Discourse Prosody in Production and Perception

by

Joseph C. Tyler

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Linguistics)
in The University of Michigan
2012

Doctoral Committee:

Assistant Professor Ezra Russell Keshet, Chair
Professor Julie E. Boland
Associate Professor Robin M. Queen
Lecturer Hannah Rohde, University of Edinburgh

Acknowledgments

I'd like to start by thanking my dissertation committee: Ezra Keshet, Robin Queen, Julie Boland and Hannah Rohde. They have been a great team. My thanks begin with Ezra Keshet, my dissertation chair, who has been a wonderful advisor. He has helped me navigate diverse topics and concerns, making new connections and keeping each task in perspective. It was with great joy that I walked into our meetings, excited to discuss new data or a new analysis, to ask questions, and see where our discussion takes us.

Robin Queen has worked with me from the very beginning of my time in graduate school. I feel lucky to have gained from her insights into my work as well as how to work. Julie Boland has provided critical practical advice at many stages of my research, helping me make decisions about research design and implementation. She has welcomed me into the psycholinguistics community and helped introduce me to new topics and questions. And Hannah Rohde has been a great advisor, bringing insight and analysis from outside Ann Arbor, asking questions and making connections between my work and new areas.

A special acknowledgment is also due to Jennifer Arnold and Jason Kahn at The University of North Carolina, Chapel Hill. They have been an important source of feedback and inspiration for my work and a sounding board for ideas. We have developed and run projects together. I give them my sincere appreciation.

In addition, I am grateful for the larger community at Michigan, which has provided a climate of linguistic exploration in many areas. This has of course involved the Linguistics department, but also Philosophy, Psychology and Anthropology. While I cannot thank everyone individually, I'd like to specifically mention Andries Coetzee, Sam Epstein, Steve Abney, Barb Meek, Deborah Keller-Cohen, Anne-Michelle Tessier, Sally Thomason and Rick Lewis. My fellow graduate

students have also been integral to my development, and for this I am grateful. Special thanks to Anna Babel, Chris Odato, Lauren Squires, Brook Hefright, Eric Brown, David Medeiros, Jon Yip, Erica Beck, Harim Kwon, Mike Opper, Tim Chou, Terrence Szymanski and Stephen Tyndall. You've all helped make grad school rich, rewarding and fun.

And, of course, my friends and family beyond the university. My parents Ilene and Norm Tyler have been a solid support as I navigate grad school, listening to me recount developments at different stages. And my cousin Lori Burkall has been a wonderful listener and friend. And dear friends, for your ability to keep me grounded in things besides school: Jon Humphrey, Sasha Kimel, Mary Liu, Jason Krol, Julia Raskin, Alex Lee, Sarms Jabra and more.

Thank you to everyone who has been there for me over these years in graduate school.

Table of Contents

List of Tables

List of Figures

List of Appendices

Abstract

A well-formed discourse is more than just a series of well-formed sentences. While often left implicit, this structure to discourse is sometimes overtly cued. And though most attention in this area has focused on lexicalized cues like discourse markers, prosody can also convey information about the structure of discourse. This dissertation explores the relationship between prosody and discourse in production and perception, helping to identify what information about the structure of discourse is in speakers' prosody and what prosodic variation listeners use in discourse interpretation.

First, a production study examines prosodic correlates of discourse structure in readings of a newspaper article. Prosodic measures of pause duration, pitch, intensity and speech rate were correlated with discourse structural measures of boundary size, discourse coordination/subordination, and their interaction. The prosodic measures were correlated with both structural measures and their interaction. This interaction shows that the effect of boundary size on an utterance's prosody often depends on whether that utterance is coordinated or subordinated, and vice versa.

Second, a series of perception studies examine the ability of synthesized manipulations of prosody to bias the interpretation of ambiguous discourse. For example, the discourse "I sat in on a history class. I read about housing prices. And I watched a cool documentary" could be interpreted as describing three separate, independent events (coordinated interpretation) or that the events of the second and third sentences took place during the event of the first (subordinated interpretation). Rising pitch at the end of the first sentence led to more coordinated interpretations compared to falling pitch.

These results suggest that one meaning for rising pitch can be to mark coordination in discourse. This proposal is motivated by research on listing intonation. The potentially contradictory claim by Pierrehumbert & Hirschberg (1990) that high terminal pitch indicates elaboration, a subordinating relation, is discussed and re-analyzed to bring their data in line with these results. Finally, these results are discussed with respect to prosodic disambiguation of syntax, and comparisons are made between prosodic disambiguation of syntactic and discourse structures.

Chapter 1

Introduction

Language is clearly structured in many different ways. Established areas of linguistics have for decades studied the systematic organization of sounds (phonology) and parts of a sentence (syntax). Similarly, the sentences of a discourse are structured, and a well-formed discourse is more than just a series of well-formed sentences. But the structure of discourse may be difficult to see because of its very familiarity. One way to reveal this structure is to remove it, perhaps by re-ordering the sentences of a discourse. For instance, if you were to read the sentences of this paragraph from last to first, the resulting discourse would be quite hard to follow. Even the two possible orderings of two sentences can lead to different interpretations of the events narrated.

(1.1) John banged his head. He fell over.
(1.2) John fell over. He banged his head.

