore, since the
local variations in speech rate seem to coincide relatively well with topic structure for at least some speakers, we might say that it is actually varied independently of other initial strengthening phenomena, rather like the height of F0 peaks and the size of F0 falls were found to vary independently of one another in the previous chapter.

The combination of the boundary phenomenon of topic-unit-initial increase in speech rate and the internal phenomenon of decreased speech rate within new topic utterances could also be a perceptual cue in that it would create a heightened contrast (just as between local final lengthening and local initial speech rate increase); that is, in comparison with the surrounding speech, the increased rate on the initial word could appear even faster, and the slower rate of the sentence overall would be even more noticeable after the quick production of the first few syllables. In this way, the two types of variation could reinforce one another, and perhaps also contribute to the unity of the overall utterance, by keeping the internal portion consistent and more drastically contrasted with the boundaries. Also, the presence of multiple cues would mean that if a listener happened to miss one cue, the other would still be available to aid in comprehension, and might even allow the listener to guess in retrospect what prosodic cues were present on the missed portion. A perceptual test in noisy or otherwise adverse listening conditions might be able to suggest the extent to which this is true, but is beyond the scope of this thesis.
Chapter 6

Other Prosodic Cues to Topic:
Phonation Quality

Parts of the analysis in this chapter were previously reported in Zellers & Post (2010).

6.1 Introduction

Another type of variation that has been demonstrated to be a cue to prosodic boundaries is variation in phonation quality, and particularly the production in many languages, including English, of creaky voice (or vocal fry) at the ends of prosodic units. For instance, Henton and Bladon (1987) showed that glottalization was likely to occur, though not obligatory, at the ends of utterances (cf. also Redi & Shattuck-Hufnagel 2001). Other studies have shown that glottalization of some form is much more likely at prosodic boundaries (Pierrehumbert & Talkin 1992), as well as in other prominent locations such as pitch accents (e.g. Pierrehumbert 1995). Glottalization may occur both before and after the location of the prosodic boundary. Dilley et al. (1996) suggest that the glottalization of phrase-initial vowels in American English is a strengthening phenomenon, like the segmental phenomena investigated by Fougeron & Keating (1997). Huffman’s (2005) finding that syllable coda glottalization before obstruents is more likely in IP-final position is consistent with this account, if the higher incidence of glottalization is interpreted as evidence for an avoidance of coarticulation across these phrase boundaries. Similarly, Redi & Shattuck-Hufnagel (2001) show that final glottalization is more likely in utterance-final position, as well as in
utterance-medial position where there is a full IP boundary.

However, the fact that glottalization is optional in final position (Henton & Bladon 1987), as well as the observation made by Fougeron & Keating (1997), among others, that different speakers use different strengthening strategies at prosodic boundaries, raises the question of whether the distribution of glottalization is that simple in real production. One complicating factor in studying the incidence of glottalization at phrase boundaries is that glottalization or creak are associated with a number of other segmental or prosodic characteristics that occur independently of phrase boundaries. For instance, syllable-final voiceless stops in English may trigger the production of creaky voice on preceding segments if not to a complete allophonic substitution by a glottal stop (cf. Gordon & Ladefoged 2001), as is often the case with syllable-final /t/. Furthermore, creaky voice by definition has an F0 which is lower than that of modal speech, and this may be used by listeners as a perceptual cue to identify glottalization even in the absence of other acoustic cues, although other cues such as damping between vocal cycles might also be necessary in this case (Gerratt & Kreiman 2001). Similarly, some speakers may use a lowering of intensity instead of glottalization in comparable speech contexts (Dilley & Shattuck-Hufnagel 1995). Dilley et al. (1996) also found that glottalization was more likely in contexts following pauses. Since pause length can be a cue to boundary strength (Swerts 1997), this may present another possible factor which must be taken into account when investigating glottalization at prosodic boundaries. It may be difficult to separate the effect of prosodic boundaries versus that of simply pausing, if there is in fact a difference between the two, on production of glottalization in these contexts.

Despite the co-occurrence of these cues in many contexts, however, there is evidence that glottalization can occur independently of F0 and amplitude changes. For example, Pierrehumbert and Talkin (1992), Dilley et al. (1996), and Redi and Shattuck-Hufnagel (2001) all show glottalization in areas of mid to high pitch, which is not what would be expected if there was a necessary link between glottalized productions and low F0. Languages other than English show canonical productions of glottalization with High tones (cf. Gordon & Ladefoged 2001), which also suggests that there is no intrinsic link between glottalization and low pitch.

Furthermore, even within the class of phenomena that can be coherently identified as
Glottalization phenomena, there is still scope for differences. Gerratt & Kreiman (2001), in their review of studies on nonmodal phonation, list “low-frequency, damped pulses” (p. 370) as characteristic of creak, as well as identifying “low-frequency aperiodicity, [the alternation of large and small glottal pulses, [and] high-pitched phonation with intermittent subharmonics” (p. 371) as characteristic of vocal fry/creaky voice. This is identified as a separate phenomenon to creak itself, which may be assumed to be creak at a specific location rather than a modulation of phonation quality over a longer stretch of speech. Redi & Shattuck-Hufnagel (2001), on the other hand, identify glottalization as “…a region in the speech signal characterized by irregularly spaced pitch periods and often accompanied by other characteristics, such as full damping, low F0, breathiness, or low amplitude…” (p. 408). They argue that all of these acoustic features can contribute to an auditory impression of glottalization. They also point out that different speakers often use very different strategies to produce the acoustic effect of glottalization, even within the same dialect. The lack of a unified acoustic definition for glottalization can make analysis of such phenomena more challenging.

Despite these difficulties, the evidence reported above for glottalization as a cue to phrase boundary strength is suggestive. Since even untrained listeners are able to identify creaky voice, based on a variety of acoustic parameters, with 95% accuracy (Blomgren et al. 1998), it would appear that glottalization is a readily available acoustic parameter for use as a cue to prosodic phrasing. If this is the case, it should be possible to observe variation in glottalization running parallel to the topic structure categories noted above.

6.1.1 Hypotheses

Glottalization has been explicitly identified as a cue to degree of prominence or finality in English (Redi & Shattuck-Hufnagel 2001). On the basis of these previous studies, I hypothesize that the presence or absence of glottalization could be a cue to distinguishing discourse structure boundaries of different strengths. In particular, glottalization will be more likely to occur at boundaries of higher strength. That is, the ends of utterances preceding a New Topic utterance will be most likely to show glottalization, and ends of utterances preceding Continuations will be least likely to do so. It is also possible that the length of glottalized stretches will be greater (either absolutely or in proportion to the word or phrase length) at
boundaries of higher strength. That is, stretches of glottalization will be longest in utterances preceding a New Topic, and shortest in utterances preceding a Continuation. These hypotheses may be valid either simultaneously or independently of one another.

6.2 Aperiodicity and Topic Structure

6.2.1 Methodology

The data used for this analysis were the same as those used for the studies on F0 and speech rate production. A subset of five speakers (4 female and 1 male) from the above study were used for the current investigation. The speakers were those who fairly consistently produced identifiable aperiodicity in utterance-final position as an aspect of glottalized productions. While most speakers produced audible glottalization in utterance-final contexts, not all produced aperiodicity as an aspect of that glottalization.

Although aperiodicity is only one of the correlates of glottalization reported by other studies, it is consistently cited as one of the most common features of glottalized production. It is also relatively easily visually identified in a waveform, which contributes to ease of analysis. Therefore the analysis below will focus on the presence or absence of aperiodicity. This should not be understood as an assertion that aperiodicity is the only or even the primary cue to glottalization in speech production; it is simply an experimental simplification.

Following Dilley et al. (1996), two criteria had to be met for a stretch of speech to be labeled as glottalized (or specifically in this case, aperiodic). First, there had to be an auditory perception of glottalization, identified as creak, roughness or unevenness. Second, there had to be identifiable aperiodicity, or unevenness, in the waveform; the “Pulses” feature in Praat (Boersma & Weenink 2009) was used as an aid to visual identification. Instances of aperiodicity were excluded from the study if final glottalization could be related to the presence of (1) a voiceless plosive consonant at the end of the utterance, or (2) an initial vowel in the following utterance, since in these cases, it would be difficult to determine whether the glottalization was related to the phrase or topic structure boundary or to the segmental

---

1This function uses automatic methods to give a rough estimate of the location of glottal pulses in the waveform display, drawing marker lines at what it identifies as the beginning of each new period.
make-up of the tokens. In all instances that were kept, the utterance was followed by a silent pause. (See section 6.4 below for discussion of why pause length was not analyzed.) The resulting dataset consisted of 520 utterances.

One common measure of aperiodicity in sound signals is jitter, a calculation of the variation in length of successive periods in the signal. For all speakers investigated, the jitter of stretches labeled “aperiodic” was approximately double the jitter of the modal stretches preceding them, as calculated by Praat. The jitter of modal stretches was approximately 0.0123, while the jitter of aperiodic stretches was 0.0237 (t-test, t=-6.18, df=208.36, p<0.001). All other acoustic measurements were also taken in Praat.

6.2.2 Results

6.2.2.1 Occurrence of aperiodicity at boundaries

Aperiodicity as a cue to glottalization was present in 80.1% of the utterances studied (after discarding items as described in 7.3.1). For the male speaker in the study, 92.9% of the utterances were produced with final aperiodicity. For the four female speakers, final aperiodicity was produced in 74.8% to 82.4% of the utterances.

The set of data contained 91 utterances preceding new Topics, 150 preceding Additions, 135 preceding Elaborations, and 144 preceding Continuations. Although the absolute number of items with aperiodicity preceding each topic structure category varied for each speaker, the overall trend was similar, and therefore all speakers are considered together (see figure 6.1). Stretches of aperiodicity were possible in all boundary conditions (that is, preceding all four topic structure categories) and in any sequential position in the group of utterances (that is, initial, medial or final). Aperiodicity was common in positions preceding new Topics (85%), Additions (90%) and Elaborations (91%), which did not vary significantly from one another. However, in utterances preceding Continuations, 44% of utterances had no aperiodicity; this difference from the other three categories is statistically significant ($\chi^2=72.93$, df=3, p<0.001). There was no main effect of the position of the utterance in the group of utterances, nor any interaction with this factor; in particular, utterances in final position in the group of utterances were no more likely to contain aperiodic stretches than utterances in other positions, except that utterances in final position were by definition
never followed by Continuations.

### 6.2.2.2 Length of aperiodic stretches at boundaries

To compare the length of aperiodic stretches in different topic structure conditions, a subset of directly comparable target words\(^2\) was analyzed further (N=120). These words were all three syllables long, with lexical stress falling on the first syllable, and they were composed only of voiced sonorant segments. Differences between the target words were not found to be significant, so they are treated together in the following discussion.

There was no effect of topic structure category on the length of aperiodic stretches in the utterances. In the overall results, there were no significant differences between groups

\(^2\)See section 3.2.2 for more detailed description of the matched target words used in the study.
Figure 6.2: Length of aperiodic stretches preceding topic structure categories

Length of Aperiodicity by Topic Structure Category
CHAPTER 6. OTHER PROSODIC CUES TO TOPIC: PHONATION QUALITY

137

(see figure 6.2).

For some speakers, there were significant differences between two or three categories, but these did not follow any consistent pattern, and may therefore be assumed to be incidental, unsystematic variation (at least with regard to the topic structure categories). There was also no effect of the position of the utterance in the utterance group; utterances in final position in the utterance group did not show longer or shorter stretches of aperiodicity than utterances in initial position in the group. This was true when measuring the absolute length of aperiodicity, the length of aperiodicity in proportion to the word, or the length of aperiodicity in proportion to the intonational phrase. There was a significant effect of speaker, with the speakers falling into three groups: the male speaker produced the longest aperiodic stretches, and the female speakers clustered into two pairs producing longer or shorter aperiodic stretches (ANOVA, F(4, 119)=5.93; p<0.01).

6.2.2.3 Discussion

Neither of the initial hypotheses proved sufficient to explain the aperiodicity data presented here. The first hypothesis was shown to be only partly true: although utterances preceding a Continuation showed a different pattern with regard to the presence or absence of an aperiodic stretch, the other topic structure categories were not distinguishable using this variation. The second hypothesis was shown to be completely incorrect: there was no apparent relationship between the topic structure categories and the length of the aperiodic stretches.

It would seem, then, that aperiodicity may not be the best choice of cue to signal subtle differences between boundary strengths, at least not on the level of discourse structure. However, the varying presence of aperiodicity in different contexts, as well as the perceptibility of the modal-creaky contrast by untrained listeners, suggests that aperiodicity may still be varied meaningfully, perhaps to aid in some other contrast. Since the second analysis compared different tokens of the same target word in similar contexts, we may discard segmental/lexical effects for the moment. However, phonation quality changes are often tied to F0 changes (cf. Gerratt & Kreiman 2001). It is possible that variation in phonation quality could either result from or possibly contribute to an F0 or tonal (phonological) contrast. Dilley et al. (1996), for example, have suggested that glottalization could be associated with
an L* in the tonal structure; this possibility will be explored in section 6.3.

### 6.2.3 Alternative boundary characteristics: aperiodicity and syntax

Another possibility is that the varying distribution of aperiodicity in the Continuations was conditioned by the varying syntactic constructions involved (Anne Wichmann, personal communication). Since Continuations of previous utterances had different characteristics, they might also relate to different boundary strengths which were unrelated to the topic structure categories. For example, a subordinate clause to be followed by a main clause might be less likely to have aperiodicity than a main clause followed by a conjoined main clause, since in the latter case the syntax of the first utterance would be complete independently of any follow-on.

In order to test this possibility, the Continuations were labelled according to their syntactic characteristics as shown in Table 6.1. Then the varying distribution of aperiodicity, and the length of aperiodic stretches, in the utterances preceding them was compared across these categories.

<table>
<thead>
<tr>
<th>Syntactic Type</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coordinated Main Clause (CMC)</td>
<td>He had to walk, <em>and he memorized every step on the way.</em></td>
</tr>
<tr>
<td>Main Clause Following Subordinate (MCFS)</td>
<td>When an animal escaped from a laboratory, <em>he spent hours helping to chase it down.</em></td>
</tr>
<tr>
<td>Subordinate Clause (SC)</td>
<td>The other men hid <em>while Emory faced down the creature.</em></td>
</tr>
<tr>
<td>Non-Sentence Coordinate (CC)</td>
<td>The King honoured him with a speech <em>and an invitation to dine at the palace.</em></td>
</tr>
<tr>
<td>Adverbial (ADV)</td>
<td>The Napoleonic wars continued in Europe, <em>annually seeming to gain momentum.</em></td>
</tr>
</tbody>
</table>

Table 6.1: Syntactic categories of Continuations with examples (Continuations shown in italics).
Figure 6.3: a. Presence or absence of aperiodic stretches across categories (in percent of preceding sentence tokens with aperiodicity). b. Length of aperiodic stretches across categories.
since the utterance in question would have been syntactically complete, there was not a statistically significant difference between the groups ($F(5, 100) = 0.3493, \text{n.s.}$).

The lack of any significant effect of the syntactic structure on the use of aperiodicity preceding continuations may be interpreted as being a result of the fact that the text used was a read text. The existence of the full written text meant that speakers were aware of whether there was more to come at any given point, regardless of whether the structure of the preceding utterance was syntactically complete. In a spontaneous interactive situation in which the speaker must signal the (potential) end of a turn, there might be a clearer indication of what the speaker considers to be “complete.”

### 6.3 Aperiodicity and F0

#### 6.3.1 Analysis and results

The aperiodic stretches in utterances did not normally extend all the way to the end of the utterance; in most cases, there were several regular pitch periods at the very end, within a similar pitch range to that of the voiced stretch preceding the aperiodic region; these stretches, though occurring with low amplitude, were still clearly audible. The pitch movements before and after the aperiodic stretches were identified as either Rising or Falling on the basis of F0 measurements and auditory judgments. (Note that the label “Rise” in the post-aperiodic stretch could also be applied to high pitch, and “Fall” in that context could be applied to non-high or non-rising pitch.) If we now examine the length of aperiodic stretches in these different contexts, we find an interacting effect of the F0 movements preceding and following the aperiodic stretches (see figure 6.4). Rising movements followed by low or falling movements show the shortest aperiodic stretches. This is fairly unsurprising, given that (at least in English) glottalization is not generally associated with high pitch, and the change from a rising to a falling movement would by definition include a pitch peak. However, the change from a falling movement to a high or rising movement, where we would expect an F0 valley, does not show the longest stretches of aperiodicity. Instead, the longest stretches of aperiodicity are found when the directionality of the pitch movement is consistent: a fall to a fall, or a rise to a rise. Statistically, there is no main effect of either the preceding or the following pitch movement; only the interaction is significant.
6.3.2 Discussion

The correlation of longer aperiodic stretches to stretches where the directionality of the pitch movements is consistent suggests that variation in the length of aperiodicity may be a phonetic effect which is modulated to help create acoustic space for more complex pitch movements (and by extension, more complex tonal configurations). Perhaps the length of an aperiodic stretch aids listeners in identifying what kind of F0 movements are occurring at the ends of phrases. This could be especially relevant given that the falling off of amplitude utterance-finally and the relatively short length of the post-aperiodic stretches could create adverse listening conditions for perception of the tonal configuration. A more detailed study of the interaction between F0 and aperiodicity (and other correlates of glottalization), as well as a perceptual study of this phenomenon, could test this hypothesis, but these are beyond the scope of this thesis.

Alternatively, the length of aperiodic stretches could be related to the amount of attention on the vocal folds by the speaker. Creaky phonation could be seen as a correlate of there being less attention to the vocal folds; more complex pitch movements would require the
speaker to attend more strongly to the phonation mechanism, thus decreasing the length of stretches of aperiodicity and/or other phonetic cues to creakiness (Francis Nolan, personal communication).

6.4 General discussion: non-F0 prosodic characteristics

In this chapter and the previous one we have seen two types of non-F0 variation which may each contribute, to a varying extent, to the marking of topic structure in spoken discourses in SSBE. First, contrasting variations in speech rate mark both boundaries of topic structure units and serve as an internal marker of their topic structure status. Second, variations in the presence of creaky phonation at the ends of utterances appear to be a weak cue to the topic structure level of the following unit; specifically, whether or not an utterance will be followed by a Continuation. Since the definition of a Continuation is that it completes a speech act begun in a previous utterance, aperiodicity/glottalization could be seen as a “completeness” marker rather than a boundary cue as such.\(^3\)

There are a number of other non-F0 cues to topic structure status that have been examined in the literature which were not investigated here. These include pause length, degree of final lengthening, and loudness/amplitude, among others. Pause length appears to be a relatively uncontroversial cue to boundary strength in many languages (cf. Lehiste 1970 and many others for English, Swerts et al. 1994 for Dutch, inter alia), and there is no reason to expect that it is not present as a cue in speech, and specifically in this data. However, the method in which the text was presented to readers (that is, divided into paragraph groups) would have interfered with the accuracy of the data collection in this study, since the boundaries of the topic groups were visually “forced” on the readers\(^4\). This means that there were other, topic-unrelated factors which could have caused long pauses, since in practice written paragraph boundaries often have much more to do with ease of visual interpretation than with actual topic organization (see discussion in Wichmann 2000). For this reason, no analysis of pause length was made in the current data; but this should not be seen as a rejection of pause length as a cue.

\(^3\)The results of the syntactic analysis (section 6.2.3) suggest that syntactic completeness, at least, is not what is at stake in this particular case.

\(^4\)A better methodology might have been that employed by Tseng et al. (2004; reported in Tseng 2010), in which paragraphs were much longer and the only punctuation was commas.
The cases of loudness/amplitude and degree of final lengthening are a bit more complex. A rough analysis on a subset of the current data suggested that there was no systematic variation in RMS amplitude either at the beginnings or ends of topic structure units. This is contrary to the findings of Herman (2000), who found that the RMS amplitude trace of discourse-medial units was of a greater magnitude than the trace of discourse-final units. Her study compared the same sentence in different contexts, and so she was able to compare more global trends across these sentences, whereas my study compared only target words in different positions; the fact that the sentences were segmentally different meant that their amplitude contours were not comparable. Furthermore, my target items were spread out over 7 recordings for each speaker; although the recordings were made with the same settings and all in one session (see section 3.2.4), it is likely that speakers were not entirely consistent in their amplitude across all the recordings, if only because of e.g. shifts in body position bringing them nearer to or farther from the microphone. It is possible that the design of my study obscured variations in RMS amplitude in relation to either topic structure or position in the topic structure group. However, the systematicity of variation in RMS amplitude as a cue in spontaneous speech seems debatable. Herman (2000) found a systematic effect of RMS amplitude related to her topic structure categories, but amplitude remains relatively unstudied in this context, so it is not clear to what extent these findings may extend beyond her study. It is important to note that in other contexts where amplitude has been tested as a measure of prominence, the findings have been variable: intensity has been shown to be a component of lexical stress in English (Beckman 1986; Kochanski et al. 2005) and Spanish (Llisters er et al. 2005), but it appears to be less reliable in Japanese (Beckman 1986) and Dutch (Sluijters & van Heuven 1996, Sluijters et al. 1997). It is unclear to what extent these differences are based on methodology versus language-related differences; Beckman’s (1986) study would suggest that the primary difference is language-related, but the differences between English and Dutch are surprising in this regard, since these two languages often show relatively similar prosodic behavior.

The lack of findings regarding RMS amplitude in this study is inconclusive; more evidence, preferably in more spontaneous speech situations, would be required to be able to make any useful claims about RMS amplitude as a cue to discourse structure. Some research has also focused on the interaction between F0 and RMS amplitude, particularly in terms of
increasing the perceived prominence of F0 (Herman et al. 1996). Turk and Sawusch (1996) suggest that pitch and amplitude are perceived together and that ultimately any independent contribution to prominence from amplitude is negligible. It is possible, therefore, that RMS amplitude is not so important on its own, while still having a strong effect on the perception of the F0 variations reported in chapter 3. On the other hand, some automatic systems trying to identify prominent syllables, such as that reported by Silipo and Greenberg (2000) for American English, have found amplitude to be useful for this task. More research is therefore required to establish the relationship between F0 and amplitude in perception.

The case of final lengthening is also complex. White (2002) reports interaction between nuclear-accent-related lengthening and final lengthening in English, as well as noting that word length and structure can influence the degree of lengthening that is possible. In particular, the magnitude of lengthening (or possibly the degree of flexibility) of unstressed syllables is greater than that found in stressed syllables, meaning that the rhythmic structure of words in phrase-final position may have a strong effect on the degree of final lengthening observed. This appears to be the case cross-linguistically as well; a study on Taiwan Mandarin (Lin & Fon 2009) found that the degree of final lengthening is not highly correlated with the strength of a discourse structure boundary, despite the apparent variation within IPs and smaller phonological constituents. Therefore it seems possible that the degree of final lengthening is not so important as its mere presence at a final boundary. Since all utterance boundaries were also IP boundaries, final lengthening was present in all cases in the data reported in this thesis: a rough analysis based on the ASD data from section 5.4.2 indicates that the mean duration of IP-final syllables was 0.276 seconds, while the mean duration of other syllables was 0.156 seconds (t-test: t(210.133)= -17.61, p<0.001). However, an ANOVA showed no significant difference between the duration of the IP-final syllables on the basis of the topic structure of either the current or the following utterance (F(3,181)=1.2, n.s.). As this is consistent with the findings reported above, the degree of final lengthening in relation to the topic structure was not investigated further.
6.5 Conclusion

This chapter and the previous one reported experimental results on two non-F0 types of prosodic variation, speech rate and aperiodicity, and their relationships to the topic structure of discourse. The results of these studies show the importance of investigating non-F0 (or non-intonational) cues in addition to the more canonical pitch-related variation, since they appear to interact with the F0 as well as playing an independent role in the prosodic signalling.

Taken together, the results from the production study (chapters 3-6) reveal a rich prosodic signalling system for topic structure, involving a variety of cues occurring at different times during the discourse, and allowing for a high degree of recoverability; that is, even if a listener misses one cue or it is obscured by interaction with some other prosodic characteristic, other signals are available so that the listener should still have the topic structure information available. As yet, it has not been shown to what extent listeners make use of this information. Before addressing this question, however, we will turn briefly to the question of the validity of the topic structure categories in question.
Chapter 7

Validation of Topic Structure Categories

It is possible that some of the phonetic variation reported in chapters 3-6 could be due to the readers having a different interpretation of the topic structure of the text than that assumed by the author. This means that it is important to determine to what extent the topic groups and the categories themselves are used in the same way when applied by different people. This chapter reports several post-hoc analyses of the topic organization of the text used in chapters 3-6, with the aim of determining to what extent the categories used were applicable to the text and can therefore be assumed to have been used by the readers.

7.1 Topic structure in different types of discourse

The topic structure categories adopted in chapters 3-6 were adapted from studies using spontaneous task-oriented dialogue (Nakajima & Allen 1993) and BBC newsreaders’ news story monologues (Wichmann 2000); see discussion of this process in section 3.2.2.

One of the benefits of adapting topic structure categories from these previous studies for use in the current research was that it would be theoretically attractive to be able to use similar types of categories for different types of discourse. However, the adoption of these categories for a narrative text could be criticized on the grounds that narrative texts have been traditionally described by different types of structural categories than those proposed in the current study. For example, Labov and Waletzky (1967) argue that narratives are a
unique type of text, being perhaps the only kind of spoken discourse with a definable beginning, middle and end. Propp (1968) comes to a similar conclusion, pointing out that folktales have a consistent structure or series of events. Labov and Waletzky (1967) call the elements of a narrative abstract, orientation, complicating action, result/resolution, evaluation, and coda. Chafe (1994) narrows these slightly into orientation, complication, climax, denouement, and coda. Chafe suggests that the “topic” of a narrative, as an unexpected idea, likely coincides with the climax of the narrative. This link appears to be problematic, though, for any but the shortest narratives in light of the other definition of discourse topic that he provides. If discourse topic is “...an aggregate of coherently related events, states, and referents that are held together in some form in the speaker’s semiactive consciousness” (1994: 121), and which need not be made overt, then the idea of connecting the topic of a narrative with its climax seems to make sense only in a situation where a narrative is fairly minimal. Otherwise, verbalizing this aggregate (as must be done if it is the climax of the story) could potentially be overwhelmingly difficult due to its large scope. Therefore a longer narrative, while it may be divided into the five macro-elements mentioned above, is likely to also have a more complex internal structure, containing a variety of different topics representing the ever-changing set of active “events, states, and referents,” as Chafe (1994) acknowledges himself. Chatman (1978) divides narratives into “kernels” (major events) and “satellites” (minor events), where each satellite has its own internal structure as well as being hierarchically subordinate to a kernel. Again we see a distinction between macro-elements and micro-elements: the overarching narrative is the series of kernel events, which may not be modified without disrupting the narrative; but satellites may go off in their own direction (i.e. on a different topic) without doing so.

It is therefore important to consider this internal structure as well as the macro-structure of a narrative. To this end, we may examine the case of what Chafe has called “episode shifts” (1994: 138): if the scene changes, or a character’s point of view changes, or there is a jump in the time-course of a narrative, there is by definition a small discontinuity. Sometimes more than one of these shifts can occur at once. Chafe argues that when this occurs, the more shifts that are present at one location in the text, the stronger the boundary between episodes. This kind of episode shift could be seen as congruent with the type of topic shifts addressed in the production experiment, since a new Topic requires a strong
discontinuity from what has come before, while Additions, Elaborations, and Continuations require increasingly strong continuity with previous utterances. Although the topic structure categories proposed in the current study do not make explicit links to shifts in setting, time, or character perspective, they could be interpreted as the following:

- **New Topic**: multiple shifts of any type (character, setting, time) possible (or even obligatory)
- **Addition**: one-two shifts of any type possible
- **Elaboration**: one shift possible; may be a restriction on type e.g. no character shift
- **Continuation**: no shift possible

In this way, the current categories could be brought roughly into line with some previously proposed theories of narrative structure, creating greater continuity between descriptions of types of texts (see also section 9.2.2 for further discussion of the different theoretical approaches available and how they might be reconciled).

### 7.2 Motivation

With any kind of category applied to a text, it is important to show that speakers are sensitive to the kind of organization that is proposed. The following sections report a series of experiments comprising a post-hoc validation of the category structure used in chapters 3-6 as it was applied to the Emory story. The goal of these experiments was to determine how closely the category structure applied matched readers’ own intuitions about the structure of the text, and to investigate how readers’ alternative structures might be reflected in the prosodic characteristics of the spoken text.

There are three issues at stake in validating the categories used in chapters 3-6. The first is whether or not the topic boundaries and topic-structure categories can be shown to be present in the data on the basis of independent evidence. This raises the sub-question of what kind of evidence is sufficient to demonstrate that the categories are or are not there. The second issue is to what degree the evidence for the presence or absence of the categories changes the results reported in the preceding chapters. The third, closely related to the
second, is whether the results of a validation shed any further light on the phonetic results (either the original results, or the results of a modified analysis).

Since the categories used in chapters 3-6 are not those that have typically been applied to narratives, there is no precedent or standard for how they ought to be applied beyond the descriptions given in section 3.2.2. One way to establish such a standard is to ask a group of raters to apply the categories to the text and to compare their responses. A rating task addresses both whether the category definitions are interpretable at all (i.e. if readers can understand them), and also whether the categories represent a psychological reality in terms of the organization of the text (i.e. if readers apply them consistently with one another).

One drawback of a rating task is that it is somewhat difficult to determine what degree of agreement constitutes evidence for or against the use of the topic structure categories. Given a large number of raters, 100% agreement would be highly unlikely in any study of this kind. Furthermore, we do not expect readers to agree on the interpretation of the structure of a text, particularly in reading: Swerts (1997) found that people who read a text were less consistent in marking boundaries than those who listened to the same text. Therefore, we must allow for a certain amount of disagreement even if the categories are valid for the text. On the other hand, agreement that is only minimally above chance level would not constitute particularly convincing evidence for the categories in question. Therefore, unless agreement is extremely high or extremely low, we must be cautious in our interpretation.

Despite this caveat, the validation of the topic boundaries and the topic structure categories on the basis of a rating task allows for the establishment of a baseline against which to compare the original labels. This makes it possible to test whether the phonetic analyses in chapters 3-6 are in fact valid. If the topic organization given by raters is sufficiently similar to the organization originally used, we can accept these results as accurately reflecting the way in which prosody is used to signal topic structure in SSBE. If the topic organization given by raters is different than the organization used in the production study, a comparison of the phonetic data in the original versus the validated organization will shed light on the ways in which the interpretation of the data might be (somewhat) different than that originally discussed in the preceding chapters.
7.3 Validation Part 1: Topic Boundaries

In the experiment reported in chapters 3-6, participants read from a text that was divided into paragraphs. The visual paragraphs corresponded with the intended topic structure of the text; that is, each visual paragraph represented one topic unit. The topic units were of approximately equal length in order to make the phonetic analysis more straightforward. However, there is no linguistic requirement that topic units have similar lengths, or indeed that they correspond to written paragraphs (which are often divided up for aesthetic purposes, or to ease reading, rather than on the basis of topic alone, cf. Wichmann 2000). Therefore, although the first utterance in a paragraph was considered a new Topic in chapters 3-6, this is not necessarily the case. The first validation experiment investigated how readers of the text with no paragraph or punctuation cues divided the text into different topics.

7.3.1 Methodology

The topic boundary validation experiment had two versions, each employing a slightly different methodology. In the first version, the text was presented in line-by-line form, with each line being an “utterance” (as defined in the text description in section 3.2.2). Participants were asked to label each line with a topic number (e.g. T1, T2, T3; see figure 7.1a). They could label as many lines as they wanted with the same topic number, and if they felt that a line belonged to more than one topic, they could give both topic numbers. In the second version, the presentation was as one continuous text, with no paragraph boundaries (line breaks would have been determined by the width of the participant’s computer screen). Participants were asked to select (contiguous) chunks of text corresponding to topics and paste them into text boxes (see figure 7.1b). In both cases, the text was presented without sentence punctuation (except apostrophes) or capitalization.

The first version was designed for ease of analysis; the utterances could be easily matched up with the original text labels from the production experiment and directly compared. The second version was designed to give participants more freedom, particularly in areas where there was potentially ambiguous syntax. This meant that participants were able, if they wished, to identify a new topic starting in “the middle” of what the original
Figure 7.1: Screenshots of topic boundary task in (a) line-by-line format and (b) long form.
CHAPTER 7. VALIDATION OF TOPIC STRUCTURE CATEGORIES

experiment had analyzed as a complete utterance. A comparison between responses to the
two versions should show whether there was any effect of the task on participants’ decisions
about the text organization.

The experiment was presented online via a Google Documents form. Participants were
self-selected according to the criteria that they were native speakers of English, aged 18-40,
with no known linguistic impairments. Upon completion of the study they could submit
their email address to be entered into a drawing to win a £25 or £15 gift voucher to Amaz-
on.co.uk. All three tasks discussed in this chapter (two here and the third in section 7.4)
were presented via one webpage, with participants selected semi-randomly to take part in
one of the tasks by clicking a link on the basis of the first letter of their surname; i.e. each
participant completed only one of the three validation tasks.

7.3.2 Results

7.3.2.1 Version 1: line-by-line topics

Twenty participants responded to the line-by-line topic structure experiment; of these, one
was excluded due to an apparent failure to understand the task.\textsuperscript{1} Thirteen of the participants
were female; ages ranged from 18-33. Ten reported being speakers of southern British En-

This participant only used three labels (T1, T2 and T3) and jumped between them every two or three lines.
\textsuperscript{2}Landis and Koch (1977) do not provide any empirical basis for their descriptive interpretations of kappa
values; this is simply a convention.
which the relevant contrast is only between topic changes and topic holds; this is because the value calculated for agreement by chance is high when there are only two possible labels (cf. Rietveld & van Hout 1993). This means that the resulting value returned for agreement is a very conservative estimate in such cases.

Second, a strict definition of “topic” was not given in the task instructions. This means that readers may have used different strategies to determine what they thought the scope of a topic was; and indeed, a visual inspection shows this to be the case. A hierarchical cluster analysis and a multidimensional scaling analysis were used to compare participants’ patterns of response. Of the nineteen participants included in the analysis, six (the left side of the cluster analysis, figure 7.2a, removing labeller 16) labelled a total of eleven topics or fewer, while the other thirteen (the right side of the cluster analysis, figure 7.2a) identified between twenty-seven and fifty-five topics. Thus we can group the participants into those who identified “macro-topics” and those who identified “micro-topics”. For those who used micro-topics, $K = 0.359 \,(p<0.01)$. This agreement level is despite the fact that the labeller identifying the most micro-topic boundaries (53) identified almost double the number of topic boundaries than the labeller identifying the fewest number of micro-topic boundaries (27). Among the macro-topic labellers, agreement about the beginnings of new topics was much lower, achieving only $K = 0.16 \,(p<0.01)$, or “slight” agreement on Landis and Koch’s (1977) scale. However, a visual inspection of the data suggests that the overall scope of their topics was similar and that they disagreed about the specific location of the topic beginnings rather than the scope of the topics themselves; the topic beginnings were in similar regions of the text for all six macro-topic labellers.

The characterizations of the dimensions in the multidimensional scaling analysis are difficult to define, but Dimension 1 seems to roughly correspond with the number of topic boundaries identified (see figure 7.2b).

In addition to comparing agreement between the labellers in the line-by-line task, it is also important to see whether the topic boundaries they identify are consistent with the topic boundaries used in the studies in chapters 3-6. Twenty-two of the utterances in the study received a majority vote (at least 50%) as new topics. Of these twenty-two, twenty-one had been labelled as new Topics in the production study; only one utterance was consistently rated as a new topic which had not been considered a new topic in the production study.
Figure 7.2: (a) Clustering of line-by-line labellers in a hierarchical cluster analysis. (b) Results of a multidimensional scaling analysis of line-by-line labellers. Labeller 16 was excluded from the final analysis due to having performed drastically differently from all other labellers on the task, as these analyses show, while labellers 2 and 6 gave (apparently) almost identical characterizations of the text.
The utterance in question was the following, originally categorized as an Addition (given in context):

he made great contributions to science / yet history has forgotten him / [also voted as new topic] emory was born in 1810 / it was a year that would live in memory / the napoleonic wars continued in europe...

It appears that either the utterance in question or the preceding utterance was sufficiently disjunct from what went before as well as what followed to create an ambiguity as to their topic membership. Four of the raters labelled both of these utterances as the beginnings of new topics.

Taking the reverse point of view, of the 47 utterances identified as new Topics in the production study, 21 had a majority vote (10 or more) as topic changes, 9 had at least 33% votes as topic changes (7 to 9 votes), and the other 17 had six or fewer votes. One utterance, the following, had no votes as a topic change:

mr rinnering and the other men hid / while emory faced down the creature / with great shouts and bellows / emory forced the creature away...

Although in the production study context the syntactic link of this utterance with the following was forced by the orthography, in the validation experiment, the link was ambiguous, and it is probably not surprising that readers associated it with the preceding utterance rather than the following. Furthermore, if a topic shift is actually present at this location, it is probably better characterized as a “stepwise transition” (cf. Jefferson 1986) than an obvious topic discontinuity.

7.3.2.2 Version 2: long-form topic

Eight participants responded to the long-form topic experiment. It is likely that the reason for lower response to this task was a perception that the task was harder than the line-by-line task (since the text was presented in one chunk and may have looked longer). Of the eight participants, five were female; ages ranged from 18-32. Three were native speakers of Southern British English; the other five were speakers of American English.

It appears that the longer text biased readers to having longer topic units in general; all except two had between 10 and 15 topics (though all had more than the “macro-topic”
labellers in section 7.3.2.1 above). Furthermore, there was a much higher degree of agreement between the labellers than in the line-by-line task. Overall K = 0.363 (p<0.01) for all labellers, and if the two labellers who had far more topics than the others are removed, K increases to 0.404 (p<0.01), or “moderate” agreement according to Landis and Koch (1977).

Five locations in the text received unanimous or near-unanimous (six or seven) votes as topic changes; all five coincide with locations that had been marked as topic changes in the original analysis, and with the locations receiving majority votes as topic changes in the line-by-line analysis. Given the higher agreement level, locations that did not have unanimous or near-unanimous votes generally had only one or two votes if any as being topic boundaries. For the two locations that had four votes (50%), both coincided with majority vote locations in the line-by-line task and with topic boundaries used in the original analysis. No location that had not been marked as a topic boundary in the original task received more than one vote as a topic boundary, and most received zero; that is, even when participants did not agree with each other about the number of topic boundaries, they tended to place them in similar locations to one another and to the original labelling scheme from the production study.

One of the goals of the long-form task was to allow participants to place topic boundaries at locations other than the utterance boundaries defined in the line-by-line task (and the original analysis). However, in practice, only one of the readers placed a boundary at a location that would have been impossible in the line-by-line task, and this occurred only once. This suggests that the syntax was not too ambiguous, and that both the line-by-line task and the long-form task were successful in doing away with the biases inherent in the original production task due to the presentation of the text in paragraphs.

### 7.3.3 Prosodic characteristics at validated topic boundaries

In order to examine the validity of the phonetic results of the production experiment, we may now observe the prosodic characteristics of utterances which were rated as topic changes by the majority vote. For this purpose, the 22 items which received a majority vote as starting new topics in the topic boundary tasks were labeled as representing new topics, while the other 25 items which had been considered new topics in the original analysis were discarded. These items were discarded, rather than moved to a different category, since there was no
immediate evidence available as to which category they should belong in (see section 7.4.2 below for discussion of the redistribution of utterances into categories).

7.3.3.1 Results: F0 analyses

The set of validated new Topic utterances contained many more instances of F0 rises than F0 falls; the full data set produces 63 pitch rises in Topic position and only 10 falls. (This number excludes tokens which do not provide useful measurements due to problems such as creaky production interfering with one or both pitch measurements.) This is consistent with the results reported in chapter 4, showing that new Topics were much more likely to contain rises than falls.

The peak accent distribution analysis arose out of the necessity of dealing with the data on F0 peak alignment (cf. section 3.2.6.4), which were not entirely consistent with Wichmann’s (2000) report that new Topics showed later F0 peak alignment than other topic structure categories. Although chapter 4 provided one way of resolving this discrepancy, namely by explaining the difference in peak timing in terms of a difference in phonological analysis of the pitch accents in question, it is also possible that items that did not actually constitute topic changes were included in the analysis, thus contributing noise to the data. Therefore it is desirable to recreate the original F0 peak delay analysis using the new set of topic boundaries as identified in the validation task. As in the original analysis, F0 peak alignment is measured as a ratio of the time between the onset of the stressed vowel and the F0 peak compared to the length of the stressed vowel. The data in figure 7.3a appear to reverse the results of the original dataset, with items in the anacrusis condition showing a delayed F0 peak in new Topics, while items in the initial condition showing slightly earlier F0 peaks in new Topics; in other words, the result here is similarly ambiguous to the original result, even though the direction of the pattern is reversed. Figure 7.3b compares the validated new Topics to the other utterances that had originally been labelled new Topics, suggesting that there is no consistent difference between the two categories. An ANOVA found no significant differences between any of the topic structure categories (F(3, 471)=3.55, n.s). There was also no significant difference in this case for the anacrusis condition, but since there were only 4 tokens of validated new Topics in the anacrusis condition, this is not surprising. The results of the validation experiment therefore do not appear to
Figure 7.3: F0 peak alignment (delay ratio) by topic structure category. A = anacrusis, I = initial (no anacrusis) (a) shows the validated new Topics compared to the other categories (with no changes); (b) shows the validated new Topics (T) compared to the other items formerly labelled new Topics (M = mid).
change the overall conclusions that may be drawn from the F0 peak alignment data.

While the distribution of rises and falls is consistent with the results reported in chapter 4, it means that it is impossible to conduct a statistically valid verification of the F0 fall range results reported in chapter 3. As can be seen in figure 7.4, the pattern for the size of the F0 fall range is still nearly identical to the pattern reported in chapter 3. Due to the small number of tokens of new Topics, it is not possible to attain a statistically significant effect separating the new Topics from the middle categories (Addition and Elaboration), although a significant effect was obtained for the difference between new Topics and Continuations (ANOVA: F(3, 423)=12.88, p<0.05 with the Bonferroni correction applied). The validated new Topics had a marginally smaller mean pitch range (4.15st, sd=1.47) than the remaining utterances that were originally labelled as new Topics (4.49st, sd=1.34), but the difference in means did not attain statistical significance (t-test: t(14.5)=0.6477, n.s.).
7.3.3.2 Results: local speech rate

In the case of the local speech rate analysis (chapter 5), there are more data available for a reanalysis using the validated topic boundaries. Figure 7.5 shows the word length of items in the validated new Topic category compared to the other (unchanged) categories. As in the analysis reported in chapter 5, new Topics have the shortest word length; i.e. they are produced at the fastest speech rate. An ANOVA confirms this result statistically ($F(2, 915)=15.17, p<0.05$ with the Bonferroni correction applied). The mean length of the validated new Topics (0.304, $sd=0.058$) was slightly shorter (i.e. faster) than the mean length of the remaining original new Topics (0.324, $sd=0.056$); this difference nearly attained statistical significance ($t$-test: $t(159)=2.0893$, unadjusted $p=0.038$). \footnote{This test is not statistically significant when the Bonferroni correction is applied, but is reported due to its nearness to the critical value for $t$, 2.268 in this case.}

The data reported here are consistent with the data reported in chapters 3-6, despite the use of a much smaller subset of new Topic utterances to obtain the results. Although there
were too few items in the F0 fall range case to conclusively confirm the results reported in chapter 3, the consistency of the local speech rate data with the previous analysis in chapter 5 suggests that it is unlikely that the overall findings of the two studies would show any major discrepancies. Furthermore, the difference in speech rate between the originally-labelled new Topics and the validated new Topics strongly suggests that despite the strong cue of the visual paragraph boundaries in the production study, the readers may still have been sensitive to the difference between a paragraph boundary and a topic shift.

7.4 Validation Part 2: Topic Structure Categories

As discussed in section 7.1 above, the topic structure categories used in the production study in chapters 3-6 have not previously been applied to a narrative text, and are not necessarily congruent with the types of discourse structure that have been applied to narratives (although they are also not necessary in conflict). Because these categories have not been applied to a narrative before, it is important to know whether they can be considered an accurate characterization of how readers interpret the structure of a narrative. The second validation experiment therefore tested the degree to which readers were able to apply the four topic structure categories consistently to the text.

7.4.1 Methodology

As in the first version of the topic boundary validation, the text was presented in line-by-line form. The names of the four topic structure categories, along with their definitions (as in section 3.2.2) were given. Participants were asked to apply a category label to each line, as in figure 7.6. If they felt that more than one label could apply to a line, they were permitted to give more than one response; they could also respond “none” if they thought that none of the labels were appropriate. As before, the text was presented without sentence punctuation (except apostrophes) or capitalization.

The same procedures were followed as for the topic boundary validation (cf. section 7.3.1). None of the participants in this task had completed either of the topic boundary validation tasks.
7.4.2 Results

Eighteen participants responded to the topic structure categories task. Thirteen were female; ages ranged from 19-36. Eight reported that they spoke southern British English; other dialects reported were northern British English, American English, Scottish English, Irish English, and South African English.

A Fleiss’ kappa test was used to compare agreement among the raters for the four category labels. Agreement was higher for new Topics ($K = 0.317$) and Continuations ($K = 0.35$) than for Additions ($K = 0.137$) and Elaborations ($K = 0.149$); overall $K = 0.22$ ($p<0.01$). While still falling under the category of “fair” agreement (Landis and Koch 1977), this is lower agreement than was observed for the line-by-line topic data. However, the lowest observed agreements were in the categories of Addition and Elaboration. This is consistent with the observations in the prosodic data (chapters 3-6) in which these two categories were difficult to distinguish from one another on the basis of the prosodic variation. Therefore, a second analysis was conducted in which Additions and Elaborations were combined into one category. In this analysis, the agreement for the combined category increases to $K = 0.249$, and the overall $K = 0.294$ ($p<0.01$).

Observing the rater distributions in a multidimensional scaling analysis, the participants appear to roughly cluster into two groups in the analysis using all categories (figure 7.7a).

---

4Participants 1, 8 and 17 may be best ignored as outliers in this case. Participants 8 and 17 pattern very differently from the others in both versions of this analysis, although there is no immediately apparent reason
Table 7.1: Confusion matrix showing majority-voted categories (down) compared to original labels (across).

<table>
<thead>
<tr>
<th></th>
<th>New Topic</th>
<th>Addition</th>
<th>Elaboration</th>
<th>Continuation</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Topic</td>
<td>19</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td>Addition</td>
<td>23</td>
<td>54</td>
<td>10</td>
<td>3</td>
<td>90</td>
</tr>
<tr>
<td>Elaboration</td>
<td>2</td>
<td>8</td>
<td>40</td>
<td>1</td>
<td>51</td>
</tr>
<tr>
<td>Continuation</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>60</td>
<td>69</td>
</tr>
<tr>
<td>No majority</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>47</td>
<td>66</td>
<td>55</td>
<td>64</td>
<td>232</td>
</tr>
</tbody>
</table>

Although it is not possible to determine exactly what causes this pattern in the clustering, we may hypothesize that these clusters were due to similar interpretations (and therefore applications) of the Addition and Elaboration categories. When we observe the data with these two categories combined (figure 7.7b), the two clusters have disappeared, and the overall dimensional spread between the participants (ignoring the outliers) has decreased substantially. This, in combination with the Fleiss’ kappa results reported above, suggest that it is the Addition and Elaboration labels that participants found specifically problematic.

An overlap in performance between the Addition and Elaboration sets would be consistent with the phonetic results reported in chapters 3-6 as well as in the validated-topics version in section 7.3.3 above. That is, in almost all of the phonetic variables investigated, Additions and Elaborations were not distinguishable on the basis of the phonetic cues. Although it was proposed in the discussions of those analyses that it might require multiple cues to make a distinction between the two categories (e.g. speech rate combined with F0 fall range, section 5.2), it is equally or perhaps more likely that there was no observed phonetic difference because there was no actual category difference. On the other hand, the new Topic and Continuation categories, which enjoyed much higher agreement in the category labelling task, were also much more easily distinguished on the basis of their phonetic characteristics. Therefore the results of the category validation seem to be overall consistent with the phonetic results reported in chapters 3-6.

The set of categories which received a majority vote for each sentence was compared with the original category labels using Cohen’s kappa. Overall $K = 0.656$ ($p<0.01$), a “substantial” (Landis & Koch 1977) level of agreement between the majority vote from the
Figure 7.7: Plots of multidimensional scaling analysis for category data. (a) With all four category labels. (b) With Additions and Elaborations combined into one category.
raters and the original labels used. Table 7.1 shows a matrix comparing the majority votes for each sentence compared to the original labels. One very noticeable difference between the two categorizations is that the majority-vote categorization led to a much heavier use of the Addition category. In fact, items that had been categorized as new Topics in the original categorization tended to be categorized as Additions in the majority vote (if they were not voted as Topics). As can be seen in the confusion matrix, the items voted as new Topics were a subset of those originally labelled as new Topics; items that had been originally labelled as new Topics might be voted into a different category, but items that were originally labelled as a different category were not voted as new Topics.

After the redistribution of new Topics into (primarily) Additions, the second clear area of confusion is between Additions and Elaborations, although this confusion is apparently not as serious as might have been expected on the basis of the rater distributions. However, it is important to remember that this result represents the majority vote; that is, only 50% agreement for a rating was necessary, and therefore the level of disagreement among raters (cf. figure 7.7) was not taken into account in this analysis.

7.4.3 Discussion

The overall agreement for the category labels was lower than the agreement in the line-by-line topic boundary task, despite the fact that in the topic boundary task the labels were binary, which makes the statistical test weaker in that case than in the current case where there were four possible labels (plus the fifth option of “none”). This indicates that it was probably easier for participants to identify sections of text sharing the same topic than it was for them to apply the topic structure category labels. However, their application of the topic structure labels was still more consistent than would have been predicted by chance. This suggests that the category distinctions among the different types of utterance were still consistent, even if the category labels themselves were not always the best possible descriptions of the category distinctions. Furthermore, there was strong agreement between the majority vote labels and the original labelling scheme, indicating that for at least new Topics and Continuations, the labelling scheme adopted in the analysis in chapters 3-6 appears to have validity for the text.

for this upon a visual inspection of the data.
One notable area of confusion in the labelling task, the distinction between Additions and Elaborations, was the area of greatest overlap for many of the prosodic characteristics analyzed in the production experiment. For the F0 fall range (chapter 3), many speakers had no significant difference between the size of F0 falls in Additions and the size of F0 falls in Elaborations, despite the fact that the size of the F0 fall range was otherwise predictive for the topic structure categories, and despite a general trend in the expected direction (i.e. Additions having larger F0 falls than Elaborations). Similarly, for many speakers, the variation in local speech rate (chapter 5) was not significant between Additions and Elaborations. Although it was proposed that the two cues could work together to distinguish between Additions and Elaborations, the data supporting this distinction showed a trend rather than a strongly statistically significant result.

There are at least two possible interpretations of this finding. First, it is possible that there is no real distinction between what in this study were defined as Additions and Elaborations. That is, perhaps both could be subsumed under one category comprising additional information with a more distinctive status than Continuations, but not disjunctive enough from previous information to comprise a topic change. Alternatively, Topics, as comprising a large shift, and Continuations, as comprising highly-related information, could be separate categories, while what were identified as Additions and Elaborations could simply be an elsewhere case: utterances that are not distinctive either for their high degree of disjunction or their high degree of connectedness. This latter description could also account for any apparent variation in the phonetic characteristics in this middle category; there is no reason to suppose that an elsewhere case would require distinctive prosodic marking, even if very highly related or very highly disjunctive cases would.

The second possible interpretation is that there is a category distinction within this large group of utterances, but that the category distinction is different than the distinction adopted in the production study between Additions and Elaborations. Given that the text in question is a narrative, different types of categories might be better defined on the basis of episode shifts, as proposed by Chafe (1993; see discussion in section 7.1). The trend towards Additions and Elaborations showing stronger or weaker prosodic features (despite the lack of statistical significance) could be explained by a degree of overlap between these hypothetical episode-shift categories and the categories in the current study; that is, locations with
stronger episode shift characteristics might have been more likely to be labelled as Additions, although there was not an exclusive distribution.

Another way of explaining the current data is not a theoretical interpretation but rather a question of individual variation. It is clear from the higher-than-chance agreement among raters that the category definitions were understandable and applicable to the text at least to a degree. However, different readers of the text may simply have understood the structure of the text differently, assigning a different degree of importance to the same utterance, and therefore labelling it differently. Since different readers would approach the text with different interests and background knowledge, it is unsurprising that they would give greater importance to different portions of the text. This is likely to be the case regardless of whether the proposed categories are the most appropriate ones for the type of text.

The major area of disagreement between the original labels and the majority-vote labels, the redistribution of Topics into other categories, particularly Additions, was foreseeable on the basis of the topic-boundary tasks reported in section 7.3. The majority vote preference for these items to be ranked as Additions reinforces the ranked interpretation of these categories, consistent with a theoretical approach that combines a hierarchical and a category-based approach (see section 9.2.2 for further discussion).

7.5 General Discussion

The validation experiments reported in this chapter demonstrated independent evidence for the existence of the topic structure categories used in chapters 3-6 via a series of tasks in which untrained raters marked topic boundaries or topic structure categories in the text. The results of these rating tasks show a fair to substantial degree of agreement with the original labelling scheme, and a reanalysis of some of the phonetic data on the basis of the validated topic boundaries does not appear to contradict the results of the production experiment. The results of the category labelling task in particular also shed light on an area of confusion in the production experiment results by providing an explanation for the overlap of the phonetic characteristics of Additions and Elaborations.

One clear conclusion that may be drawn from the data reported in this chapter is that it is easier for participants to define topic groups than it is for them to categorize utterances
into topic structure categories. This appears to be true despite the fact that participants had widely differing ideas of how great a scope a “topic” should have, as well as the fact that they were able to apply the category labels with a degree of consistency greater than chance. Agreement on the location of topic boundaries apparently varied more on the reader’s interpretation of how specific a topic could be than on the things that formed the topics, as evidenced by the high levels of agreement about fewer topic boundaries, and the fact that locations that had not already been identified as topic boundaries were only very rarely rated as boundaries by the readers. Agreement about the identities of Continuations was at a similar level to agreement about Topic boundaries, suggesting that this category, representing little to no change in topic material (completing something previous, or having no episode-shift characteristics), was also fairly easily identified by the readers. Continuations also showed the highest level of agreement with the original topic-structure labels (90.3%).

The agreement on the topic boundary locations and the identity of Continuations suggests that the topic structure analysis applied in chapters 3-6 was valid for at least these two categories. The case for the existence of Additions and Elaborations is weaker, but since this is consistent with the results of the prosodic investigations, this finding in the validation study does not appear to be directly problematic for the overall structure of the production experiment. However, it is important to note that due to the distribution of the voted topic boundaries in the constructed text, only a partial validation of the production data was possible. This means that the results reported in chapters 3-6 may not be as definitive as the initial investigation may have suggested. However, a descriptive investigation of the data with the new topic structure analysis applied shows that the adjusted categories are not likely to lead to very different results. Furthermore, the result reported in chapter 6 about the distribution of aperiodicity appears relatively secure, since it was only in the context of Continuations that a different likelihood of aperiodicity was found.

In terms of applying the proposed discourse topic structure categories to a narrative, which had previously not been attempted, it appears that there is scope for such a theoretical application, although the categories themselves require some modification. Still, these validation experiments suggest that similar discourse-structure categories may be useful in describing the structure of a wide variety of types of texts. Future research in this area might return to Nakajima and Allen’s (1993) list of topic structure categories with their internal
constituents, to attempt to come up with a better characterization of utterances in a narrative. Similarly, these definitions may be coordinated with the idea of episode shifts as proposed by Chafe (1994). Although both of these theoretical modifications are beyond the scope of this thesis, the latter proposal will be briefly discussed in section 9.2.2.

Researchers in Conversation Analysis remind us of the importance of interpreting utterances in context, and point out that regardless of the rules we may try to apply or the frequency with which patterns occur, it is important to account for how an episode of speech occurs in a particular context at a particular time (cf. Local & Walker 2005). In other words, rules or categories might be better considered as generalizations than as being strictly definable and unbreakable. This means, among other things, that it is probably not reasonable to expect 100% agreement among readers (or speakers) about how a text is structured. Each person will bring their own background knowledge and interpretation to a text. However, there are still strong commonalities within groups as to how structures are to be interpreted in general; and the results of the validation reported here suggest that readers are broadly in agreement about what constitutes (or could constitute) a topic change.
Chapter 8

On-line Perception of Topic Structure Prosody

8.1 Introduction

The research reported in chapters 3-6 investigated a number of ways in which speakers of Standard Southern British English (SSBE) modify their production of prosody in parallel with variations in the topic structure of the discourse they are producing. These variations in prosody include changes in the F0 contour, the speech rate, and the phonation quality of the utterances in question. We saw that the signalling system for topic is very rich and contains a variety of cues which are potentially recoverable by a listener over the course of the whole utterance. However, the degree to which listeners recover and make use of these cues in processing discourse, and particularly discourse topic structure, is not yet clear. While some previous studies have shown that listeners are able to use prosodic information to identify topic boundaries, it is not known whether listeners make use of this information outside of meta-linguistic tasks. The purpose of the following experiment is twofold. First, it aims to test whether the previously-identified prosodic topic signalling are recovered and used in the online processing of SSBE spoken discourse; second, it develops a novel methodology in order to conduct this kind of investigation.
8.1.1 Literature review

8.1.1.1 Metalinguistic studies

Listening studies have been used to identify prosodic characteristics of topic structure organization. Swerts and Geluykens (1993), found that variations in F0 and, to an extent, pause length affected listeners’ judgments about when a discourse unit in spontaneous speech had ended. Swerts and Geluykens’ perceptual task was metalinguistic; participants were asked to give a verbal indication of when they thought a “major discourse unit” (1993: 194) had ended. In this case, the loss of F0 information had the most detrimental effect on participants’ ability to identify the boundaries, although they were still above chance level on the task, suggesting that F0 was not the only source of this information, consistent with the production study findings reported in chapters 3-6.

Swerts (1997) examined the perception of discourse structure boundaries by participants who both read and listened to a text, and the perception of boundaries by participants who only read the text. Swerts’ goal was to identify the hierarchical structure of a discourse by using group segmentation data, with the assumption that

“...there will be less disagreement about stronger breaks [in discourse structure] since they present clearer transitions in the flow of information. It can thus be expected that more people will feel inclined to mark these as a boundary. Therefore, instead of taking the variance between labelers as a disadvantage, one can rather exploit it to specify hierarchically different discourse boundaries.” (1997: 515)

He found that people who listened to a text were more consistent in their placement of discourse boundaries than people who only read the text. Swerts then used these results to identify some prosodic correlates of the discourse structure (cf. section 2.3.3). At the moment, however, the interest in this study is primarily methodological. Swerts’ assumption, borne out by his prosodic findings, was that listeners are sensitive to prosody and are able to use it in identifying discourse organization.

A similar study was conducted by Grosz & Hirschberg (1992), specifically testing the topic-structure predictions made by the theory presented in Grosz & Sidner (1986). Grosz and Hirschberg were interested in identifying prosodic cues to the beginnings and ends of
CHAPTER 8. ON-LINE PERCEPTION OF TOPIC STRUCTURE PROSODY

discourse segments. Their study had seven participants, four labelling from a written text and three labelling from both the written text and a spoken version. Unlike Swerts (1997), Grosz and Hirschberg report that identification of discourse segment beginnings and endings was no more consistent among the listeners than among all the labellers together, and in fact, their labellers who had only the written text available may actually have been more consistent than the group who had the spoken text as well. However, they note indirectly that this may be a result of the small number of participants; one of the labellers in the spoken-text group was very inconsistent with the other two, and this could have skewed their results in this case. Despite the relatively small number of labellers and the inconsistency in this case, Grosz and Hirschberg still found very good labelling agreement, particularly for the beginnings of discourse segments, and they identified a number of intonational correlates of the discourse structure on this basis. They further argued that using these intonational correlates, it should be possible to predict the general pattern of segmentation that listeners will use, thus providing a precursor to the study reported below.

These studies give evidence that listeners are able to use prosodic information to impose an organization on a spoken text, and that prosodic information is relevant possibly even beyond the lexical content of the text. However, they do not show that listeners actively make use of this knowledge in normal speech processing, i.e. when they are not being specifically asked to find discourse boundaries. It may be the case that the task itself is unnatural, and that listeners are simply making the best of the situation. However, both their systematic performance as well as the systematic phonetic effects related to topic structure in this thesis and in other research (cf. section 2.3.3) suggest that the prosodic information related to topic structure could in fact be useful for listeners during the comprehension process. The long-term availability of prosodic topic-structure cues and therefore their potentially easy recoverability add credence to this.

Similarly metalinguistic studies have been performed on other, sentence-level aspects of prosodic interpretation, and may be useful to consider in the context of developing methodologies for studying the processing of topic structure. Bock and Mazzella (1983) found that when prosodic accent placement matched the information structure of a sentence, participants were quicker to say that they had understood the target sentence than when the accentuation was inappropriate. Birch et al. (2000) found that in some cases, items which
CHAPTER 8. ON-LINE PERCEPTION OF TOPIC STRUCTURE PROSODY

were syntactically focused were more easily recognized afterward than items which had not been, although there was some interaction with the syntactic position from which the item might have been displaced. Schafer et al. (2000), dissatisfied with previous work focusing on read sentences which they found unnatural, studied disambiguation in relatively spontaneous discourses. After conducting a production study, they used recordings from this study to see whether listeners were able to correctly categorize sentences having early- or late-closure around a target word according to the prosodic cues. In addition to using sentences with the appropriate prosodic cues, they also cross-spliced some of the recordings, in order to have items in which the prosody was ambiguous or even created a conflict. They found that appropriate prosody aided listeners greatly in making correct categorizations, but interestingly, ambiguous prosody did not necessarily cause listeners problems, though conflicting prosody did to an extent. These results can be interpreted to mean that listeners, though sensitive to prosody, also attend to other information, in this case the lexical information, before making a decision about the interpretation of the sentence. It should be noted that this task, like the labelling tasks discussed above, was metalinguistic; Schafer et al. (2000) presented participants with a forced choice between two potential sentence completions from which they had to choose after listening to the first half of the sentence. In other words, although the results are suggestive, the task was not a very natural one.

8.1.1.2 On-line processing of prosody

A few studies have investigated the role of prosody in on-line processing at the level of the sentence, although these studies often have to do with sentence disambiguation and syntactic processing. Prosody has also been shown to be a reliable cue for the on-line resolution of the kinds of syntactic ambiguities discussed above. For instance, Frazier et al. (2006) review a series of results showing that listeners are sensitive to pause length in determining which phrases fall within the scope of a modifier. They point out specifically that it is not only the presence of a pause or other prosodic boundary that determines how modifier scope is interpreted, but also the size of the boundary in context; instances of interpretation of high attachment decreased as the size of an early boundary increased relative to the size of a late boundary. This means that listeners had to not only identify that there was a boundary, but maintain during processing some kind of information about the size of the boundary. Oth-
erwise they would not be able to compare it later to a second boundary in order to make a
decision about how to interpret the modifier attachment. Frazier et al. interpret these find-
ings as indicating a central role for prosody in processing in terms of solving the “binding
problem”; that is,

“...the problem of how to identify the correspondence between the different ap-
pearances of, say, a past tense or a possessive in different representations in
which it occupies very different locations. If each (level of) linguistic repre-
sentation is indexed to the prosodic representation, then it would explain how a
single unit (e.g. the past tense) can be identified across representational types”

It is difficult to test listeners’ comprehension of the topic structure of a discourse on the
basis of prosody in part because there are so many other linguistic factors at hand which can
interfere. First, we know that the availability of prosodic information is not obligatory in
order for listeners to be able to comprehend a discourse; although punctuation is commonly
seen as a substitute for prosody in written speech, in fact the prosodic system of a spoken
language is far richer than the punctuation system of any written language of which the
author is aware. Despite this, literate persons are able to glean meaning and structure from
written texts lacking the information provided by prosody. This means that much structural
information is either directly available or else inferable from the lexical-semantic content
of the text. Things like the scope of pronoun/anaphor reference may provide information
as to how a text is to be chunked (although this could be a circular definition, since one of
the expected functions of topic grouping is to delimit the space within which an anaphor’s
antecedent may be found, cf. Grosz & Sidner 1986). From the point of view of an ex-
perimental test for the role of prosody in marking topic structure, this means that any kind
of linguistic stimuli could be prone to an inherent bias towards topic structure. In other
words, they could lend themselves to a specific topic-structure interpretation without regard
to prosodic cues.

One possible way to test the contribution of prosody to the message would be to use
some kind of incomplete or degraded signal; for example, speech in noise, sine wave speech,
low band-pass filtered speech, or use of pitch contours only. However, this could also be
problematic in that it would not necessarily address the question of how these prosodic manipulations are used in speech perception. Repp (1982) gave an early review of research suggesting that speech perception is a special mode of perception differing from other auditory perception, including evidence for trading relations and context effects, as well as the integration between auditory and visual input, which cannot be accounted for by theories of simple auditory perception alone. More recently, Moore (2003) covers this ground in greater detail. He points out that

“...listening in a specific perceptual mode is not unique to speech perception...

We may listen in an analytic mode, hearing the pitches of one or more individual partials, or we may listen in a synthetic mode, hearing a single pitch corresponding to the fundamental frequency... However, the speech mode is unusual in that it operates for an entire class of highly complex and varied acoustic signals, whose main common feature is that they were produced by a human vocal tract” (2003: 315).

Another question which is yet to be addressed is the extent to which topic affects information processing as a whole. It is possible that topic is used by listeners primarily in a local way; for example, to constrain the scope of pronoun/anaphor reference. A topic shift could force a boundary over which it might be more difficult to maintain such a reference (Grosz & Sidner 1986, Blakemore 2002). The extent to which topic structure might have an influence on processing beyond the creation of boundaries has not yet been explored. It is possible that topic’s main role in processing is to make processing easier by grouping like things together; this might be a prediction of a hierarchical model. Grosz & Sidner (1986) approach discourse structure as being relevant to maintaining versus clearing pieces of information on a stack. The location of information on a stack determines its “accessibility” at any given time; however, the accessibility of certain referents may depend on how the referents are presented in the utterance rather than a simple overall discourse structure (Baumann & Grice 2005). The structure itself is primarily relevant in that things are grouped or not grouped by whether or not they remain in a stack at the same time (Grosz & Sidner 1986; cf. section 2.3.2.1). However, if we take a categorical approach to topic, in which the semantics of the utterances themselves matter for the topic structure, topic
might be considered to play a role in the semantic processing, as in Chafe’s (1994) notion of maintaining a set of semiactive information.

The word-based nature of many psycholinguistic paradigms is also problematic in the face of the question of the global versus local relevance of discourse topic structure in utterance and discourse processing. One common methodology for testing how people comprehend sentences is a self-paced reading task, wherein participants read a sentence one word at a time, pressing a button to cause the next word to appear once they have finished reading the current word. This task allows researchers to identify points in a sentence where processing may be slower, or potentially to gain more accurate reading times for whole sentences than could be obtained simply by asking participants to respond when they have finished reading/comprehending (as in Bock & Mazzella 1983). However, in the case of studying discourse processing, this is problematic because there are too many other things that can influence word reading. A less-common word, or a word with a large neighborhood, may be read more slowly because of its lexical nature, without any reference to the discourse structure (c.f. Harley 2001). Furthermore, it is difficult to tell where readers are focusing their attention, or to ensure that they are performing this task as requested; they may move quickly through several words on the screen and then process them later (Ferreira & Henderson 1990). Therefore this kind of word-based paradigm presents difficulties for interpretation when it comes to studies of utterance or discourse comprehension.

8.2 Methodology Development and Pilot Studies

In order to overcome the difficulties described above, a methodology involving three tasks was developed to test the degree to which listeners use topic-structure prosody in on-line processing. These tasks were (a) a pause detection task, (b) an acceptability rating, and (c) a memory task. Although this methodology was somewhat complex, each task contributed information about a separate hypothesis regarding the process of perception.

8.2.1 Simultaneous Task: Pause Detection

The first question to be addressed was whether listeners are differentially sensitive to the cues previously identified as topic-change versus topic-hold prosody at the moment that they
hear it, regardless of whether they actually use this information to immediately identify topic change. In other words, we want to know whether it is possible to observe different behavior solely on the basis of different prosody. This is important because it addresses listeners’ sensitivity to cues, which is a prerequisite for recoverability and potential subsequent use of these cues.

One way of determining whether sensitivity to different cues varies is to test listeners’ performance on another task simultaneous to listening, to see if there is variable performance on that task. Decreases in performance on a measurable task can be interpreted as evidence for an increased cognitive load caused by performance on the unmeasured task (in this case, listening to the prosody of the utterance). However, it is important to test task performance at the right time, particularly given that phonetic information is considered to be prone to rapid decay; it may no longer be available to listeners after an interval as short as 250ms (Andruski et al. 1994). Therefore, if we wish to measure listeners’ sensitivity to prosody specifically, a companion task must occur within this time frame in order to be certain of capturing an effect, assuming it is present.

8.2.1.1 Probe-dot detection

The initial pilot for the study used a probe-dot detection task as the simultaneous activity (cf. MacLeod, Mathews & Tata 1986). A colored dot, either red or blue, was displayed at the center of the screen at the end of the targeted pitch contour (or at other locations in the auditory sentences in filler items). The color of the dot was randomized across target and filler items; all items had a dot presented at some point. Participants had to press a button on the response box indicating the color of the dot that they saw. However, no effect of the intonation contour on this task was observed during the pilot study. One possible explanation for this was that the task was too easy, either in general, or due to the fact that it was completely non-linguistic, and that therefore any cognitive burden placed on the perceptual system by the intonational variation was not enough to affect performance on the dot detection task. A search for a more linguistically-linked task led to the selection of the pause detection paradigm for the main study.
CHAPTER 8. ON-LINE PERCEPTION OF TOPIC STRUCTURE PROSODY

8.2.1.2 Pause detection

One kind of task which can be related to a specific point in time is a pause detection task (cf. Mattys & Clark 2002, Mattys et al. 2005). In a pause detection task, participants listen to speech and are asked to respond, e.g. by pressing a button, when they hear an unnatural pause. Mattys & Clark show that pause detection is useful as an “on-line measure of global lexical activity during speech processing” (2002: 355); that is, if there is a large amount of activity, there will be longer pause detection latencies than if there is less activity. Pause detection has been shown to be sensitive to other non-lexical aspects of speech as well, including prosody; Duez (1985) reported that French speakers’ difficulty in identifying pauses was the same when listening to only the prosodic contour as when they listened to the full speech signal. Pause detection is also by its nature fixed in time; the pause can be placed anywhere in an utterance, even in the middle of a word if necessary.

In the current experiment, the pause detection task was used in place of the dot detection task to test whether or not listeners were sensitive to the difference in intonation contours between topic-change and topic-hold contexts. Listeners heard one or two context-setting sentences, followed by a third utterance which had had the intonation contour manipulated to signal either a topic change or a topic hold. In the pilot experiment, the different intonation contours were produced by the speaker, as opposed to the main experiment, in which one version of each sentence was resynthesized in Praat to modify the F0 contour appropriately. The intonation-manipulated utterance also had an unnatural pause of 200ms duration synthesized into it; listeners were asked to press any button on the response box when they heard the unnatural pause. The value of 200ms was taken from Mattys et al. 2005, and was preferable to the 150ms pause used by Mattys & Clark 2002 due to the fact that in the current experiment the pause occurred between words, although not in a location that would be predicted by the phrasing, and therefore a longer pause might be necessary in order to give the impression of unnaturalness necessary for the task. No pilot experiment was conducted with the pause detection task, but two trained phoneticians listened to earlier versions of the target stimuli and commented on the degree of naturalness of the pauses.
8.2.2 Sentence Acceptability Rating

One potential difficulty with the pause detection task is that it might lead listeners not to pay attention to the semantic content of the utterances at all, but rather to concentrate on rhythmic cues or use some other strategy in order to make their decisions about the pauses. Furthermore, the pause detection task only tests listeners’ sensitivity to variations in prosody without explicitly testing whether the cues are used in processing topic structure. Therefore, at least one additional task is necessary to make sure that listeners attempt to interpret the semantic information conveyed by the prosodic variation. For this purpose an acceptability rating task was employed. Subjects had to decide whether a final sentence (hereafter the “follow-up sentence”) which they read on the screen was appropriate in the context of the sentences that they had previously heard. They therefore needed to attend to the previous content of the context sentences in order to be able to make a decision about the appropriateness of the final sentence. Each set of context sentences plus an intonation-manipulated sentence had two follow-up sentences associated with it. One follow-up sentence would be more appropriate to a context in which the previous utterance had contained a topic change (and that utterance was therefore the new topic), and the other follow-up sentence would be more appropriate to a context in which the previous utterance had contained a topic hold (and that therefore the topic of the first context sentence(s) was still in effect). In some cases, the follow-up sentence presented would match the prosodic cue given in the previous utterance (e.g. topic-hold prosody with a topic-hold follow-up sentence), while in other cases the follow-up sentence would not be consistent with the prosodic cue (e.g. topic-hold prosody with a topic-change follow-up sentence). Consistent pairs should be more acceptable than inconsistent pairs.

8.2.2.1 Pre-testing of follow-up sentences’ acceptability

It was important to ensure that the follow-up sentences were in fact acceptable in context. Although the experimental comparisons were made within sentences (i.e., responses to the same follow-up sentence in different prosodic conditions were compared), it would still be problematic to have follow-up sentences which were not generally judged as acceptable in the context.
An internet-based pilot study assessed 32 potential items and their follow-up sentences. The context sentences plus both follow-up sentences were presented in written form, and participants were asked to indicate which follow-up sentence was more appropriate in the given context. No definition of “appropriateness” was given. There was a response option for “equally plausible” in addition to the options for the two follow-up sentences, and a text box allowing participants to add any comments they might have about each specific item overall.

Nineteen native speakers of various varieties of English participated in the pilot study via the internet. All their names were entered into a draw for a £20 gift certificate for Amazon.co.uk. No time limit was placed on the pilot study.

When the results were collected, each response was coded as 1 = first sentence (topic hold) was more plausible, 2 = both sentences were equally plausible, or 3 = second sentence (topic change) was more plausible. All scores for each item were averaged, giving a value between 1 and 3 for each of the 32 items. A mean value close to 2 suggested that both follow-up sentences to the item were considered equally plausible.\(^1\) A mean value close to 1 or 3 suggested a bias towards one of the follow-up sentences.

T-tests were performed on each of the data sets to compare the mean value to 2 and test for significant differences (see table 8.1). 17 of the item means were found to not differ significantly from 2; all of these items were used for the final version of the experiment. In order to make up the remaining total of 24 target items for the final experiment, the other pre-tested items were modified with reference to the comments made by the participants in the pre-test. In a large number of cases, items had been marked as implausible due to participants’ dislike of the way a sentence was phrased, or due to a term used in the context sentences, rather than due to a semantic problem. If several commenters agreed on a specific issue, that change was made and the item was adopted for the final experiment.

\(^1\)The mean value does not take into account the amount of variability in the responses; that is, a mean value of 2 could indicate that all participants thought both sentences were equally plausible or that half of the participants thought one sentence was more plausible and half thought the other. To an extent this variability is dealt with in the t-test; however, since the goal was to have overall agreement across a population, a mean value of around 2 is still a good indicator regardless of the amount of variability. The mean score also does not take into account the fact that participants may have considered both follow-up sentences equally unlikely. One of the goals of the text box was to allow participants to indicate if they thought something like this was the case; in practice, however, no one made any comments to this effect, so it seems that this possibility can be discarded.
Table 8.1: Mean response value for pre-tested items; df=34 in all cases. C = topic hold (i.e. Continuation); T = topic change (i.e. new Topic).

<table>
<thead>
<tr>
<th>Item</th>
<th>Mean</th>
<th>SD</th>
<th>T-value</th>
<th>p&lt;0.05</th>
<th>Bias direction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.000</td>
<td>0.000</td>
<td>(n.a.)</td>
<td>Y (n.a.)</td>
<td>C</td>
</tr>
<tr>
<td>2</td>
<td>1.882</td>
<td>0.332</td>
<td>1.5074</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2.059</td>
<td>0.429</td>
<td>0.5838</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1.706</td>
<td>0.772</td>
<td>1.6163</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2.176</td>
<td>0.529</td>
<td>1.4126</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1.235</td>
<td>0.437</td>
<td>7.423</td>
<td>Y</td>
<td>C</td>
</tr>
<tr>
<td>7</td>
<td>1.706</td>
<td>0.588</td>
<td>2.1506</td>
<td>Y</td>
<td>C</td>
</tr>
<tr>
<td>8</td>
<td>1.294</td>
<td>0.470</td>
<td>6.5115</td>
<td>Y</td>
<td>C</td>
</tr>
<tr>
<td>9</td>
<td>1.765</td>
<td>0.562</td>
<td>1.7804</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>2.471</td>
<td>0.624</td>
<td>3.223</td>
<td>Y</td>
<td>T</td>
</tr>
<tr>
<td>11</td>
<td>2.353</td>
<td>0.606</td>
<td>2.4714</td>
<td>Y</td>
<td>T</td>
</tr>
<tr>
<td>12</td>
<td>2.000</td>
<td>0.612</td>
<td>0</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>2.235</td>
<td>0.437</td>
<td>2.2815</td>
<td>Y</td>
<td>T</td>
</tr>
<tr>
<td>14</td>
<td>1.824</td>
<td>0.529</td>
<td>1.4142</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>1.529</td>
<td>0.514</td>
<td>3.8877</td>
<td>Y</td>
<td>C</td>
</tr>
<tr>
<td>16</td>
<td>1.353</td>
<td>0.702</td>
<td>3.9102</td>
<td>Y</td>
<td>C</td>
</tr>
<tr>
<td>17</td>
<td>2.118</td>
<td>0.485</td>
<td>1.0322</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>2.353</td>
<td>0.702</td>
<td>2.1334</td>
<td>Y</td>
<td>T</td>
</tr>
<tr>
<td>19</td>
<td>2.176</td>
<td>0.636</td>
<td>1.1741</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>1.882</td>
<td>0.781</td>
<td>0.641</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>1.176</td>
<td>0.393</td>
<td>8.8955</td>
<td>Y</td>
<td>C</td>
</tr>
<tr>
<td>22</td>
<td>1.235</td>
<td>0.437</td>
<td>7.427</td>
<td>Y</td>
<td>C</td>
</tr>
<tr>
<td>23</td>
<td>1.824</td>
<td>0.529</td>
<td>1.4115</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>1.471</td>
<td>0.717</td>
<td>3.1302</td>
<td>Y</td>
<td>C</td>
</tr>
<tr>
<td>25</td>
<td>2.412</td>
<td>0.870</td>
<td>2.0092</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>1.882</td>
<td>0.781</td>
<td>0.641</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>1.588</td>
<td>0.618</td>
<td>2.8284</td>
<td>Y</td>
<td>C</td>
</tr>
<tr>
<td>28</td>
<td>1.824</td>
<td>0.636</td>
<td>1.1741</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>1.941</td>
<td>0.556</td>
<td>0.4502</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>1.824</td>
<td>0.728</td>
<td>1.0257</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>1.706</td>
<td>0.772</td>
<td>1.6157</td>
<td>N</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>2.412</td>
<td>0.795</td>
<td>2.1987</td>
<td>Y</td>
<td>T</td>
</tr>
</tbody>
</table>
8.2.2.2 Sentence acceptability pilot

The sentence acceptability rating task was piloted concurrently with the dot detection test (cf. section 8.2.1.1). Participants were presented with one or two context sentences followed by a target sentence with either topic-change or topic-hold prosody. In the pilot, the sentences were produced differentially by the speaker rather than being manipulated in a consistent way, so a variety of topic-change cues were presumably available to listeners, including high pitch, a longer pause, and possibly variation in the pitch accent structure; these were not analyzed. The speaker had been asked to read the topic-change items “as though they were starting a new paragraph with that sentence”; the experimenter was present during the recording session to request repeats if necessary.

Participants were asked to rate on a scale of 1 to 5 how likely it was that this was the next thing the speaker was going to say, with 1 being very unlikely, and 5 being very likely. There was no time limit on their responses but they were asked to respond as quickly as possible.

It became apparent very quickly that the task of ranking the follow-up sentence on a scale of 1 to 5 was too complex. Most participants used only two or three levels of ranking rather than using the whole scale (see table 8.2). Also, the responses fell overwhelmingly into the “likely” or “very likely” categories, and the responses of “unlikely” seemed to be at random, perhaps reflecting individual differences, attentional differences, or simply a feeling that too many sentences in a row had seemed likely. Given the apparent difficulty that participants had with using the 5-level scale for the rating task, the decision was taken to use a binary distinction between likely and unlikely in the final experiment, in order to create less task-based confusion.

The pilot showed almost no effect of the prosodic match or mismatch on the acceptability response to the following sentences, so it might seem preferable to simply discard this task. However, in examining the participants’ response times to “acceptable” utterances, a different response latency became apparent in the match and mismatch cases. That is, when the follow-up sentence did not match the prosodic characteristics of the prosody-manipulated sentence, response latencies were longer than when the prosodic characteristics matched (see figure 8.1). This suggests that even though the semantic content, which was
Table 8.2: Sentence rating responses in pilot study (statistics not reported due to small number of participants).

<table>
<thead>
<tr>
<th>Pilot1</th>
<th>Pilot2</th>
<th>Pilot3</th>
<th>Pilot4</th>
<th>Pilot5</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>0</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>0</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>0</td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

Figure 8.1: Response times for “acceptable” ratings by match or mismatch with prosody in pilot study. (Statistics not reported due to small number of participants.)
acceptable in either case, was overruling any effect of the prosodic mismatch, there was still a point in processing at which the prosodic mismatch caused a problem for listeners which had to be overcome. Furthermore, the sentence rating task required participants to interpret the utterances that they heard, which was necessary for the design of the memory task. Therefore the acceptability rating task needed to be maintained.

8.2.3 Memory Task

The weakness of the sentence acceptability rating task is that it is metalinguistic; stating whether a sentence is appropriate in context is not a very normal language task (except perhaps in cases of misunderstandings in interaction). Therefore it was desirable to have another task that could further investigate the way in which the topic-change intonation was being interpreted by the listeners.

A number of studies on the role of focus (both prosodic and syntactic) in affecting people’s memory have shown that items that have been appropriately focused in context are better remembered than items that are either not focused or inappropriately focused. Birch and Garnsey (1995) showed that prosodic focus led to improved memory for concepts. They believe that the role of focus is to increase the strength and particularly the specificity of the memory trace associated with a word, in order to make integrative processing across sentences (e.g. of anaphor reference) easier for listeners. Birch et al. (2000) also found a similar effect for syntactic focusing constructions, suggesting that syntactic focus and prosodic focus play similar roles in speech perception. Birch et al. furthermore make the observation that the memory effect is only relevant after working memory had been cleared; before this, focused and neutral items seem to be equally accessible.

This could be seen as a parallel with Grosz & Sidner’s (1986) idea of “clearing a stack” of information with the shift of discourse topic; that is, when a topic change occurs, the “stack” of information which is held in working memory is no longer needed and can be discarded. Following up on this research, Almor & Eimas (2008) found that (syntactic) focus interacts with a repetition effect for nouns in speech; that is, repetitions of full NPs when a pronoun or anaphor would do instead can be rendered more or less appropriate on the basis of the focus structure. Almor & Eimas did not find this effect in a lexical decision task, but rather when they asked subjects to recall what NP they had heard. They argue that
Table 8.3: Example stimuli for perception experiment (full stimulus list provided in Appendix II). Italics indicate the location of the prosodic manipulation.

<table>
<thead>
<tr>
<th>Stimulus type</th>
<th>Context sentences plus intonation-manipulated sentence (heard)</th>
<th>Read sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target (topic change)</td>
<td>She arrived at the hotel. She found her room. She \textit{started to} [pause] unpack her bags.</td>
<td>She wanted to hang her clothes in the closet.</td>
</tr>
<tr>
<td>Target (topic hold)</td>
<td>She arrived at the hotel. She found her room. She \textit{started to} [pause] unpack her bags.</td>
<td>She needed to leave soon to get to her meeting.</td>
</tr>
<tr>
<td>Filler (alternative pause)</td>
<td>There was a surprise [pause] party. Everyone was hidden. The guest of honor walked through the door.</td>
<td>Everyone jumped out and yelled, “Surprise!”</td>
</tr>
<tr>
<td>Filler (semantic mismatch)</td>
<td>They skied all day. They were really [pause] tired. They slept well that night.</td>
<td>In the morning the post was delivered.</td>
</tr>
</tbody>
</table>

“...recall accuracy reflects the outcome of the integration of anaphors into the representation of the discourse, which shows the effects of the balance between the processing cost associated with maintaining distinct active representations that are overlapping, and discourse function” (2008: 221).

The concept that improved memory can result from better integration of information is the basis for the memory task in this experiment. If the follow-up sentence that a listener reads matches the intonation of the intonation-manipulated sentence, and if this is indeed a salient cue to topic structure which the listener is expecting to hear and use, then information (the content of the sentence) should be better integrated and therefore better remembered in this case than in a case where there is a mismatch between the intonation and the follow-up sentence.

8.3 Full Experiment Methodology

8.3.1 Materials

The stimuli were 24 intonation-manipulated sentences, each preceded by one or two context sentences, and followed by a read sentence, as shown in table 8.3. The experiment specifi-
cally tested the role of the F0 fall range in signalling topic change to listeners. F0 fall range was chosen as the strongest cue to topic structure identified in the production experiment (cf. chapter 3), and one which had not been previously used in any perception study. As in the production experiment, there is a certain amount of confound in the case of new topic utterances: F0 fall range was found to vary alongside topic structure, but new utterance groups (and therefore new topics) were also signalled by a high F0 peak in relation to the speaker’s baseline. In order to test whether there was an independent contribution of the F0 fall range in signalling topic structure, two sets of stimuli were created. Both had F0 falls of 6 semitones (topic change) and 4 semitones (topic hold), as per the mean fall sizes observed in the production study (section 3.3.2). However, as shown in figure 8.2, one set, the “high” set, had both contours ending at the same pitch level, but the topic-change contour peak was higher than the topic-hold contour. In the other set, the “low” set, both contours had their peaks at the same pitch level, but the topic-change contour ended lower than the topic-hold contour. The difference between the “high” and “low” sets was intended to test whether participants would respond differently to a higher peak versus a larger pitch movement without a higher peak; however, in the data reported below, participants who heard the “high” set showed the same response patterns as participants who heard the “low” set. Therefore, the two topic-change intonation sets are treated as one set in the analysis.

All versions of the pitch contour were resynthesized from the same recording. The
contours were automatically stylized to a resolution of 2 semitones by Praat, and in some cases smoothed further by hand in order to remove spurious pitch points such as octave errors. The location of the pitch peak in the target word was not changed, but the height of that point and the low valley following it were both standardized to the pitch levels given above. Any other particularly high or low pitch peaks in the context utterances were also modified at this time in order to ensure that the context sentences’ prosody did not license another topic change, although this was only necessary in the case of two of the utterances. The pause was placed directly after the end of the F0 contour in question (i.e., after the low turning point of an F0 fall), and therefore was medial to an intonation phrase, although it was sometimes at a location which would have been appropriate as an intermediate phrase boundary. A 200ms pause was inserted between two words (cf. Mattys et al. 2005 and discussion in section 8.2.1.2), in the same place in both the topic-change and topic-hold versions of the stimuli.

A 2x2 Latin square design ensured that all items appeared in all possible match and mismatch conditions, but each subject saw a given item only once, and with only one possible follow-up. The 24 experimental stimuli were presented in a randomized order among another 26 filler items. The filler items were constructed in a similar format to the target items, i.e. two or three sentences were presented auditorily followed by another written sentence. Of these items, 16 had 200ms pauses in locations that corresponded to the context utterances in the target items. The remaining 10 had no synthesized pauses. The pause locations and/or absence in the filler items were designed to prevent listeners from predicting the location of the pause in the target items, since in those cases it always fell early in the intonation-manipulated sentence (immediately after the first pitch accent). Also, some of the follow-up sentences for the filler items were deliberately nonsensical or unrelated to the auditory sentences. This was done in order to help participants feel comfortable responding that some of the follow-up sentences were inappropriate for their context, since all of the target follow-up sentences should have been appropriate given the right intonational context, and the sentence acceptability task might have given spurious results if participants did not feel able to judge some sentences as obviously inappropriate.

The memory task used a list of the 24 target sentences, along with 8 filler sentences that had not been heard in the first part of the experiment. These sentences were presented on
paper, in the same order for each participant (which should not affect the results since the auditory stimuli were presented in a different randomized order for each participant).

8.3.2 Procedure

The integrated version of all the experimental tasks was as shown in figure 8.3.

Participants heard two context sentences and an intonation-manipulated sentence, during which time they performed the pause detection task. They were asked to identify unnatural-sounding pauses by pressing a button on a response box as they listened to the stimuli. They were given four practice items, which were supervised by the researcher to make sure that they were able to correctly identify the unnatural pauses. Most subjects were able to identify all the unnatural pauses on the first attempt; those who could not were taken through the practice items again, though most said that they had simply not been sure about when to respond on the first attempt.

Each participant, after hearing the context sentences and the intonation-manipulated sentence, saw only one of the two possible follow-up sentences and had to perform the sentence acceptability task; that is, they responded to “whether it seemed likely that this was the next thing the speaker was going to say.” In some cases, participants were presented with follow-up sentences that matched their prosodic context; e.g. they heard the intonation-modified item with topic-change intonation, and then read the follow-up sentence that was appropriate for a topic-change context. In other cases, participants were presented with a follow-up sentence that did not match the prosodic context; e.g. they would hear the intonation-modified item with topic-change intonation but read the follow-up sentence that was appropriate for the topic-hold condition.

The order of presentation was as follows: the spoken sentences were presented first, with a blank screen. 200ms after the end of the spoken utterances, the written follow-up sentence appeared on the screen. The follow-up sentence was visible for 2000ms in order to allow sufficient time for participants to read the sentence and respond, and then the screen was cleared with a 1000ms pause before the beginning of the next item. Responses were either the far right (acceptable) or far left (unacceptable) buttons on the response box; participants were asked to use both hands, although they could choose which fingers were most comfortable for them. Participants were asked to respond as quickly as possible while
<table>
<thead>
<tr>
<th>Participant experience</th>
<th>Possible conditions</th>
<th>Question tested</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Hear context sentences</td>
<td>(n/a)</td>
<td>(n/a)</td>
</tr>
<tr>
<td><em>The Tube station was closed. He had to get off at the next stop.</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Hear target sentence</td>
<td>Topic change or topic hold prosody (larger or smaller pitch movement)</td>
<td>Are listeners sensitive to different prosody?</td>
</tr>
<tr>
<td><em>He got lost trying [pause] to walk back.</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Do pause detection task concurrently</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Read follow-up sentence</td>
<td>Consistent with either topic change or topic hold having occurred (therefore either consistent or contrasting with prosody heard in previous stage)</td>
<td>Do listeners process prosody on-line, affecting immediate interpretation of topic change?</td>
</tr>
<tr>
<td><em>He eventually arrived at the party.</em> OR <em>He ended up in a neighbourhood he'd never seen before.</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rate acceptability of follow-up sentence after reading</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[10-minute break]
Read target sentences
Do memory task (write follow-up sentence)
Sentence was either consistent or inconsistent with topic change having occurred
Is memory affected by better integration of information? ("Better integration" defined as consistent prosody and lexical content)
still remaining accurate in their judgments. Responses and response times were recorded; any response which took longer than the 2000ms that the sentence was on the screen was not collected, but participants normally did not have difficulty responding within this time frame.

After the pause detection and sentence acceptability portion of the experiment had been completed, which took just over 10 minutes, participants were given a 5-minute break during which they could get up and move around if they chose. After the break, they were given the memory task, which they could take as much time to complete as they needed. For this task, the participants, who were not informed prior to the start of this portion of the experiment that there would be a memory task involved, were given a list of sentences, some of which had occurred in the first portion and some of which had not. They were asked to identify which sentences had been present, and to write down the following sentence or any part of it that they could recall.

The study was conducted in a quiet room at the Research Centre for English and Applied Linguistics, University of Cambridge. The first half of the experiment, comprising the pause detection and sentence acceptability rating tasks, was conducted using E-Prime (Psychology Software Tools 2007), which presented the items in a randomized order and collected all the responses. The auditory stimuli were presented via a pair of Philips SBC HP400 headphones; participants were able to adjust the volume at their discretion. The visual stimuli were presented as a single line of black text on a white background on a 17-inch monitor. Subjects responded to the stimuli on a Psychology Software Tools 200A 5-button response box. The second half of the experiment (the memory task) was conducted with pen and paper.

8.3.3 Participants

Forty native speakers of a variety of dialects of English, all of whom had been resident in Cambridge for at least a year, took part in the study. Ages ranged from 18-26; 24 were female. Most were students at the University of Cambridge. None reported any hearing or language impairments; this was based on self-selection in recruitment and was not tested in this study. Three were bilingual; all bilinguals reported English as their primary language, but the data from one participant was ultimately discarded due to some uncertainty about
whether this was actually the case. Twenty-eight of the participants reported at least an intermediate level in one foreign language (bilinguals were not counted in this number unless they reported studying a third language). Participants received an honorarium of £7 for participation in this study along with participation in a second, unrelated study for a different researcher.

8.4 Hypotheses

The specific hypothesis tested was that F0 fall range, the most reliable topic-structure cue identified in the production study, would affect listeners’ perception of whether there had been a topic change or not. Since part of the purpose of the study was to develop a methodology, the complex distinction between different levels or categories of topic structure was not taken into account in this instance. The effect of the intonational manipulations was expected to manifest itself in a different form in each of the three tasks.

8.4.1 Pause detection

Pause detection latencies are said to be linked to the cognitive burden caused by other tasks. If the informational load provided by topic-change intonation is different from that provided by topic-hold intonation, then pause detection latencies after the different intonation contours should be different. It is not immediately clear, however, which should lead to longer latencies. On the one hand, topic-change intonation could be seen as providing additional information, in which case the detection latencies should be longer. On the other hand, topic-change intonation could be seen as clearing the working memory (cf. section 8.1.1), which would open up more processing “space”, and might lead to shorter detection latencies following topic-change intonation. In either case, there should be a difference in response latencies between the topic-change and topic-hold cases.

8.4.2 Sentence acceptability rating

The most straightforward hypothesis regarding the sentence acceptability rating task would be that in cases where the prosodic signalling and the follow-up sentence were matched, acceptability should be higher than in cases where there was a mismatch. However, as dis-
cussed in section 8.2.2.1, all of the follow-up sentences were potentially acceptable in a prosody-neutral context (e.g. in reading). Therefore it is possible that the semantic content of the sentences would override the prosodic signalling in the acceptability rating; and this is in fact what was observed in the pilot study. Therefore, participants’ response times could also be relevant. Even if follow-up sentences are judged to be acceptable in a mismatched context, participants may be slower to accept them when there is a mismatch with the prosody than when the prosody and the follow-up sentence match.

8.4.3 Sentence memory

Following Almor & Eimas’ (2008) finding that better-integrated information is remembered better, in cases where there was a match between the prosody and the follow-up sentence, participants’ memory of the follow-up sentence should be better than when there was a mismatch.

8.5 Results

8.5.1 Pause detection

Button presses for pause detection were measured automatically in E-Prime from the onset of the stimulus; then the length of time between the stimulus onset and the pause onset was subtracted to come up with a response latency for the pause detection. Any latencies shorter than 200ms (including those less than zero) or greater than two standard deviations away from the subject’s mean were discarded, as they presumably reflected a response to something other than the target pause (cf. Mattys et al. 2005).

The results for the pause detection task varied dramatically between listeners, and there was no overarching difference between response times after topic-change intonation versus after topic-hold intonation (df=427.46, t=1.317, n.s.). Some listeners’ mean responses were faster after topic-change intonation, while the mean responses of others were faster after topic-hold intonation. Within those two groups, some of the listeners showed obvious differences between their response times for topic-change and topic-hold items, and some did not (cf. figure 8.5).

There was a weak relationship between which intonation pattern led to faster responses
Figure 8.4: Pause detection latencies by subject. Dark grey bars show items with topic-change prosody, light grey bars show items with topic-hold prosody.
and the version of the experiment that participants took, suggesting that the pauses were more easily identified in some sentences than others, regardless of the intonation contour (ANOVA: F(7,414)=1.89, p=0.06). There was no difference in performance globally between the High and Low stimuli sets (F(3, 570)=0.9093, p=0.4362 n.s.); therefore the two sets were collapsed in the data analysis.

Given that some of the differences appear to be based on the individual sentence, it is important to look at the differences in performance on the same sentence (in both intonation conditions) across different listeners to determine whether the intonational differences had any effect on response times at all. However, in this case we once again see that for the most part, mean response times to the same sentence in different prosodic conditions were similar (cf. figure 8.5).
Figure 8.6: Acceptability of sentences in conditions where they matched and mismatched the prosody of the previous utterance

8.5.2 Sentence acceptability

8.5.2.1 Ratings

For all the target items, a response that the follow-up sentence was acceptable in context was ten times more likely than a response that it was not acceptable, regardless of the match or mismatch with the prosody of the intonation-manipulated sentence heard previously (see figure 8.6). The difference in distribution of responses was not found to be statistically significant ($\chi^2(1, N=887)=0.0089, p=0.925 \text{ n.s.}$)

8.5.2.2 Response time

Since no effect was found in the acceptability ratings themselves, an analysis was performed on the response times for items which were rated as acceptable.\(^2\) This allowed for a comparison between the ease of acceptance when there was a match versus a mismatch of the follow-up sentence with the prosody. Response times of less than 500ms were discarded as this would not have provided enough time for participants to read the sentence and hence

\(^2\)Items rated unacceptable were excluded from this analysis so that only equivalent responses were compared to one another.
the response must have been to something else (cf. Harley 2001). As noted above (section 8.2.2), response latencies of more than 2000ms were not recorded. Since the response time measure reflects only “acceptable” ratings and were therefore all responses on the same button on the response box, there should be no effect of participants’ handedness. While there was little correlation between the sentence length (in words) and the response time ($R^2=0.0336$), longer follow-up sentences did tend to have longer response times, and the longest sentences tended to have fewer responses overall as a result of the 2000ms cutoff. Since all sentences occurred in both match and mismatch conditions, however, this should not affect the overall results.

In the case of the response times, it was possible to observe an effect of the prosodic manipulations (figure 8.7). As in the pause detection task, the High and Low stimulus sets did not give different results, so the two have been collapsed in the data reported here. There is a statistically significant effect of the interaction between the prosodic condition itself (topic-change or topic-hold) and the match or mismatch of the follow-up sentence with the prosody (see table 8.4). When listeners heard topic-change intonation (that is, a larger pitch movement, independent of whether it was in the High or Low stimulus set), they were slower to rate the sentence as acceptable than when the prosody matched. How-
Table 8.4: Linear mixed-effects model for acceptable sentence response times.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Estimated (change in) response time</th>
<th>Standard Error</th>
<th>t-value (p-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prosodic mismatch, topic-hold prosody (Intercept)</td>
<td>1403.72</td>
<td>41.13</td>
<td>34.13 (0.000)***</td>
</tr>
<tr>
<td>Prosodic match</td>
<td>47.83</td>
<td>31.02</td>
<td>1.54 (0.1236)</td>
</tr>
<tr>
<td>Topic-change prosody</td>
<td>54.57</td>
<td>31.06</td>
<td>1.76 (0.0794)</td>
</tr>
<tr>
<td>Prosodic match, topic-change prosody</td>
<td>-109.40</td>
<td>44.14</td>
<td>-2.48 (0.0135)*</td>
</tr>
</tbody>
</table>

ever, when they heard topic-hold intonation, there was no statistically significant difference between the match and mismatch conditions. The difference in the overall response times is small (ranging from 30-70ms) in all cases, but only the difference between the topic-change+mismatched-prosody condition and the other conditions is statistically significant in the linear mixed-effects model.

8.5.3 Memory task

Scoring the memory task was somewhat tricky since there was a great deal of individual variation in the kinds of memory errors made. In the end, three different scores were assigned for each sentence. The first score was based on whether or not participants remembered hearing the stimulus sentence at all; a score of 1 was given if they correctly remembered hearing a target item, and 0 if they did not. The filler items were not included in the scoring so there was no penalty for incorrectly remembering that they had heard one of the filler items, which were not actually present in the initial listening tasks.\(^3\) The second score was based on how well the follow-up sentence was remembered. A score of 4 indicated a perfect word-for-word memory of the sentence. A score of 3 represented a nearly word-perfect memory. A score of 2 was given for memory of the basic idea of the sentence or for a word-perfect phrase. A score of 1 was given when one or two words were remembered. A score of 0 was given for either failure to remember, or else a completely inaccurate

\(^3\)Some speakers showed a small amount of interference from similar semantics in the filler items (i.e. a noun that was used in a target item and also in one of the fillers); this interference is not analyzed here. Given the relatively low number of apparent occurrences of this interference compared with the overall reliability of the results, it is unlikely that including a correction for these cases would noticeably impact the outcome of the task.
memory of the follow-up sentence. The third score was an average of the first and second score, designed to take both passive (i.e. sentence recognition) and active (i.e. reproduction of the follow-up sentence) memory into account; so even if a participant did not remember a follow-up sentence, the fact that s/he recognized the stimulus sentence was still taken into account in the same score. Although results were computed for all three scores, the differences between them did not yield any interesting insights, so the third score (the mean memory) is reported here.

Although the mean memory values varied slightly, the only significant difference between how well items were remembered was how recently the items had been heard in the task, which was scored based on whether the item had been presented in the first half or the second half of the listening experiment (table 8.5). There was a small, non-significant difference between items where the prosody matched or did not match, with mismatched items being slightly better remembered than matched ones.

Three of the items were extremely poorly remembered in general, with averages of near zero in all cases. However, even when these items are removed from the analysis, the finding reported here still holds.

8.6 Discussion

8.6.1 Pause detection

The prediction for the pause detection task was that listeners would have different response latencies to the unnatural pauses after topic-change prosody compared to topic-hold
prosody, which would then indicate a different cognitive load for processing a topic-change versus a topic hold. However, this was not found to be the case; listeners varied widely in their responses and there appeared to be no overarching pattern. Furthermore, even if we take these results as reflecting a systematic response by each individual listener, it is difficult to draw a simple conclusion because some listeners were slower in responding to the topic-change items, while others were slower in responding to the topic-hold items. Several explanations might be put forth for these results, including individual differences in strategy, the overall acceptability of all contours, or problems in the task design.

8.6.1.1 Individual differences in strategy

Although the task was designed in an attempt to force listeners to attend to the linguistic content of the utterances and therefore to process the utterances in whatever way would be normal for them, the fact remains that as an experimental task listeners may have had different strategies for identifying and responding to the unnatural pauses. The individual differences in response latencies to the pauses may have reflected a more prosodically-oriented versus a more lexically-oriented strategy, for example. Also, some listeners were better in general at identifying the pauses than others were (i.e. they had overall faster response times and/or more responses in total). Again, this may have had to do with where listeners focused their attention, or it might be related to individual differences in proficiency in dealing with prosody.

Depending on the priority which they assigned to each of the tasks, listeners may have been paying more or less attention to the pause structure versus the sentence acceptability task. This might also have depended to a certain extent on when they heard filler items with obviously unacceptable follow-ups; if they expected the acceptability rating task to be relatively easy as a result of hearing several of these early on, then they might have focused more on the pause identification. On the other hand, if they felt the acceptability task would be difficult, their attention might have been more on sentence comprehension and less on the correct identification of the pauses.
8.6.1.2 Overall acceptability of contours

In this experiment all of the intonation contours presented were acceptable; it was only the follow-up utterances that might be actually unacceptable. The processing burden measured by the pause detection task may have been the additional processing burden of having to deal with something that was inappropriate in context, but since all of the intonation contours were acceptable at the point they were heard, the pause detection task was unable to find a difference between them.

If this is the case, it raises the issue of to what extent a methodology is useful if it can detect only the difference between “normal” and “abnormal” processing, rather than the differences between two “normal” cases when the informational burden might still be considered to be different. One of the assumptions of the pause detection task was that it is indeed possible to detect differences between two “normal” cases; that is, although both contour types are part of normal language use, they still lead to a different cognitive load.

This assumption may be an important one in determining the role of topic in language processing, since whether or not it is true affects how we interpret the results of the pause detection task. If it is possible to detect differences between two “normal” cases of processing assuming that the task is different, and such a difference was not detected, this may mean that the task was not in fact different in the two cases. In other words, the interpretative process was not affected by the topicality or non-topicality signalled by the prosodic variation. This might suggest that topic structure does not have psychological reality for (some or all of) the listeners.

On the other hand, if a task like pause detection is only able to detect differences in performance between “normal” and “abnormal” processing, then the results reported in section 8.2.1.2 are neutral as to the use of topic structure by listeners. The lack of a difference for many speakers between the two conditions would suggest simply that both topic-change and topic-hold prosody signal something that listeners are accustomed to interpreting in the usual course of comprehension.
8.6.1.3 Problems in task design

There is a chance that the processing burden was in fact different between the topic-change and topic-hold intonation conditions, but that the task design obscured the measurement of this difference. Firstly, some subjects reported a great deal of uncertainty regarding what constituted the unnatural pauses in the experiment (although they nonetheless responded to the pauses with high accuracy). Duez (1985) showed that prosody can constrain pause detection in context, so a pause at the end of an intonation contour might have been considered acceptable even when listeners were trying to identify unnatural pauses. Conversely, it is also possible that the unnatural pauses, as part of the prosody of the utterances, interfered with listeners’ ability to process the intonation contour normally. While the choice of a pause detection task was in part specifically intended to be tied to the prosody, this may have caused too much confusion for listeners.

It is also possible that listeners responded to the pauses as though they were hesitation pauses, which might be the most similar phenomenon in natural speech. Chiappetta et al. (1987) found that native language influenced how listeners perceived pauses in speech stimuli, suggesting that it is difficult or maybe impossible to separate the processing of unnaturally-located pauses from the interpretation of other pauses simply by nature of the fact that language processing is occurring. Research by Marslen-Wilson and colleagues (Marslen-Wilson & Welsh 1978, Marslen-Wilson 1987) also suggests that linguistic processing is obligatory in the context of encountering linguistic stimuli. Lickley (1994, 1995) suggests that listeners are generally unaffected by hesitations in natural speech, to the point of being nearly unable to detect disfluencies even when asked. If normal linguistic processing is obligatory, this could be a serious problem for using the pause detection task in the context of longer stretches of speech. Furthermore, Collard (2009) reviews a body of literature suggesting that disfluencies may even affect a listener’s attitude towards a speaker. This means that the disfluencies might have been interpreted as bearing an informational load of their own independent of the experimental task. Again, if linguistic processing is obligatory, the experiment participants might not have been able to avoid these interpretations even though they were told explicitly that there were unnatural pauses in the stimuli.

Another possible difficulty with the pause detection task itself is the timing of the pause
in relation to the intonation contour. As noted above, the pause was placed as soon as possible following the target contour’s completion, though not at a place where there was already a phrase boundary (as this would license a pause and make even a longer pause sound relatively natural, in addition to creating the percept of a stronger boundary). The close time-point was chosen in consideration of evidence that phonetic information is lost quickly, possibly within 250ms (cf. Andruski et al. 1994). However, it is possible that the pause was actually placed too close to the end of the intonation contour, and that placing it later might have been a better choice in terms of being able to measure an effect of the processing of prosody. Even if the phonetic information was lost by this point, it is likely that the information about topic structure would have still been available; and in fact perhaps it was not even available until after this time (i.e. interpretation had not yet taken place). In this case, since the pitch contours were both acceptable in context, identification of a different response to them would only have been possible later than the time at which the pause detection task occurred. The results of the pause detection task would then only reflect the perceptual portion of processing, not the interpretative portion.

8.6.2 Sentence acceptability

An effect of the match or mismatch between prosody and the content of the follow-up sentence was observed in the response time data, but in interaction with the prosodic characteristics of the spoken sentence. When participants heard topic-change intonational characteristics but read a follow-up sentence that would have been consistent with a topic hold, they were slower to rate these items as acceptable than in other conditions. This difference suggests that hypothesis 2 was correct, and that listeners are using topic structure information to guide their processing as they listen, rather than simply having it available to them to refer back to if any sort of post-hoc and/or metalinguistic tasks are required of them. Of course, the acceptability rating task itself is metalinguistic; while listeners need to pay attention to whether sentences make sense in context, the response in a communicative situation would be to ask for clarification, or possibly to wait for clarification through further production by the speaker, rather than to simply identify unacceptability. Still, the fact that the observable effect was in the response times rather than in the responses themselves suggests that this task succeeded in getting at some of the underlying processes in terms of discourse under-
standing and the role of prosody as a signal to topic structure. Although the acceptability task was metalinguistic and under listeners’ conscious control, the speed of their response was less so and could therefore be considered as representing differences in the automatic processes of interpretation.

It is possible that the time at which this measurement was taken was too late in relation to the overall processing. Since topic is characterized as being pertinent to whole utterances rather than simply a few words at a time, the task was designed to measure the response to a whole utterance. However, it is likely that listeners have already become aware earlier in a sentence that a topic change is relevant or not; for example, changing a referent or even moving something into subject position from object position could suggest that this item has now been moved into the background. Birch et al. (2000) found that when a referent was presented in focus, participants were more likely to return to it later in a discourse, though not necessarily outside of that context. From a more theoretical standpoint, Büring (2005) proposes that an S-topic can “narrow down” a D-topic, leading to more constraint rather than more freedom within the D-topic (cf. section 2.3.1).

This narrowing-down could in turn imply (or perhaps confirm) that a topic change has taken place, well before the sentence has ended. If this is the case, then the bulk of the prosodic information about topic change would be relevant not after reading the entire follow-up utterance, but perhaps only a few words into it. If this point could be identified, it seems likely that the effect of the prosodic mismatch might be better measured there than after reading the whole utterance. However, it is difficult to identify without further empirical study where such a point might be, and without careful controls, it could end up in circular reasoning: that is, it might argue that the point where processing is slowest is the point where a topic change is relevant, simply because processing is slowest at that point. This relates back to the difficulty with using word-based paradigms as discussed in section 8.1.1; they could be useful for identifying interesting time-points in utterance processing, but less so for diagnosing the problem. The solution to this problem may be the combination of a word-based task with a more global task. Alternatively, a methodology such as eye-tracking which permits a continuous measure to be taken might be a better choice.
8.6.3 Memory Task

The result for the memory task, in light of other experiments in the literature using a similar paradigm, is somewhat disappointing; no significant differences were found between groups except for an effect of how recently the item was heard.

A small, though non-significant, difference was observed between how well matched versus mismatched items were remembered (see section 8.5.3), with items that did not match the prosody of the auditory sentences being remembered slightly better. Although this is not consistent with hypothesis 3 above, it could be an indicator that listeners were in fact aware on some level of the mismatch in the prosody. However, instead of disrupting their memory for these items, it might have instead made them more salient due to their unexpectedness, therefore slightly helping participants to remember the items (Sven Mattys, personal communication). This must remain merely speculation, since the differences in how well the items were remembered were not statistically significant and may simply be an artifact of the particular data set collected.

From the results of the memory task, it seems possible to make a distinction between the kind of information provided by the prosodic focus structure of an utterance, and the prosodic topic cues. As has already been well-established, appropriate prosodic focus (at least in intonation languages) is, if not essential, at least highly influential in terms of the perception of an utterance as well-formed. When focus is inappropriate for the sentence context, listeners’ comprehension ability suffers, although they are still able to understand utterances. However, the effect is great enough that when asked to recognize or recall the words that they heard, the task is more difficult in the case of a prosodic mismatch with the context.

For the topic structure information, however, this was not the case. In part it seems likely that this is because both of the prosodic contexts presented were “acceptable”; that is, neither one was actually highlighting something that should not have been highlighted, or vice versa. Therefore, the mismatch of the follow-up sentences with the prosody may have been interpreted as some kind of speaker variability or idiosyncrasy, rather than an actual error. This is reflected in the memory task; if Almor and Eimas (2008) are correct that better integration leads to better memory, it seems apparent that participants’ ability to
correctly integrate the utterances into the topic structure was unaffected by a mismatch of the sentence with the preceding prosody.

One implication to be drawn from this is that the two systems of focus and topic marking, though using the same phonetic means (i.e. prosodic variation), are not part of the same level or tier of linguistic structure. While focus structure in the sense of accenting versus deaccenting is now widely acknowledged to be part of a language’s phonological composition (cf. section 2.2.5), with a clear requirement for well-formedness on a phonetic level, the results from this task do not suggest a parallel requirement for topic structure. Instead, we may consider topic structure cueing to be part of a different level of linguistic structure, to be discussed in section 9.3.1.

8.6.4 General discussion

The observed effect of the mismatch between prosodic signalling and the semantic content of the follow-up items indicates that the question of the role of topic structure in speech comprehension is worth further study, though the methodology requires some improvement.

The main goal of this study was to investigate on-line processing of prosodic cues to topic structure. The only task which showed an effect in this case was possibly the most metalinguistic, that is, the sentence acceptability rating; but the effect measured was in response times, suggesting that the experiment was successful in identifying underlying differences in processing. The second goal of the study was to develop a methodology for testing the effects of topic-structure prosody on processing, which has resulted in the identification of several new issues to be taken into account in future experimental designs.

One ongoing question which still needs to be addressed is where in the time-course the topic structure information provided by prosodic signalling becomes relevant. From the point of view of efficiency or limitations on memory (cf. section 2.1.1.1), it seems probable that topic structure information is only maintained in working memory or considered “active” until such time as it can be resolved in some way. This resolution might be as simple as confirming that a topic shift suggested by the prosody does seem to have taken place in the lexical content as well. Given that the prosodic signals involved can potentially have many other functions, “resolution” might even be considered as a way of linking the prosodic signalling to a specific system; that is, identifying it as being related to topic struc-
ture rather than e.g. turn-taking or emotional involvement of the speaker. Alternatively, if the listener can be relatively certain that the signalling involved is linked to topic structure and not something else, for the sake of discourse coherence it would still be important to confirm in some way that the topic shift has in fact happened. However, the time at which the topic shift becomes relevant could vary. If the idea of topic shift is primarily concerned with facilitating or blocking things like anaphor reference, then topic might not matter at all until a pronoun or anaphor occurred in a following sentence, which would not happen at a fixed distance from the topic boundary, and potentially might not occur at all. This raises the question of to what extent a listener holds the topic information in working memory (if at all), and for how long. The evidence from the sentence acceptability task seems to suggest that the topic-shift information is actually not retained for very long, although it should be remembered that only one prosodic cue to topic shift, the size of the F0 fall, was presented to the listeners, and that having multiple cues might strengthen the cue and/or the degree to which the information is retained and/or accessible.

On the basis of this experiment, using only one prosodic cue, it appears that topic-change or topic-hold information from one utterance may not be retained through the following utterance. This is on the one hand unsurprising, as each utterance presents a new opportunity for topic-structure information to become available, and listeners have a limited capacity in their working memory (Baddeley & Hitch 1974; Baddeley 1990). As soon as a new utterance begins, listeners have information about that utterance’s topic structure available to them; and it is likely that whatever information is necessary to integrate that utterance with the previous utterance is also available long before the later utterance ends. If the purpose served by the topic structure information is complete, there is no reason for listeners to continue to clutter their working memory with that information. (Unless, of course, they are being asked to do a metalinguistic task involving identifying the topic structure or rating sentence acceptability.) Therefore, it seems likely that topic structure information from a preceding utterance is discarded as soon as it has served its purpose; as Grosz and Sidner (1986) have suggested, the “stack” of active information can be cleared.

This raises the difficult question of identifying the point in a sentence where prosodic topic-structure information from a previous utterance is discarded. At first glance, one choice would be to discard this information as soon as an anaphor or pronoun has been
resolved, since this might seem to be the purpose of topic structure. However, this ignores the possibilities that (1) there might be more than one pronoun or anaphor requiring resolution, and (2) topic structure might serve some other purpose beyond simply constraining these referents. Therefore, it is difficult for a listener to predict where in an utterance the topic-structure information from a previous utterance is no longer useful.

The duration of relevance of the topic structure information provided by the prosody may also vary depending on what is signalled, as indeed is apparent in this data. Participants' behavior was different depending on whether they had heard topic-change or topic-hold prosody. When they heard the (probably more salient) topic-change prosody, their responses were different depending on whether the follow-up sentence matched or did not match the prosody they had heard. However, when they heard the (probably less salient) topic-hold prosody, their responses did not differ. This suggests that either more information was provided by the topic-change prosody, or that the information lasted longer, or possibly both. The signal for a topic-change could carry a larger informational burden in that it breaks up the discourse into groups, and specifically signals the end of one group and the beginning of another. The longer-lasting effect of topic-structure prosodic cues could be seen as simply the perseverance of this new information, whereas the topic-hold prosody contributed no new information and there was therefore no additional burden on the listener.\(^4\) On the other hand, the greater salience of the topic-change prosody (i.e. a larger pitch movement) might simply have created a stronger perceptual effect and therefore a stronger memory trace, which lasted longer than the smaller, less salient effect of the topic-hold prosody. As Birch and Garnsey (1995) point out, sometimes memory for a specific linguistic form, in addition to memory for the meaning, is important in processing; and they show that words which are prosodically focused are better remembered, suggesting that items which are phonetically more salient are simply better remembered. This effect alone could probably explain the difference in the effect of the prosodic mismatch in the cases of topic-change versus topic-hold prosody. In any case, though, it seems unnecessary to force

\(^4\)While this argument fits the data, a topic-hold might also be interpreted as bearing more information (Anna Tristram, personal communication), since the listener must continue to keep in mind things like previous possible referents, or more items on the focus stack in Grosz & Sidner's (1986) terminology. In the context of this experiment, however, it seems most helpful to consider these phenomena in terms of continuations versus changes, with the assumption that a change always creates a greater processing burden than a steady state, since it must be interpreted in a new way (cf. Kluender et al. 2003).
a choice between the two options. The use of more salient and therefore more memorable
prosody in the topic-change case could be accounted for by the need to help the listener
retain the additional information provided.

8.7 Conclusion

The methodology presented here has proven promising in terms of opening up new avenues
to the investigation of the perception of topic structure and addressing some of the diffi-
culties inherent in conducting such a study. The results reported here of a slower response
time in the condition of mismatch between prosodic cueing and semantic content suggests
that this information is indeed recovered as part of speech processing; in other words, it
appears that topic structure’s phonetic correlates are used by listeners in on-line language
comprehension. Future studies of prosodic variability in perception should therefore take
topic structure (and the related prosodic variation) into account.
Chapter 9

General Discussion

9.1 Review of results

The research presented in chapters 3-8 addresses the role of prosody in signalling the topic structure of long discourses in SSBE, from the point of view of both the speaker and the listener. My research makes several unique contributions in the investigation of both prosody and discourse structure. First, I investigate prosodic signalling of topic structure taking into account a more complex model of this structure than most previous phonetically-oriented studies had used, and also taking a phonologically-informed view of the phonetic variation aspects. Second, I address a neglect of non-F0 prosodic cues in the context of discourse structure signalling by examining speech rate and phonation quality cues in addition to F0. Third, I conduct the first experimental investigation of the perception of topic structure cues by listeners in an on-line processing situation rather than via a metalinguistic task, opening up new methodological avenues for the study of prosody and discourse structure.

The production study shows that a variety of prosodic cues, both cues that have been previously investigated by other researchers and cues that are identified for the first time here, vary systematically in conjunction with the topic structure of the discourse. In this study written text which was controlled for the topic structure and for segmental characteristics was read aloud by 16 speakers, and the recordings were analyzed with regard to the topic structure. Pitch reset (i.e. the production of the first F0 peak at an extra-high level) was found to occur at the beginnings of new topic groups, as had already been found in numerous other studies (Lehiste 1976 *inter alia*). Similarly, a reduced global speech rate
was also found to correlate with the introduction of new topics in this data (cf. Brubaker 1972). F0 peak delay, although not found to be directly related to topic structure as hypothesized (cf. Wichmann 2000), was found to be potentially related to pitch accent category, depending on the model of intonational phonology chosen, with late peak (i.e. rising) pitch accents more likely to occur in new topic utterances than in other topic categories.

Several production cues were also identified that do not appear to have been investigated in the previous literature. Most notable is variation in the size of the first F0 fall in an utterance, which varies in relation to the internal topic structure, not only the boundaries of the topic group. In the current data, this was the most reliable cue to discourse topic structure for most speakers. Similarly, although global speech rate appears to vary only between new topics and non-new-topic utterances, the local speech rate of the first few syllables of each utterance also varies in relation to the topic structure categories, with faster speech rates occurring at stronger boundaries. This variation is similar to that reported by Li & Zu (2008) for Mandarin, as well as to “rush-through” as described in Conversation Analysis literature (e.g. Schegloff 1982, 1987). The local speech rate variation is particularly apparent in speakers who made less use of the F0 fall range cue, suggesting a system in which cue-trading is possible. The use of aperiodicity at the ends of utterances was also a weak cue to the topic structure category of the upcoming utterance, with aperiodicity less likely before Continuations than in other contexts. Although aperiodicity has been well-studied at the phrase boundary level, it had not previously been investigated in terms of higher-level organization.

A post-hoc validation of the topic boundaries and topic-structure categories used in the production study addressed the question of whether these categories were applicable to the Emory text, since the categories identified for narrative texts normally take a different form (cf. Labov & Waletzky 1967; Chafe 1994). The existence of the Topic and Continuation categories were confirmed by rater agreement, while the validation also provided an explanation of some of the ambiguous data in the form of greater variability among raters as to the characterization of the Addition and Elaboration categories. A reanalysis of some of the phonetic data using revised topic boundaries confirmed that the basic form of the phonetic results should remain the same even with a new topic-structure analysis applied. The variability among raters raised the important issue of individual variation in the characteri-
zation of discourse structure, and indicates that the lack of robustness in some of the original phonetic results may be due to readers having characterized the discourse topic structure in different ways.

Numerous earlier studies, notably those by Swerts and colleagues (Swerts & Geluykens 1993; Swerts et al. 1994; Swerts 1997), used a listening paradigm to identify prosodic correlates of topic structure. However, these studies showed only that listeners are capable of using prosody to identify this kind of discourse organization, without demonstrating that it is regularly used in processing. The perception experiment presented in chapter 8 gives evidence that topic structure prosody is relevant in normal discourse processing, even when listeners are not actively asked to attend to grouping. Although the experiment suffered from some weaknesses in design, there was still an observable effect of the match or mismatch of the prosody to the semantic content on the speed at which listeners were able to process what they heard, indicating that they were taking the topic structure cues into account as part of comprehension. Specifically, a mismatch between the topic structure indicated by the prosodic signalling and the topic structure indicated by the semantic content of the utterance caused listeners to be slower to respond than when the two cues were congruent with one another. To the best of my knowledge, this is the first study to provide evidence that listeners are sensitive to topic-structure cues in online processing when they are not explicitly asked to identify topic changes.

The results of these experiments bring new evidence to bear on a number of open questions in both prosody-related and discourse-structure-related research. The structured prosodic variation in the production experiment as well as listeners’ sensitivity to topic-structure cues in the perception experiment lend support to theories of discourse structure that make topic relevant to the overall organization of discourses. This relationship in turn leads to the interpretation of the prosodic variation observed as falling within linguistic systems, rather than being simply a general communicative device. The experimental approaches used in both the production and perception studies also address some methodological issues that may be identified in previous research on the prosody of discourse structure. The discussion below addresses the implications of the current research for research on discourse topic and prosody, and for experimental methodologies in these fields.
9.2 Implications for topic

Although the above studies were conducted primarily in the domain of prosody rather than of discourse analysis, the results have implications for theories of topic structure. Prosodic signalling in speech can suggest ways in which the spoken text is organized. While it is possible that the organization is more complex than what speakers signal (or that the same signals have different purposes in different contexts), the variation in prosodic signalling gives a baseline for the minimum amount of variation that we should be looking for in topic structure, assuming that speakers do not produce variation meaninglessly (cf. Relevance Theory, Sperber & Wilson 1995).

9.2.1 Prosody and existing theories of topic

The existence of the grouping of discourses into units of related information seems uncontroversial, regardless of the theoretical status of “topic” as an overt element in the discourse. However, the existence and form of finer structure within these units is a source of more debate. Blakemore (1992, 2002) argues that any “structure” we may identify within this kind of unit is an artifact; that is, ideas that are more similar are naturally presented more closely together in a discourse, but beyond that speakers do not have any structure in mind for their discourses. This view is appealing particularly because it eliminates the need to account for planning, which intuitively seems as though it ought to be harder in spontaneous discourse.

However, on the basis of the prosodic variation observed in the production study, it is possible that defining items as “more” or “less” related to one another is not specific enough. First of all, Blakemore’s (2002) argument implies that the relationships involved are linear; that is, the relatedness of an utterance is defined in relation to the preceding utterance and nothing else. This might be contradicted by the results of the production study, however, which suggest that items in similar topic structure categories show similar behavior, regardless of their location within a discourse (cf. also Wichmann 2000). The validation tasks reported in chapter 7 demonstrate that untrained raters were able to use at least the Topic and Continuation categories consistently in describing the structure of the discourse. While this result might be attributed to having defined Topic as “very unrelated” and Continuation as “very related”, the fact that raters appeared to cluster in their interpretations of Additions
and Elaborations leaves the possibility of the existence of categories open.

Of course, the existence of categories is not the only possible way to interpret the data; it could equally be explained by a hierarchical structure (cf. Grosz & Sidner 1986). The basic concept is the same: items in the hierarchy dominated by the same node should be produced similarly, taking into account locational factors within the group of utterances. Since the items categorized as Additions should all be directly dominated by the Topic, their prosodic realizations are expected to be similar. The difference between the hierarchical theory and the category theory is that in the hierarchical theory, it is the height of the utterance in the hierarchy that matters, so the depth of embedding, or change between levels of embedding, should affect the production of the utterances, with larger changes carrying more salient prosodic signalling.

Applying a simple hierarchy to the text used in the production study suggested that the categories, rather than the hierarchy, provided the best explanation for the variation observed in the data (cf. section 3.3.2). However, it is difficult to make a conclusive statement about whether a category or a hierarchy approach is most appropriate for describing discourse topic structure. Applying either approach entails making an *a priori* decision about the categorical versus gradient nature of the prosodic variation and structure. A categorical theory of discourse structure could lead to the strong hypothesis that the prosodic variations should have categorical interpretations; that is, each category should be uninterpretable as any other on the basis of the prosody. There is some evidence that this is possible, at least for the distinction between new Topics and other categories. However, the overlap in the production of the other three categories, and the results of the category validation (cf. section 7.4), as well as the need to take into account other factors such as declination, means that this strong hypothesis is unlikely to be true for all the categories.

This conclusion is further supported by a return to the idea that the set of categories available depends upon the type of discourse. Assuming that this is true, either every possible category type would have to have its own specific set of prosodic characteristics, or there would have to be a complex many-to-one mapping of categories to sets of prosodic characteristics that would be shared among members of a language community. The first option seems extremely unlikely unless it becomes possible to identify a small set of categories that are nearly universal. The latter is slightly more likely, but it too depends on speaker
and listener agreement as to stylistic possibilities and conventions; and furthermore, these conventions must be accessible at least across dialects (as shown in the perception experiment, where listeners from a variety of dialects of English were sensitive to the topic cues found in SSBE). Therefore, it seems only reasonable to discard the strong version of the hypothesis that the prosodic variations are inherently categorical. Of course, this does not prohibit a category-like interpretation, as can be seen in the case of vowel perception, where categories are difficult to identify experimentally despite being evident in speech perception (cf. Strange 1999).

Despite the indefensibility of the strong hypothesis, the patterning of the variation with the topic structure categories, as well as the ability of raters to consistently assign category labels, suggest that it is not sufficient to discard the notion of internal topic structure altogether. Instead, it would appear that in order to account for the prosodic variation, it is necessary to posit that the speaker intends some kind of internal structure, whether categorical or hierarchical. The systematic variations in the prosodic realization support the lexical-semantic information in the utterances in helping the listener reconstruct this organization in the interpretative process. The importance of the topic structure in interpretation of at least some kinds of discourses is supported by the wide variety of prosodic cues that are available. These include both direct cues, such as the size of initial (prenuclear) F0 falls and variations in speech rate at initial utterance boundaries, and less direct cues, such as the distribution of pitch accent choice or the use of aperiodicity at the ends of utterances. In fact, the system of available cues is extremely rich and often redundant, which means that information about topic structure is available throughout the course of an utterance; even if a listener misses cues at the initial boundary, for example, this information is still available later in the utterance, to be made use of if and when it becomes necessary. This ready accessibility indicates an important role for the topic of the discourse.

9.2.2 Topic and discourse purpose

In general, the explicit identification of what the topic of a discourse is may be relatively rare (cf. Chafe 1994); in normal conversation it might be most likely to be important in the context of reported speech (e.g. “He was talking about the party.”). Furthermore, not all kinds of discourses make the idea of topic relevant. Sequences of greeting, for exam-
ple, do not have information exchange as their main point; instead, the goal of this kind of interaction is the act of opening the conversation itself (Richard Ogden, personal communication). In this kind of interaction, or portion of an interaction, although it may be possible to identify a topic (e.g. discussing the weather), it is unlikely that this would be meaningful for purposes of discourse understanding. However, topic structure cues might be relevant in signalling a change from one kind of interaction (e.g. conversational openings) to another, in which topic structure might be more relevant (e.g. telling a story, sharing information). Therefore the use of the term “topic” with reference to “aboutness” could still be misleading. It might instead be better to think of topic shift as having scope over a shift in a variety of important characteristics of the text more generally, such as the discourse purpose, as Grosz and Sidner (1986) suggest, or an episode shift in a narrative, as described by Chafe (1994).\(^1\)

Language on all levels undergoes grouping processes; whether these grouping processes are of phonemes into words, words into syntactic constituents, syntactic constituents into utterances, syllables/words into phrases, or small phrases into larger phrases. Grouping seems to be related to memory processes; since we can only hold a limited number of items in working memory (cf. Baddeley & Hitch 1974; Baddeley 1990), it behoves us to make those items as rich as possible.

Topic structure provides another grouping mechanism, of utterances into utterance groups (paratones). This grouping would seem to have at least two obvious purposes. First, the grouping helps a listener to know when a discourse purpose has been completed, or rather when the speaker thinks it has been completed; this means that the listener knows when to make inferences to fill gaps, or alternatively, when to ask questions if his or her inferences do not seem to have resulted in a complete discourse representation. Second, and related, the grouping constrains reference. It is widely recognized that topic change can constrain e.g. pronoun and anaphor reference; these elements generally refer back to things that have occurred within the current unit. This relieves a burden on the listener’s memory by allowing extraneous referents to be discarded, and probably makes processing faster since a search process for referents does not have to go through so many items. Thus the

\(^1\)It may not be necessary to make an absolute distinction between discourse purpose and narrative elements, since telling stories also performs a social action (cf. Johnstone 2001 for a brief discussion).
speaker’s intention that the listener understand the topic of a discourse (or discourse portion) contributes to understanding by defining the boundaries of what must be retained for optimal understanding, or, as Büring (2005) suggests, what may be taken for granted. Although as previously discussed this is not strictly essential, it helps ease processing, perhaps particularly of more complex discourses.

It turns out that expanding the scope of discourse topic beyond “aboutness” provides for a clearer account of the production and perception data reported in earlier chapters. First of all, it requires that the topic structure still be marked, so that the listener can make use of it. Since the topic is linked to the purpose of the discourse or discourse segment, or to the episode currently in question, the long-term recoverability of topic information is not surprising; the speaker must ensure that the listener has access to the relevant information. Grosz and Sidner (1986) argue the discourse purpose is always relevant to the interpretation of the discourse, and so its recoverability is of high importance. Similarly, the kinds of content that may change in an episode shift (e.g. the setting or the point of view) form part of the context or environment that is obligatory for the successful interpretation of a narrative (Chafe 1994).

9.2.2.1 Discourse purpose, episode shifts, and individual variation in production

The expanded scope of “topic” can also account for the wide variation between participants in the production study, for the simple reason that their interpretations of the text could have been drastically different (cf. Brown and Yule 1983). In the production study, participants were seated in a sound-attenuated booth, reading into a microphone. Their instructions requested that they read in any manner that felt natural to them. They were not alone in the booth, but neither were they reading to a visible audience, since the experimenter was seated behind them. This means that they had to form their own concept of their audience. This idea would in turn be affected by their imagination, experience with other experiments, and/or general engagement with the task. Some readers were very expressive, and it seems clear that they imagined some kind of audience to whom they were performing; other speakers may have been reading the text to the experimenter and/or with the sense that style did not matter.

The environment of the experiment, a small booth with technical equipment, probably
did not contribute positively to naturalness (although this is not to say that it necessarily had a detrimental effect for all speakers). All of these environmental factors would have contributed to very different interpretations of the purpose of the discourse. A speaker who felt s/he was reading to an audience would be more likely to prioritize factors involving understanding; prosodic signalling of the discourse structure would be more relevant for this speaker. A speaker who assumed that the message of the discourse was unimportant would be more likely to either ignore the structure of the discourse, or to signal it in minimal ways. A speaker who was more engaged with the task might put more emotional signalling in, which might make the prosodic signalling easier to identify; for example, larger pitch movements in all contexts might make variation in the size of pitch movements more apparent.\(^2\) All of these things are related to the perceived purpose of the discourse (and possibly even to whether or not speakers thought it had a “real” purpose at all). If the speakers did not think that it was important to communicate the topic structure or the context (i.e. because they were not communicating with anybody), then we would expect only minimal prosodic variation.

Johnstone (2001) summarizes a body of research indicating that the structuring of a spoken narrative is a skill that speakers develop as part of the language acquisition process. This implies that individual differences in ability to represent a discourse structure are probably to be expected, although it is impossible to know whether the experiments reported in this thesis would have gone beyond any of the participants’ abilities to structure discourse. In any case, differences of ability need not be the main cause of differences in topic-structure characterization. Chafe (1980, 1994) acknowledges that individuals differ as to judgments about topic organization and even how long in terms of time “semiactive information” (or topic) may be retained. The topic boundary rating task (cf. section 7.3), in particular, emphasized the differences among readers as to how long a single topic could or should last. The multidimensional scaling analysis of the category boundary ratings, emphasized differences in judgment about structure, with speakers clustering together apparently on the basis of their interpretation of the two more ambiguous categories, Additions and Elaborations (cf. section 7.4.2). This freedom of interpretation of texts, however, is still ultimately

\(^2\)The text was not intended to evoke an emotional response from the readers, and so this was not included as a factor in the analysis of the production study; however, readers’ engagement with/interest in the text could be seen as an affective characteristic of their production.
hemmed in by the need to communicate effectively; an expanded understanding of “topic” may allow for more flexibility in interpretation.

9.2.2.2 Discourse purpose and perception

A purpose-based account of topic structure also accounts for the results of the perception experiment. All of the sentences involved would be comprehensible on their own, so prosodic signalling was not essential to semantic comprehension (leaving aside for the moment the question of reference interpretation); nonetheless, listeners were slowed by a mismatch of the prosodic information to the semantic content of the utterances, particularly when the prosody suggested a topic change but the topic change did not occur in the lexical semantics.

If the goal of topic signalling was simply to flag a part of the discourse with a particular topic structure as a kind of formal designator, it seems likely that listeners could learn to ignore the prosodic signalling: it would not necessarily be important to them from an interpretative point of view, and would in fact be problematic in the prosodic-mismatch contexts. However, if we understand the topic signalling to be part of the discourse purpose, or necessary for providing an environment for what is said, then the fact that listeners continued to attend to the prosodic signalling makes more sense. If understanding the topic is part of the discourse purpose, and understanding the discourse purpose is part of understanding the discourse, then listeners cannot afford to ignore the prosodic cues, even if they prove unreliable in the context of the experiment.

9.2.2.3 Re-examining “topic structure”

Thinking of topic structure in terms of the wider discourse purpose helps eliminate the choice between a category theory and a hierarchical theory. Instead, the wide variety of categories are always available, but the type of discourse constrains the kinds of categories that will be relevant. For a narrative, like that used in the production experiment, the category elements might be defined on the basis of episode shifts (cf. Chafe 1994), which keep relevant information active at the appropriate times. For example, a new Topic could be defined as a location with a critical number of episode-shift characteristics: perhaps a change in location, time, and character point of view occurring simultaneously. Conversely,
a Continuation could be identified as a location in which no episode-shift characteristics are present. In other kinds of discourses, the relationships used in a narrative might not be ideal; Nakajima and Allen’s (1993) grouping of topic structure categories suggests some possible sub-variations under each category (see section 2.3.2.2), which would allow for the definition of categories to be somewhat flexible given different kinds of texts.

In some cases, having categories of information relationship may not be useful. This might be the case in, for example, a setting in which one speaker was asking questions and the other answering them. In this case the relevance of the responses should be clear from the context in which they are provided (i.e. as responses to the questions), although prosodic cues to the utterance relationship may still be relevant if, for example, the person answering questions wishes to request a clarification before answering. Some kinds of discourses may also not make topic relevant, as in greeting sequences (cf. section 9.2.2); in this case the topic-structure cues we have identified may be co-opted for some kind of similar purpose, for instance providing boundaries for a greeting sequence, helping to move it on into some other conversational function. The role of the cues, in signalling the speaker’s purpose or intent for a portion of the discourse, can remain the same in this case.

On this basis it seems most expedient to adopt a model which takes into account discourse purpose and text type (e.g. task-based dialogue versus narrative), with the unifying factor of a categorical description, which can help link the ways in which different types of texts with different purposes may be described. Rather than a simple hierarchy that is unrelated to the content and/or the type of the discourse, the hierarchy should be sensitive to the internal context, with topic-structure categories, if they are relevant to the type of discourse, able to influence how the structure is formed. In other words, elements have a place on the hierarchy because of their topic structure category, rather than independently of it. While the hierarchy idea remains intact, the main difference from Grosz and Sidner’s proposal is that in this case it is necessary to motivate the placement of elements in the hierarchy on the basis of the topic structure information. This means in some cases that there could be “gaps” in the hierarchy; that is, levels which are left open because the topic structure does not require them to be filled. A larger distance between levels would be signalled by more prominent prosodic cues, regardless of whether the gap is filled or not. This would help explain why the prosodic signalling for category structure versus hierarchical structure does
not necessarily match (cf. section 3.3.2).

The idea of “empty nodes” on a structural hierarchy may call to mind generative syntactic theory, where empty locations on the structure are necessary to allow for “movement” or other complex features. However, the characterization of hierarchies as having gaps is not meant to motivate a system of this sort. Instead, it is meant to help account for the fact that language, particularly in conversational settings, is not strictly structured around a definitive form (in terms of topic organization). Instead, speakers cobble together bits and pieces of information, mentioning what is most salient, whether this is due to the listener’s need to know or perhaps simply due more to varying interest on the part of the speaker, and leaving out other pieces. A relatively sparse structural formulation is therefore probably best suited to the description of topic in informal speech settings, since it allows the speaker the greatest possible degree of flexibility in terms of choosing which information to include and exclude. If we adopt the assumption of Relevance Theory (Sperber & Wilson 1995) that speakers say only what is necessary to communicate their intent, then we can see this sparse structure as a framework where other information may be filled in as necessary either through pragmatic inference or by the explicit addition of further detail by any interlocutor in a discourse situation.

9.3 Implications for prosody

Although intonation, including phrasing, is now well-established as part of linguistic systems, it is clear that not all prosodic variation conveys linguistic meaning (i.e. on the basis of an abstract underlying system). Rather, some prosodic meaning appears to be universal and not linguistically structured, perhaps deriving from the Biological Codes (Gussenhoven 2004). Marking of emotional characteristics, for example, is widely considered to fall outside the bounds of the linguistic system (cf. section 2.2.5.1). It is important in intonational research to draw a boundary between prosodic variation that is or is not linguistically structured, since this will affect how we interpret systematic variation in the signal. If we believe that an observable phonetic variation is motivated by the linguistic system, then we may

This idea, in fact, calls to mind prescriptive techniques for formal writing as taught in primary or secondary education. It is not impossible for such structures to exist, it is simply the case that they are much less likely in informal speech.
model it in terms of an abstraction in the phonological system, whereas if some variation is considered to be a general communicative strategy, then it need not be modelled in this way, although there may still be meaningful systematic variation. The evidence surrounding the prosodic variation cueing topic structure suggests that it does in fact fall within the realm of linguistically structured, or “grammatical”, variation, as will be argued below. This in turn means that it is important to include it in a model of the intonational/prosodic structure of languages, although it is not part of the phonological structure (cf. section 3.3.5).

9.3.1 Linguistic structure

The prosodic variations observed in the production study would appear to be related to two different components of linguistic structure. Some of the variations occur within the intonational system (notably the choice of pitch accent). Others seem to be licensed by the discourse structure (size of F0 falls, speech rate), and are overlaid on the sentential intonational organization. There is good reason to believe that all of the variation observed is within the linguistic system, however, rather than being paralinguistic or non-linguistic. There are at least two lines of evidence which give support for this argument.

The first is simply that the variation can be identified to have a linguistic purpose, namely the organization of discourses around topics (or discourse purposes involving topics). Although the degree to which topic organization is language-specific versus general to all languages can be debated, its contribution to coherence and cohesion of discourses is part of how language is structured and how meaning is conveyed through language. The cueing system may have developed out of paralinguistic signals as suggested by the Biological Codes, for example with wider pitch excursions signalling more important information. However, these variations have apparently been grammaticalized, occurring systematically across speakers of the same dialect.

The other reason to believe that the prosodic variation is controlled by the linguistic system is that the variation occurs in units that are linguistically important. Herman (2000) made this observation about her data, finding that variations between utterances in medial and final position started in “linguistically relevant” places, for instance at phrase boundaries or at pitch accents. Nakajima and Allen (1993), although they do not make this observation explicit, also note the same thing: one of their indicators for topic structure is the relation-
ship between the F0 peaks in the utterance; these F0 peaks are pitch accents and therefore linguistically relevant points.4

The same is true of the data from this production experiment. The strongest cue to topic structure identified in this study was the variation in the size of the F0 fall in the initial prenuclear pitch accent in an utterance. Note that the topic structure did not affect whether or not there was a pitch accent in a given location;5 instead, the topic structure affected the characteristics of that pitch accent. The initial prenuclear pitch accent is an interesting location for meaning variation for a number of reasons. Although the status of prenuclear accents compared to nuclear accents has been debated in different intonational theories, recent work by Niebuhr and Zellers (to appear) indicates that prenuclear accents can bear the same meanings as nuclear accents, making them essentially “the same thing,” though perhaps still having different purposes in general. Certainly many intonational theories have indicated that nuclear accents are the main element of intonational meaning, and that they somehow bear a different (and greater) functional load. This is supported by the fact that new information often comes later in the utterance, where it can bear the most prominent (i.e. nuclear) pitch accent, whereas earlier information, even if accented, has often been already given in the discourse (although the status of the nuclear accent as phonetically more prominent is not really viable; this does not stop it from having a “special status” with regard to focus marking, cf. Ladd 1996). This means that the prenuclear pitch accent, even if it is able to have the same meanings as the nuclear pitch accent, need not do so. Carrying this argument a step further, this means that the functional load on the prenuclear pitch accent is lighter, and therefore, the prenuclear accent is an ideal location for additional information to be added to the utterance; in this case, information about the topic structure. The prenuclear accent, falling in what might be a relatively predictable environment (cf. Bock & Mazzella 1983; Terken & Hirschberg 1994), creates a prosodic space for other variations to take place; in essence, it is a way to insert additional prosodic information into the utterance without interfering with the underlying structure or conflicting with listener expectations.

If the existence and the size of the prenuclear accent should be relatively predictable on the

---

4This could be seen as contradicting Xu & Xu’s (2005) claim that every syllable in English is specified for pitch.
5It could be argued that the topic structure directly affected the speakers’ choice of pitch accent, cf. section 4.3. However, it seems more likely that the pitch accent choice was based on the informational structure of the utterance which was affected but not determined by the topic structure characteristics of the utterance.
basis of metrical and information-structural characteristics of the preceding discourse, then
its presence is a matter of course, and listeners can concentrate on using it for purposes
of gathering other information about the utterance and/or discourse rather than needing to
interpret the accentuation.

One way to test this hypothesis would be to investigate topic structure signalling in
languages with lexical accent or tone. In a tone language, potentially every word might
have a higher burden on the prosody; it would be interesting to know whether words which
were contextually more identifiable or in which the tonal information was less relevant for
understanding in some other way would be more likely to bear prosodic topic structure
cues. Similarly, it would be interesting to investigate in a lexical accent language whether
locations with intonational but not lexical-accent prosodic variation would be more likely
to bear these prosodic variations. Of course both of these hypothetical experiments would
depend on the assumption that lexical accent or tone languages would use prosodic cueing
to topic structure; this appears to be the case for the phrase-accent language Korean (Kong
2004) and the tone language Taiwanese Mandarin (Tseng et al. 2009). It is possible that
the functional burden on prosody in these languages would already be too great, and that
topic would be signalled by an entirely different part of the linguistic system (for example,
by morphosyntactic or lexical means).

It is clear from the series of experiments reported in this thesis that the linguistic prosodic
system exists on a level beyond that of the utterance, and that larger units influence the pro-
duction of intonation and other prosodic cues. However, these prosodic structures do not
necessarily have to be motivated by the internal intonational phonology system, and in fact
it seems clear that this is not the case. Instead, the discourse system has direct access to the
prosodic marking, in the sense that it can modify the output of the intonational phonology,
giving pitch accents a “non-standard” form in order to signal topic structure characteris-
tics. Discourse prosody could be seen as occupying an additional tier in the autosegmental
representation, with links to representations in the intonational tier: for example, modifica-
tions of scale of an initial pitch accent in an utterance, or speech rate information affixed to
boundaries or having scope over an entire utterance. (It seems clear that the “unit” bearing
these prosodic modifications varies depending on which prosodic feature we are consid-
ering.) The phonetic implementation of the utterance would reflect the input from all of
these tiers, producing different topic-structure variants of the same sentence-level intonation, for example; but all of these variants would be well-formed according to the rules for the sentence-level intonation (cf. Post 2011).

One way to account for the variation among speakers is to suppose that the preferred prosodic features activated by different discourse categories are different among speaker groups, even within the same dialect. Perhaps the specification is for “increased prominence” in general, which may be interpreted in a variety of ways. This would account for the alternation between speakers using F0 modification versus speech rate modification in the same contexts, for example. Also, if we see these discourse structure modifications as a gradient system, dependent upon the speaker’s decision as to its importance, then we can account for the greater or lesser degrees of contrast that we see in different contexts.

An argument supporting the presence of this cueing in the linguistic system would be the existence of other (i.e. non-prosodic) linguistic means of marking the same contrasts in topic structure, and this does appear to be the case (for example, the lexical marker -wa in Japanese, which is affixed to a sentence topic). The introduction of a new topic in an interactive context may be preceded by discourse actions like those reported by Jefferson (1993): minimal responses\(^6\) or assessments of previous conversational turns. These alternative linguistic means for indicating topic structure features, by occurring as part of the linguistic system, lend support to the argument that the prosodic cueing of topic structure does belong specifically to the linguistic system as well.

9.3.2 Multiple cues

One of the most important implications of the studies reported in this thesis is the role of multiple and possibly redundant prosodic cues to topic structure. F0 has been the most-studied cue to intonation, in part because it bears a relatively easy-to-describe relationship to pitch perception, even as it has become clear that it is not sufficient by itself to account for how listeners perceive the prosody of utterances. My research therefore contributes to a body of literature indicating that more investigation of multiple cues is necessary. At the same time, it provides an attempt at a description of the ways some of these cues interact.

\(^6\)“Minimal responses” are likely to include prosodic features indicating their minimal nature as well as being minimal from the point of view of lexical content; in CA the nature of minimalness does not distinguish between these two.
One thing that is apparent from the results of my studies is that in some cases it may be important to think of “prosody” as a unified system, even as it comprises a variety of cues (intonational, rhythmic, etc.). One of the most important pieces of support for this argument is the apparent trading relationship between local speech rate and local intonational variation in signalling the topic structure categories. As part of the prosody, these two unrelated cues appear to be substituted for one another by different speakers of the same dialect, and in some cases by the same speakers in different contexts. This means that we cannot afford to investigate the F0 variation, for example, without paying attention to the speech rate variation (at least in the context of topic structure). It appears that speakers are taking advantage of a variety of available cues to make the most of opportunities for prosodic signalling of topic structure. Since we have established that these cues fall within the linguistic system, it is difficult to argue that this is simply a paralinguistic artifact; the two different speech rate cueing systems also argue against this since they are essentially opposed in their realizations; this suggests that there has been a process of grammaticalization which has gone in two different directions (cf. Gussenhoven 1999, 2002).

One difficulty with thinking of prosody as a single system is that it is difficult to describe it in such a way that it can all be considered one system, beyond the basic definition of sound phenomena that extend beyond a single segment. One way to think of prosody as a whole would be to consider it as the sum of linguistic phenomena that have grown out of grammaticalization of Biological Code phenomena. Although this might appear to be a somewhat haphazard collection, it gains unity when thought of in this way. It also allows for different realizations of prosodic phenomena across languages or across groups of speakers via different grammaticalization patterns from the same Biological Code (Gussenhoven 1999).

In investigating prosodic phenomena, we can observe that along the lines of the Biological Codes, there are certain phenomena that we expect to see “in agreement” with one another; that is, high F0 peaks might co-occur with high intensity, or a greater degree of final lengthening preceding a longer pause. It seems obvious that these cues mutually strengthen one another by their agreement, and we might expect to see a commutative effect when multiple cues are in use: for instance, raising the peak F0 of a contour by a certain amount while holding the intensity steady might have the perceptual effect of an increase in prominence.
equivalent to that of increasing the intensity while holding the F0 steady (though obviously
the necessary variations in F0 versus intensity need not be of equal degree to lead to the
same perceptual effect).

However, we also see prosodic variations that appear to be in conflict with one another;
one notable example of this are the local and global speech rate patterns reported in chapter
5. It appears that at stronger topic boundaries, the local speech rate is increased dramatically,
even as the global speech rate decreases in new topics. One concept it is important to keep in
mind in studying all linguistic phenomena is that it is change that is relevant for perception
(cf. Kluender et al. 2003), not steady states. Following on from this, it is contrast between
different points in the signal, rather than an absolute value for any measure at a single point
in the signal, that is important. Prosody creates this contrast between different portions of
the signal, producing characteristics of different portions of the signal which are different
even if the lexical/semantic information is the same. It is valuable precisely because it can
create contrast with itself; it does not have to transform other linguistic levels in order to
do this, and it can even create contrast within a single element of the linguistic structure.
For example, changing the prosodic relationship between two elements of a compound can
signal a change in meaning (e.g. “white hóuse” versus “White House”), but it can also
maintain the same referent while suggesting a different interactional context (e.g. “Go to
the white hóuse. No, not the blue house, the whíte house.”) This contributes to keeping
lexicons smaller, since the same compound can be interpreted either as a simple collocation
by the speaker or as a reference to a very specific location. It also contributes to keeping
utterances shorter, since in the second example the speaker is not obliged to add more detail,
simply to clarify information that has already been presented, indicating the key piece to a
puzzle that the interlocutor has perhaps failed to identify.

Returning to the speech rate data from chapter 5, we can see this system at work,
wherein the prosodic variation creates contrasts with itself at relevant points in the signal;
we can assume that at least part of the purpose of this is to maintain a variety of ongoing
signals to the topic structure. The initial-boundary local speech rate contrasts both with
the global slow speech rate in new topics, which likely creates a greater perceived distance
between the two, but also contrasts with the final lengthening in the preceding utterance,
making the location and the strength of the boundary additionally clear. This latter is clearly
a redundant cue, since we know that there are strong links between final lengthening and pausing, and that pauses are almost certain to occur in this kind of context (Lehiste 1976 *inter alia*). However, this redundancy once again contributes to ease of understanding by means of increasing the recoverability of the relevant information about topic structure.

If we define the role of prosody in general as constraining contexts (cf. House 2006), then the need for a rich signalling system becomes even clearer, since looking in-depth at speech we realize that there are at least two relevant contexts. There is the context of the background information in the discourse, or the language-internal context; this is the context which defines pronoun antecedents, for instance. There is also the deictic context, that is, the context in which the discourse physically occurs, whether it is a person speaking on stage to an audience, or two friends sitting in a coffeeshop chatting. Prosody constrains both kinds of contexts. As we have seen, phrasing and probably topic structure limits the scope of reference searches both forwards and backwards within a discourse, helping the listener to make inferences about what is intended. Prosody can also constrain the deictic context, for example in interactional cues, by signalling when the speaker is finishing a turn versus intending to continue speaking. Since there are so many potential uses, and they occur simultaneously, it behoves a linguistic system to make use of as many cues as possible. For example, a late F0 peak may signal a rising pitch accent rather than a falling one. However, the linguistic system may also use a late F0 peak to signal a perceptually high peak, a pattern that might be adopted during phonologization of a contrast (Gussenhoven 1999). If these two cues conflict in terms of their interpretations, it is beneficial to the system to have supporting cues in other prosodic domains; increased loudness could increase the prominence of the pitch accent while maintaining its identity, while changes in speech rate could potentially affect how the accent category is perceived from where the F0 peak is located (cf. Niebuhr 2009; this is also relevant to the interpretation of Ladd et al.’s 1999 results, who found only minor variations in physical alignment under different speech rates, but who did not confirm perceptually that the same alignment corresponded to the same interpretation).

Therefore, at the same time as we recognize the value of multiple cues providing the same prosodic meanings (or meanings similar enough to allow the listener to make the same implicatures), it is important to remember that the cues will interact in production
and perception. The fact of this interaction has already been established in a large body of research, both in the realm of the psychology of hearing and in the realm of linguistics. It is well-known that variations in amplitude can affect the perception of the pitch of level tones, with pitch below 2 kHz decreasing by 1-5% as sound level in dB increases, and pitch above 4 kHz increasing (Stevens 1935; Terhardt 1974; Verschuure & van Meeteren 1975; see review in Moore 2003). In terms of the perception of intonational prominence, this means that we must account for the interaction of increased loudness on stressed syllables with the presence of a pitch peak. Depending on the location of the peak (i.e. within the syllable or after it), the increased loudness of the stressed syllable could increase or decrease the perceived size of the pitch movement. Although the variations in pitch based on amplitude are not very great, a variation of 5% from the value we might calculate from the F0 should be beyond the just-noticeable difference threshold for pitch variation in the range of linguistic intonation (see Moore 2003:198 for a chart of different findings for JNDs). It is important to note, however, that this research was conducted on pure tones, without a linguistic context, and therefore it is not possible to be certain to what extent the results are directly applicable to speech perception.

Work by Niebuhr (2009) has dealt with a different phenomenon of the combination of prosodic cues in a linguistic context in German. He found that if a rhythmic context was created in which listeners came to expect pitch accents to be timed a certain way relative to the segmental structure, then the interpretation of a final pitch accent could be changed. An F0 contour normally interpreted as a member of the medial accent category could be heard as early or late depending on the other accents in the sequence; if all of the accents were relatively early, the final accent in medial position would be perceived as late because it would be late in the context, even though it physically occurred in medial position. This finding again challenges the notion of a one-to-one relationship between individual acoustic measurements such as F0 and the percept of prosodic features. This particular study is additionally relevant since speakers do in fact appear to set up rhythmic patterns in speech; in Conversation Analysis studies it is possible to observe interlocutors maintaining cohesion with one another by maintaining rhythmic patterns across turn changes (Couper-Kuhlen & Selting 1996). Thus, it is not simply the prosodic characteristics, but these characteristics in relation to the context, which shape the perceptual effects.
One additional motivation for investigating cues in concert rather than in isolation only
comes from a comparison of the results of the perception experiment pilot compared to the
main experiment (section 8.2.2.2). In one of the pilot studies, utterances with naturally-
produced boundaries were used; that is, the speaker was asked to read the sentences as
though there was a paragraph boundary in the location indicated. In the main experiment,
however, the utterances underwent pitch resynthesis so that the only factor that varied be-
tween the topic-change and the topic-hold utterances was the size of the F0 fall at the first
pitch accent. The goal of the experiment was, of course, to identify the role the F0 fall
played in signalling topic change, and it was found to be a useful perceptual cue. However,
in the pilot study, where more cues were available (and in particular non-F0 cues such as
speech rate variations and/or pausing), the effect size was much larger when there was a
mismatch between the prosodic realization and the content of the following sentence, and
there appeared to be much more consistency between the listeners in the pilot study. This
suggests that the topic structure cues investigated here have a cumulative effect. The varia-
tion in the size of the F0 fall may have been just sufficient to affect listeners’ processing of
the discourse structure as containing a topic change or not, but in normal spoken language
the system is redundant, making the cueing unmissable. With only one cue, listeners might
have needed to be cautious about whether there had actually been a topic change signalled
or not, but in the pilot study, where there were multiple cues, it made more sense in terms
of ease of processing for listeners to commit to the interpretation signalled by the prosody
(cf. Vaissière 2005; Post et al. 2007).

9.4 Implications for methodology

One of the goals of this thesis was to use new methodologies to investigate the prosody of
topic structure in both production and perception. Previous production studies had tended to
either lack experimental controls for phonetic effects on prosody, ignore non-F0 variation,
or else have an insufficiently sophisticated model of topic structure, problems which the
current production study’s methodology addressed. In terms of perception studies, while
listening paradigms have been used to identify prosodic variation co-occurring with topic
structure boundaries, the experiment reported in chapter 8 is the first, to my knowledge,
which investigates the role of topic structure using online measures of language processing.

9.4.1 Production study

9.4.1.1 Experimental control

Studies of topic structure prosody have tended to center on the observation of phenomena in spontaneous discourses rather than controlled experimental texts. The advantage of this kind of study is that the experimenters can be certain that the phenomena observed actually occur in spontaneous (or non-laboratory) speech, rather than being artifacts of the way an experiment is designed, or pressure felt by participants in an experimental setting; they are likely to be aspects of conversation rather than performance. However, from a phonetic point of view this is less than ideal. It is well-known that segmental characteristics can influence the perception of F0. For example, different vowels have different “intrinsic pitches”, meaning that a close vowel will sound lower than an open vowel with the same F0; two F0 peaks which appear to be the same may be subtly different from a perceptual point of view. Furthermore, different vowels have intrinsically different lengths, which means that the timing of F0 peaks relative to the vowel onset may be different for different vowels, regardless of the intonational category (Schepman et al. 2006). This latter, in particular, is important to control for, since variations in F0 peak timing are often extremely small and within the range available for variation by different vowel length. Wichmann’s (2000) study, which could not use segmentally comparable tokens, would not have been able to identify patterns of variation like this.

Some previous studies (Herman 2000; Braun 2006) have used a paradigm in which the same sentences occur in a variety of positions in different paragraphs in order to compare prosodic realization in different contexts across longer discourses. This paradigm has the advantage of being able to compare the same (more or less) phonetic content across a variety of paragraph positions. It suffers from two weaknesses, however. The first is that constructing paragraphs in which the same sentence occurs in different locations is not necessarily

---

7 Although it is important to note that many of these “spontaneous” discourses were in fact collected in experimental or otherwise controlled settings. Swerts and colleagues (Swerts & Geluykens 1993, Swerts et al. 1994, Swerts 1997) used games to elicit natural speech in a lab setting, as did Nakajima & Allen (1993); Wichmann (2000) used a corpus of BBC newsreaders, which is probably better considered semi-spontaneous, and which in addition can be assumed to have some specific stylistic characteristics that might not be appropriate in other contexts.
the same as addressing different topic structures (although the studies mentioned above do not explicitly make this claim). This is particularly true if we assume a category-based model of topic structure, in which the relationship to preceding utterances, in addition to the position in a group of utterances, is relevant. The second weakness is more problematic from a phonetic point of view. Although the same utterance is being compared across contexts and therefore the overall prosodic course of the utterance can be compared, it is difficult to compare realizations of smaller units. For instance, comparing individual pitch accents, particularly those in different locations, may be problematic. Using the same sentence in different locations means by definition that any given word is in the same position in the sentence in all tokens. Therefore this paradigm makes it possible to compare the same word in initial position in an utterance at the beginning, middle and end of a discourse, but it is not possible to compare the same word in initial and final position in an utterance, or the realization of an accent with different stress patterns. This issue became relevant in the anacrusis versus non-anacrusis condition in the production experiment, since the presence or absence of unstressed syllables before the accent-bearing stressed one affected the alignment of the F0 peak in relation to the target vowel.

The methodology used in the production experiment reported in this thesis accounts for these phonetic issues by using the same or comparable target words in a variety of positions throughout the discourse, and furthermore by using multiple target words so that patterns could be investigated both within and across the target items. The use of the same target words also made other characteristics easier to investigate: for example, a rough measure of speech rate could be calculated simply by means of comparing the duration of the same word in different contexts. Using the target words also made it possible to rule out variation in certain locations; for example, utterance-final words did not show variation for most of the F0 characteristics studied.\(^8\)

The methodology employed in the production study also suffered from some weaknesses. In particular, finding target words which could be repeated in the appropriate contexts was extremely difficult (although using character names made this a great deal easier), and some readers noticed the repetition of some of the target words (particularly “mini-

---

\(^8\)This does not contradict Herman’s (2000) findings that the ends of utterances do show variation related to the discourse topic structure, since the studies reported here investigated different characteristics.
mum/minimal”). Similarly, repeating the target words in this way affected the style of the text, the language of which was a great deal more formal than conversational speech. The task was a reading task, not a conversational/interactional task, and therefore the generalizability of the results to different domains must be carefully examined.

The advantage of the highly controlled study, however, is that it provides a valuable methodological resource in that it allows the experimenter to rule out extraneous sources of phonetic variation. This is extremely important in prosodic research because we know that not all prosodic variation is linguistically meaningful, and the sources of this variation are not very well understood. Furthermore, many characteristics of the underlying segmental stream can influence the production of prosody. The controlled study was able to essentially eliminate F0 peak alignment from the catalog of cues to topic structure (at least as a direct cue), showing that the previously reported result from Wichmann (2000) was likely to be an artifact of segmental-phonetic or intonational variability in the data. Similarly, the controlled study meant that a much smaller set of data could be collected and compared meaningfully using statistical tools.

Future production studies ought to take into account both the controlled aspects of the production experiment reported in this thesis, and the natural aspects of spontaneous speech. Using a controlled study allows us to narrow down what kind of phenomena it would be most beneficial to examine in spontaneous speech, because it makes it possible to eliminate alternative sources for those kinds of prosodic variation. An ideal methodology would begin with results from a controlled production study and then apply them to spontaneous speech; so, for example, a follow-on study from the current production study could investigate F0 fall sizes and speech rate phenomena in a corpus of spontaneous speech. The use of the controlled experiment allows the elimination of factors which may be observable in spontaneous speech, but which are ultimately extraneous to the topic structure signalling.

The results of the text validation, particularly with regards to the degree of individual variation, also have implications for experimental design. At least two possible approaches to text design may be considered in light of these results. First, a text could be designed much as in the original production experiment (cf. section 3.2.2), and validated prior to any production data being collected, with majority votes being used to assign a topic structure. This kind of design would allow for a similar degree of comparison between items as in the
production experiment reported here, but with some added noise in the data since readers would certainly vary in their interpretations from the majority vote on the structure. Alternatively, each reader could be asked to assign their own structure to the text, and production data could be analyzed for each structural characterization. While this would provide the most robust results from a structural point of view, the important contribution of variability from phonetic sources (e.g. segmental structure or presence/absence of anacrusis) might be obscured in this case. However, either type of methodology would improve the validity, from a discourse-theoretical point of view, of data regarding topic-structure-related phonetic variation.

9.4.1.2 Phonologically-aware analysis

The other aspect of the production study which has been relatively uninvestigated until this point is the role of phonological, rather than solely phonetic, information in cueing topic structure. Previous studies have often ignored the phonological structure, or have only considered very basic contrasts e.g. observing utterances ending low versus utterances ending high (e.g. Swerts et al. 1994). Even studies which have made a more detailed phonological analysis, such as that by Herman (2000), have tended to use this analysis to narrow down the set of data to be observed, rather than taking the phonological categories into account as part of the signalling system. The sub-investigation reported in chapter 4, however, shows the importance of taking phonology into account: the specific choice of phonological system leads to important differences when modelling the prosodic phenomena observed in relation to topic structure. As discussed in that chapter, the choice of phonological model both (a) suggests whether or not there is enough “cognitive space” around the location of the F0 peak to allow it to carry meaning about topic structure, and (b) indicates whether variation of the peak location should be considered in terms of the intonational phonology versus the discourse structure. Using a phonological model helps to ensure that comparisons are made between things that are actually comparable, eliminating another potential source of confounds when we investigate prosodic variation. Of course, different phonological models will not make the same predictions, but in this case the data allow us to compare the two models to see how their results differ in the same interpretative context. Similarly, the evidence that topic structure is cued most strongly through variation in the distance between
the F0 peak and a following valley could be seen as support for interpreting this pitch accent as a fall rather than simply a peak or a rise.

An alternative account of this data, of course, could be that the relevant information is the relationship between consecutive pitch accents. However, this requires that the second bitonal pitch accent always be realized with a preceding L, which has been shown in the literature to be problematic (cf. Atterer & Ladd 2003). The cases studied were only those in which there was an unambiguous valley phonetically, but assuming that the relevant topic structure cue is the difference between consecutive pitch accents, we would expect that the low valley would need to be present in all instances of a given accent category in order to signal this difference. Whether or not this is what would be predicted by the ToBI analysis is unclear. From the point of view of maintaining a signalling system that is as simple as possible while cueing all necessary contrasts, it makes sense to have these cues either globally across the utterance, or else present in linguistically relevant locations, as noted by Herman (2000) and discussed above in section 9.3.2. If variation in the size of the F0 fall is a single cue, and not part of a global trend (which does not appear to be confirmed by the data), then it would seem that the interpretation of the fall as being a single linguistic element is a better interpretation than considering it to be a boundary across two elements; in this sense the phonological interpretation of English intonation provided by Gussenhoven (1984, 2004) would appear to be a better explanation for the phenomena observed than, for example, the ToBI analysis.9 The F0 variations observed would also fit into a contour-based theory, since the variations appear to have as their phonetic scope a whole pitch movement, rather than a single point in time (although note that in contrast, the “grouping” feature of high pitch reset appears to influence only the initial F0 peak, unless we include the raising of the bottom line from the speaker’s baseline pitch). The evidence for the varying phonological theories is suggestive rather than definitive, and it is beyond the scope of this thesis to make this type of claim about the intonational phonology of English. However, the

---

9A possible alternative, suggested to me by Mariapaola D’Imperio (personal communication), is that the L valley in these cases falls at the end of an intermediate phrase, and is an intermediate boundary tone. The scope of the F0 signal for topic structure would then be the first intermediate phrase of the utterance, rather than the first pitch accent. However, it is difficult to identify other prosodic characteristics of this intermediate phrase boundary in my data; there was no evidence of final lengthening beyond what would be expected word-finally, and in many cases, the location of the L valley was not in a location that would appear to be propitious for a phrase boundary in general (e.g. in the middle of a prepositional phrase). Still, this idea is worth consideration from the point of view of integrating this data into the ToBI framework, and it would capture the scope of the prosodic variation for topic structure in an acceptably simple way.
need for phonetic studies of prosody to take into account phonological variation is clear.

9.4.2 Perception study

Besides investigating the production patterns in language, it is essential to investigate how those patterns are used by listeners. Sacks (1984) claimed that the most important way to show the relevance of prosodic phenomena to language is by demonstrating their relevance to participants in an interaction; that is, linguistic phenomena worth studying are all and only those which interlocutors attend to in interactions.

The perception of language is also interesting from the point of view of psychology; that is, how listeners make use of what comes to them in an acoustic signal in order to come up with a linguistic interpretation which they can respond to. In this case we are interested in what the listener attends to from the perspective of interpretation only, without requiring a response. However, in some ways, the reason for this is the same as that put forward by Sacks and other conversation analysts: the acoustic signal is only relevant insofar as listeners use it to create meaning. Therefore, the differing approaches of CA and psycholinguistics can be said to converge on the same goal, though their methodologies are in some ways diametrically opposed; psycholinguistic studies are quantitative and aim to control extraneous factors carefully, while CA studies are qualitative and allow for a great deal of free variation. In fact, this is the same opposition we saw in the context of the production study above: it is relevant both to be able to isolate individual phenomena for study as well as to be able to place these phenomena in the context of how speakers and listeners behave in non-laboratory settings.

The perception study, therefore, takes the same approach as the production study, by starting from a controlled experimental environment, and leaving aside for the moment the investigation of the prosodic phenomena in a conversational setting. However, in the case of this perception experiment, we are still very interested in what listeners naturally do. For this reason, the attempt was made to use tasks which did not involve a metalinguistic assessment of the topic structure or even simply the grouping, as previous listening studies did (e.g. Swerts 1997). From a methodological point of view, this turns out to be an extremely difficult task, because there are multiple factors, not only prosodic ones, which contribute to the interpretation of the topic structure, and in fact, it was possibly the most meta-linguistic
task which was the only one to show an effect in this experiment (i.e. the sentence acceptability rating). However, since the measured effects are in the speed of response rather than in the pattern of responses, it seems that the measure was still a measure of how listeners processed the utterances naturally, rather than from the requirements of an unusual task.

The fact that listeners are responding to topic structure cues means that it is important to take these into account when we investigate prosody in general. It has been shown that listeners interpret isolated sentences differently than utterances in longer contexts (cf. Sluijter & Terken 1993), and this is probably in part due to the expectations they form about which cues will be relevant. This means that ignoring the context in which utterances are placed in normal conversation will inevitably miss some of the ways in which prosodic cues are used. Future experiments should take into account the context of utterances both in a discourse sense (i.e. longer groups) and in an interactional sense. The experiment reported in chapter 8 took a first step towards doing this by attempting to identify how listeners make sense of utterances in context from a cognitive point of view, but it was only able to make indirect claims about how the prosodic cues might be relevant in interaction. That is, the experiment demonstrated that listeners are sensitive to topic-structure cueing when listening for understanding (rather than for a meta-linguistic grouping task), and therefore we can expect to see some of these phenomena at work in casual speech. However, this is a prediction that has yet to be addressed in an interactional context.

The perception experiment also shows both some of the benefits and the difficulties of applying psycholinguistic paradigms to longer units of speech. A vast amount of psycholinguistic research has been done on individual sound segments or words, at best extending to the level of the sentence. Even in sentence-based paradigms, such as self-paced reading, there is still a focus on individual words; for instance, commonly measured effects include the speed of reading a given word, or the speed of response at a specific point in time. (This was also the case in the pause detection task in the perception experiment.) These measures can be useful in the sense that they are easy to compare; the amount of time it takes to read the same word should vary on the basis of the context only. However, they are ultimately not able to tell us very much about the overall process of language comprehension, because they are not able to take into account longer-term processes. Furthermore, many commonly-used paradigms are difficult to apply to spoken language as opposed to written language, mak-
ing it more challenging to study listeners’ responses to different prosodic phenomena; the resulting unnaturalness is likely to impact spoken language far more than written language. Things like beeps or pauses which can be used to create a response stimulus in experiments are things that outside of the laboratory have other interpretations. All of this means that it is necessary to revise our view of appropriate paradigms for studying speech in context.

Although CA has tended towards qualitative, rather than quantitative, analyses, its methods lend themselves to reinterpretation into a more controlled, quantitative type of experiment. If a laboratory experiment identifies specific prosodic features to be investigated, studies of natural discourse with CA methods (i.e. investigating how interlocutors respond to these phenomena) could be used to make clearer the ways in which these cues are interpreted, and also what kinds of combinations of cues are regularly present (cf. Post & Nolan, in press). The question of prosodic meaning could therefore be approached from two angles: both the cognitive angle, through psycholinguistics, and the interactional angle, through CA.

9.5 Conclusion

The research reported in this thesis has important implications for both the study of discourse topic structure and the study of prosody. The amount of prosodic variation available suggests constraints on the kinds of discourse structure variation that may be posited.

Although the set of cues investigated here is far from exhaustive, a unified picture of consistently-available topic structure (or general discourse structure) cueing is already emerging, with both local and global prosodic characteristics of utterances providing information that the listener can and apparently does take advantage of. The occurrence of many of these variations at linguistically relevant points suggests that they are not simply side effects of biological imperatives such as the Biological Codes, but properly part of the linguistic system, modifying the output of the intonational phonology to provide additional cues layered on top of the already rich phrase- and utterance-level prosodic signalling.

The rich system of prosodic signalling and the interaction of cues in production and per-

\footnote{This argument could of course be made for reading-based tasks as well, but we are more likely to encounter written language in unusual forms, particularly with exposure to the television and internet; the kinds of interference encountered in reading tasks are not necessarily things that would be meaningfully used to change written language, whereas the kinds of interference used in listening tasks often are.}
ception demonstrate the importance of investigating variation in multiple prosodic cues as a part of ongoing research into the intonational and prosodic structure of language. An interdisciplinary approach, combining insights from phonetics, phonology, conversation analysis, and psycholinguistics, and developing novel methodologies on the basis of these insights, allows for a more detailed investigation of these issues than has previously been possible. Future research should take into account the many possible sources of prosodic variation in order to arrive at clearer conclusions about linguistic structure in these areas.
Works Cited


WORKS CITED


Appendix I: Emory Story

This is the tale of William Emory. He was a brave and loyal man who sought adventure and knowledge. He made great contributions to science, yet history has forgotten him.

**Topic** – Emory was born in eighteen-ten. **Addition** – It was a year that would live in memory. **Elaboration** – The Napoleonic wars continued in Europe, **Continuation** – annually seeming to gain momentum. **Addition** – Meanwhile in England George III was declared insane.

**Topic** – Emory had a conventional youth. **Addition** – His mother died when he was young **Continuation** – and his father raised him and six siblings. **Addition** – They grew up in a cottage near Winchester. **Elaboration** – Annual rent was due on the cottage.

**Topic** – The family was very poor, **Continuation** – and so Emory had to work. **Elaboration** – But his father kept it to a minimum. **Addition** – Emory still had little time to play. **Elaboration** – He worked continually to support his family.

**Topic** – Manual labor bored him terribly, **Continuation** – and he longed for a more fulfilling life. **Addition** – He played mental games while he worked, **Continuation** – to try to test himself on his memory. **Addition** – He also dreamed of exploring the Amazon.

**Topic** – He knew as an explorer he should also be a scientist. **Elaboration** – In the Amazon he would investigate many new things. **Addition** – He became curious about everything he saw, **Continuation** – loyally devoting his energy to science. **Addition** - His discoveries were applauded by his family.

**Topic** – At an annual competition he would prove his worth. **Addition** – At 10 he won a contest to invent a liniment. **Addition** – At 12 he discovered a new local mineral. **Addition**
And an animal-taming project won another prize. Addition – After that he was asked to stop entering.

Locally he was now well-known. His healing liniment became popular, especially among area farmers. They used it to soothe aches and pains both in their cattle and on themselves.

**Topic** – University was always a desire for Emory. **Elaboration** – He was sure that without an education, **Continuation** – minimal opportunities would be in his reach. **Addition** – To this end, he studied constantly. **Elaboration** – At a minimum he read a book a week.

**Topic** – He made friends with the parish priest. **Addition** – The father traded Emory lessons **Continuation** – for some annual help in his garden. **Addition** – He also promised to aid Emory **Continuation** – by paying his university fees.

**Topic** – At 17 the young man left for Cambridge. **Elaboration** – Emory was finally following his dream. **Addition** – He later fondly recalled the journey. **Elaboration** – Because he was poor he had to walk, **Continuation** – and he memorized every step on the way.

**Topic** – Emory was in awe when he reached Cambridge. **Elaboration** – To him it seemed to be royally appointed, **Continuation** – and it would be like a beacon in his memory. **Addition** – The King’s College chapel inspired him, **Continuation** – minimal though his chances to visit were.

**Topic** – But Emory was disappointed by university. **Elaboration** – Annual examinations were dull, **Continuation** – and the minimal effort required was vexing. **Addition** – He had enemies among the students as well. **Elaboration** – His impoverished past was a basis for jokes.

**Topic** – Still the young man remained mannerly. **Elaboration** – He tried to be kind to everyone. **Addition** – When an animal escaped from a laboratory, **Continuation** – he spent hours helping to chase it down. **Addition** – He was annually named “Most Helpful Student.”

**Topic** – His life changed drastically one April day. **Addition** – A visitor came to speak at his college. **Elaboration** – Mr. Rinnering was an explorer and voyager. **Elaboration** – In the Amazon he had discovered many things. **Addition** – Emory was determined to join his team.

**Topic** – Rinnering wasn’t looking for more help. **Addition** – His journeys always had special teams. **Elaboration** – If a man or a youth wanted to join, **Continuation** – he needed
very particular qualifications. *Elaboration* – At the moment he had everyone he needed.

*Topic* – Emory was determined to go on the journey, *Continuation* – so he made a list of his skills for Rinnering. *Elaboration* – Mineral research was a unique ability. *Elaboration* – Emory was well-informed in that field. *Elaboration* – He was sure it must be useful in the Amazon.

*Topic* – He needed an introduction to Rinnering. *Elaboration* – If the explorer were to take him seriously, *Continuation* – he required a good recommendation. *Elaboration* – A few kind words were a bare minimum. *Addition* – Rinnering would not be easily impressed.

*Topic* – An important professor organized the meeting *Continuation* – so that Rinnering was sure to come. *Addition* – Emory would be there as though by chance. *Addition* – With his mineral research in hand *Continuation* – Emory waited with anticipation.

*Topic* – When Rinnering arrived that day, *Continuation* – animals and minerals of the Amazon *Continuation* – were already under discussion. *Addition* – Emory got to demonstrate his knowledge *Continuation* – as well as his good work ethic. *Elaboration* – In the Amazon there were mysteries to be solved, *Continuation* – and there would be plenty of need for Emory.

On a sunny day they departed from Bristol. Their ship, the Zinnia, was newly made. She was not very big, but she was fast. They hoped to arrive in thirty days, or thirty-five if the winds were bad.

*Topic* – On the voyage there was time for practical study. *Addition* – Rinnering pored over maps and charts. *Addition* – Other men prepared for capturing animals. *Addition* – Emory reinvented his prizewinning liniment. *Elaboration* – Minimal volume with maximum effect was his goal.

*Topic* – He felt quite lonely on the journey. *Elaboration* – Mr. Emory was ignored by the others, *Continuation* – who knew each other already. *Elaboration* – They mostly left him to himself, *Continuation* – and assumed that he was ignorant.

*Topic* – The ocean was calm until the last week. *Addition* – Then a storm troubled the voyagers. *Addition* – Every day was a battle to survive. *Elaboration* – The storm became a terrible enemy. *Continuation* – in the manner of a hungry monster.
Topic – The explorers finally reached land safely. Addition – Emory was elated to be in the Amazon. Elaboration – The storm they had survived left his memory. Addition – Animals and strange plants surrounded him. Elaboration – It was a whole new world to Emory.

Topic – They needed to set up a campsite for safety. Elaboration – Rinnering knew the forest was their enemy. Addition – At a minimum they needed to build a fire Continuation – so that animals would stay away. Addition – Emory and the others started to work.

Topic – Suddenly there was a huge commotion. Elaboration – Their companions raced back towards them. Elaboration – Animals also flooded the camp. Addition – Rinnering gave orders to stay together. Elaboration – Enemy creatures might be afoot.

Topic – Soon they saw the source of the trouble Continuation – as an animal the size of a ship appeared. Elaboration – It was twenty feet high at a minimum. Addition – Leaves and branches filled its fur. Addition – There were men in a panic all round it.

Topic – Emory’s heroic behavior saved the day. Elaboration – He yelled to distract the animal Continuation – which gave the others time to escape. Elaboration – Mr. Rinnering and the other men hid Continuation – while Emory faced down the creature.

Topic – With great shouts and bellows, Continuation – Emory forced the creature away. Addition – It seemed to be very timid Continuation – for an enemy of such a great size. Elaboration – Even the snapping tree branches scared it.

Topic – Such bravery could not be ignored. Addition – Mr. Emory gained everyone’s respect Continuation – despite his youth and inexperience. Elaboration – Their behavior was more mannerly Continuation – and they welcomed his presence among them.

Topic – Emory was soon seen as an expert. Elaboration – Mr. Rinnering himself sought his advice. Addition – As the men’s respect for his knowledge grew, Continuation – liniment became quite popular. Addition – In exchange they helped him research his minerals.

Topic – Meanwhile they continued their explorations. Addition – In the Amazon there was much to discover, Continuation – and they had very limited time. Addition – Experimental tasks fell to Emory, Continuation – as he was asked to analyze new findings.

Topic – Soon the explorers needed to depart. Elaboration – It was a sad goodbye for Emory Continuation – who felt he’d only just begun many tasks. Elaboration – Amazon research clearly required more time. Addition – Mr. Rinnering promised him a return trip.
Topic – Tragedy struck on their return journey. Elaboration – A storm wrecked the Zinnia, Continuation – and many men were lost. Elaboration – Mr. Emory was fortunately saved, Continuation – but others, like Rinnering, were less lucky.

Topic – When the survivors returned to England, Continuation – the other men told of Emory’s bravery. Elaboration – With the story of the massive beast, Continuation – Emory became an instant hero. Elaboration – He even eclipsed the late Rinnering.

Topic – Mr. Emory was commended royally. Elaboration – The King himself thanked him with a speech Continuation – and an invitation to dine at the palace. Addition – A holiday was proclaimed in his honor Continuation – which children hoped would become annual.

Topic – Emory’s journeys were all the rage, Continuation – winning him fame in many places. Elaboration – Soon his name was a household word Continuation – for a win or a success in any venture. Addition – It was a bit overwhelming for Emory.

Topic – He was asked to write a book Continuation – that would contain his every memory. Elaboration – The book would be called the Adventurer’s Manual. Addition – It would include his maps of the Amazon Continuation – and a mineral guide to the region.

Topic – He also gained other financial benefit Continuation – when his healing liniment was sold. Addition – Manual production was too slow for demand. Elaboration – Liniment took days to prepare Continuation – due to the careful mixing of minerals.

Topic – He planned to return to the Amazon, Continuation – to continue his travels and studies. Elaboration – Mr. Emory’s acclaimed experiences Continuation – could continue to grow annually. Addition – Animals and land remained to be studied.

Topic – He advertized for a company of men. Addition – Amazon experts raced to join him, Continuation – just like they had followed Rinnering. Addition – He could pick and choose from among them. Elaboration – Mr. Emory’s team was the cream of the crop.

Topic – However, financing was a problem. Elaboration – World exploration was now so popular Continuation – that everyone was planning journeys. Addition – Though Emory wrote to many financiers, Continuation – Amazon work was no longer funded.

Topic – Emory was very disheartened Continuation – as the way closed to the Amazon. Elaboration – He tried to remain cheerful, Continuation – but his happiness was minimal. Addition – He felt as though his dreams were slipping away.
Appendix I

Topic – He decided to take action *Continuation* – to help other men like him. *Addition*
– He created a Travellers’ Trust *Continuation* – to support those who wished to explore.

*Elaboration* – It had annual funds available.

Alas today Emory has been forgotten. When the money for his Trust ran out, interest turned to other people, and so history cheated him of his rightful dues. It was a tragic fate for poor Emory.
Appendix II: Perception Experiment

Stimuli

Target items (24):

The numbered sentences were presented auditorily, and the last sentence is the target sentence where the prosody was manipulated. The follow-up sentences were presented visually on the screen; C = Continuation (topic-hold) and T = new Topic (topic-change).

1. He needed some food, so he went to the store. A packet on the shelf caught his eye.
C: Those biscuits would make a good snack, he thought.
T: It didn’t look like it belonged there.
Pause before: shelf

2. She was terribly bored, so she went to the park. She watched some children playing.
C: Then she walked around the pond.
T: They were climbing all over the adventure playground.
Pause before: children

3. She finished the email. She closed the laptop. She got up from the desk.
C: . . . and put on her coat.
T: As she moved she felt stiff from sitting all day.
Pause before: the

4. He got off the train. The station was packed. The crowds were blocking his way.
C: . . . and he saw that no one else could move either.
T: . . . and he was hemmed in on the other side by the platform barriers.
Pause before: blocked
5. The car broke down. It was blocking the road. The driver got out to push.
C: The other cars honked their horns.
T: He felt very embarrassed.
Pause before: push
6. She flew to Rome. The plane was packed. Everyone was going on holiday.
C: She tried to sleep, squashed in her seat.
T: They were taking advantage of the bank holiday weekend.
Pause before: going
7. She arrived at the office. She sat down at her desk. She turned on her computer.
C: She listened to her voicemail.
T: She waited impatiently for it to start.
Pause before: her
8. He opened the cupboard. He took out the cereal. He needed some milk,
C: . . .since he hated eating cereal dry.
T: ...to drink with his meal.
Pause before: milk
9. It was getting dark, so she turned on a lamp. The bulb had burnt out,
C: . . .so she put on the overhead light instead.
T: . . .which made the third time that week.
Pause before: burnt
10. There was a storm. It was very windy. A tree had been blown over.
C: Some people had lost electricity.
T: . . .and it blocked the roads.
Pause before: over
11. He asked her to dance. They waltzed onto the floor. Everyone watched them dancing.
C: They spun in circles and performed fancy dips.
T: Most of the onlookers were completely mesmerized.
Pause before: them
12. The alarm clock went off. He opened his eyes slowly. The sun was shining outside the window.
C: He reluctantly dragged himself out of bed.
T: It was a gorgeous summer day.

13. They got some food. They sat down in the cinema. The lights dimmed.

C: They munched on their popcorn.

T: . . . and the film began to run.

14. The exam was tomorrow, but he hadn’t learned the material. He started to read the textbook.

C: He thought he would never learn everything in time.

T: He labored to understand the formulas in it.

15. She had been sitting at her desk all day. She felt very restless. She decided to go for a run.

C: That would help her relax.

T: She went to put on her running gear.

16. Ten inches of snow had fallen. Everyone was outside with spades. There weren’t enough plows for the streets.

C: . . . so it was everyone for himself.

T: . . . and the city government was in a panic trying to find a way to clear them.

17. The burglars crept up to the house. All the lights were off. They decided to proceed.

C: They felt safe in the dark.

T: One got out his lock-picking tools.

18. The race was very close. Two runners crossed the line together. A photograph would determine the result.

C: The runners were both gasping for breath from their exertions.

T: Everyone waited with bated breath for it to be produced.

19. The sink was blocked. It flooded their kitchen. They spent hours cleaning it up.
C: It needed to be repaired by a plumber.
T: They used every towel in the house.

Pause before: cleaning

20. The Tube station was closed. He had to get off at the next stop. He got lost trying to walk back.
C: He eventually arrived at the party.
T: He ended up in a neighborhood he’d never seen before.

Pause before: walk

21. She packed her suitcases. She wanted to bring too many things. She couldn’t lift her bags when they were full.
C: She took half of the contents out.
T: She got her boyfriend to carry them out to the car.

Pause before: bags

22. She arrived at the hotel and she found her room. She started to unpack her bags.
C: She needed to leave soon to get to her meeting.
T: She wanted to hang her clothes in the closet.

Pause before: unpack

23. She chopped up some vegetables. She put them into the stew pot. She got the sausages out of the fridge.
C: …and added them as well.
T: They needed to thaw before she cooked them.

Pause before: of

24. He was practicing the violin. He wasn’t very good. His playing made the strings squeak.
C: He had only started lessons a few weeks ago.
T: His friends winced and covered their ears.

Pause before: strings

Filler items (26):

1. His coffee was cold. He wanted it warm, so he put it in the microwave.
   …and set the timer for a minute.
2. The rain stopped and the sun came out. Birds began to sing.
   It was a perfect spring day.

3. They skied all day. They were really tired. They slept well that night,
   In the morning the post was delivered. (Nonsense)

4. A cat came into sight, so the mice ran away. They hid under the floorboards.
   They waited until the cat was gone before they came out again.

5. They cycled to school. They locked up their bikes. The shed was mostly full.
   The janitor was raking up fallen leaves. (Nonsense)

6. She cleared up the garden. She dug up the weeds. She planted some bulbs.
   She looked forward to when they would start growing in the spring.

7. The kitchen was a mess. They found some sponges. They started to scrub.
   It took about an hour to clean everything.

8. They arrived at the house. They rang the doorbell. Their host welcomed them.
   The party was already in full swing.

9. It was quiz night at the pub. They got into two teams. They raced to answer the questions.
   In the end, they had almost the same number of correct answers.

10. The children were playing. They had a tree house. They climbed up the rope ladder.
    The tree was an oak tree. (Nonsense)

11. The sun was shining brightly. They decided to go to the beach. They packed swimming
    costumes, towels and a picnic basket.
    They spent the whole day swimming in the sea.

12. The mice wanted the cheese. They scurried out of their nest to sniff it. The cat watched
    them from nearby.
    It suddenly pounced and the mice scattered.

13. The bicycle had been in the shed all winter. It needed some maintenance work done.
    The back tire was flat.
    The gear chain was rusty, too.

14. His flight arrived late. All the hotels in town were completely full. He phoned a friend
    who lived nearby.
    Fortunately, his friend offered him a place to stay.

15. She sketched the scene. It needed some color, so she got out her oil paints.
and chose a bright red for the flowers.

16. The show ended late. They went to a café afterwards, and they both ordered espresso.
   The waiter told them about his daughters. (Nonsense)

17. The game went into overtime. Both teams were exhausted. A goal was scored six minutes in.
   Fans of the scoring team cheered.

18. They went on a safari. They saw hundreds of animals. They saw a lion hunting an antelope.
   They also saw a herd of zebras.

19. There was a surprise party. Everyone was hidden. The guest of honor walked in the door.
   Everyone jumped out and yelled, “Surprise!”

20. Her team had made it to the Olympics. The match was about to start. She had never been this nervous.
   She breathed deeply to calm down.

21. His mobile phone was ringing. He couldn’t find it. It was in the bottom of his rucksack.
   By the time he found it, he’d missed the call.

22. He opened another program. The computer crashed, so he rebooted it and started again.
   Unfortunately the same thing happened again.

23. The cinema doors opened at 7 sharp. People flooded inside. Everyone raced to get to the best seats.
   They wanted to go home as soon as possible. (Nonsense)

24. She walked home through the park. Dark clouds were gathering. It started to rain.
   She wished she had remembered her umbrella.

25. The library was very quiet. She could hear people turning pages. She could concentrate on studying there.
   She turned on the TV. (Nonsense)

26. The government proposed a new tax. They put it to a referendum. Many people went to vote.
   The tax was soundly defeated.
Appendix III: Sound Files

A set of sample sound files from the production and perception experiments are available online at http://www.srcf.ucam.org/~mkz21/Sound_Files.

- Emory text F01: speaker F01 (female) reads the Emory story. 7 files.
- Emory text M02: speaker M02 (male) reads the Emory story. 7 files.
- Target stimuli: resynthesized target items (with pause) in both topic-change (T, with high and low sets) and topic-continuation (C) forms. 72 files.