A natural interpretation of the discourse in (1.1) is that John's banging his head happened before his falling over, while a natural interpretation of (1.2) is that John first fell over and then banged his head. In addition to the temporal ordering contrast, these two discourses likely also have different causal relationships. In (1.1), the banging of his head seems likely to have caused John to fall over. In (1.2), John's falling over seems likely to have led to him to bang his head.

While it seems clear there is structure in discourse, it is less clear exactly what that structure is. Sometimes aspects of discourse structure are explicitly cued, while other times a speaker leaves the structure implicit, leaving listeners to fill in the gaps

with their own reasoning. Most work that has analyzed explicit cues to discourse structure has focused on *lexical* cues, e.g. discourse markers. If (1.1) was instead produced as (1.3), with the addition of the explicit marker of temporal succession *then*, the temporal relationship between the two sentences would be explicit.

(1.3) John banged his head. Then he fell over.

In (1.3), it is explicit that John banged his head and subsequently fell over. An alternative, though dispreferred, interpretation of (1.1) could have been that it described two separate, independent events with no information about when each happened. In this interpretation, (1.1) would describe two independent events that happened to John, banging his head and falling over. With the addition of the discourse marker *then* in (1.3), the temporal ordering is explicitly encoded and this alternative is ruled out. Thus, the addition of a lexical item like a discourse marker can make explicit how the sentences of a discourse are related.

One feature of discourse identified by many theorists (Grosz & Sidner, 1986; Hobbs, 1985; Mann & Thompson, 1988; Polanyi, 1988; Van Kuppevelt, 1995) is that it is hierarchically structured. Asher & Vieu (2005) discuss the intuitions motivating hierarchical structure in the context of Segmented Discourse Representation Theory (SDRT) (Asher & Lascarides, 2003). They mention paragraph structure as an orthographic manifestation of discourse hierarchy, where paragraph-initial sentences are in some sense higher-order than paragraph-medial sentences. A paragraph-medial sentence likely provides more detail about whatever was introduced by the paragraph-initial sentence. They also argue that temporal structure motivates a hierarchical conception of discourse. If one sentence introduces an event and a second sentence describes something occurring at the same time as that first event, the second is likely providing more detail about the first event. By contrast, if a second sentence describes an event at a different time, the two events likely have equal status.

Like most theories of discourse structure, SDRT analyzes the structure of discourse by segmenting the discourse, identifying relations that hold between segments, and constructing a hierarchy from the segments and relations. SDRT

focuses on both semantic and pragmatic information for all stages of analysis (segmentation, relation identification, hierarchy). SDRT also provides an inventory of discourse relations (e.g. ELABORATION, BACKGROUND, RESULT) that are claimed to hold between the segments of a discourse. But most importantly here, SDRT builds hierarchy in discourse by classifying all discourse relations as either coordinating or subordinating. Coordinating relations link discourse segments at an equal hierarchical level while subordinating relations link a discourse segment with another segment one hierarchical level lower.

Rhetorical Structure Theory (RST) (Mann & Thompson, 1988), like SDRT, analyzes a discourse into segments, identifies relations between segments, and constructs the discourse into a hierarchical structure. RST also has a local hierarchical structure contrast in its nucleus-satellite distinction. In RST, all discourse segments are considered to be either a nucleus or a satellite. The distinction between the two is defined in terms of a segment's relative importance to the coherence of the discourse. One diagnostic test is that satellites can be deleted without harming the overall message of the discourse, while deleting a nucleus would disrupt the discourse's coherence. This test reveals one of RST's applications: automatic text summarization. If all satellites in a text were deleted, the result would be a stripped down summary of the discourse.

While RST's nuclearity principle has been compared to SDRT's coordinating/subordinating contrast (Danlos, 2010), there are points of contrast. In RST, nuclearity is a feature of a discourse segment. This means that every discourse segment is either a nucleus or a satellite. In SDRT, coordinating and subordinating relations are theorized to hold between discourse segments, but are not strictly features of the segments themselves. This means that any one segment in an SDRT analysis could be coordinated to one segment and subordinated to another. Another difference between RST's nuclearity and SDRT's coordinating/subordinating contrast is in terms of how an analyst identifies a segment's nuclearity or CoordSubord status. In RST, a central criterion for satellite status is that a discourse segment be expendable: if it can be deleted without harming the discourse's coherence, it is a satellite. In SDRT, the main point of contrast between coordination and subordination

is in terms of the level of detail. So, RST and SDRT both supply theoretical constructs that account for local hierarchical contrasts, but the nature of those local hierarchical constructs is not exactly the same.

Another influential theory of discourse that analyzes discourse into segments, relations between segments, and hierarchy is the Grosz & Sidner model (1986). Unlike SDRT and RST, which focus on the propositional content of utterances as the basis of their analyses, the Grosz & Sidner model analyzes discourse using speaker purposes, goals and intentions. In this theory, a speaker may have one overall purpose to their discourse, e.g. to give directions on how to replace a car battery. Then, this overall purpose may be subdivided into a series of subgoals, e.g. how to identify the battery, how to remove the old battery, and how to install the new battery. Grosz & Sidner propose two structural relations that organize these discourse purposes into a hierarchical structure: dominance and satisfaction-precedence. The higher-order purpose of replacing a car battery is said to dominate the three subgoals. And since the removal of the old battery needs to be complete before the installation of the new battery begins, the purpose of the battery removal portion of the discourse is said to satisfaction-precede the purpose of the battery installation portion of the discourse. These two relations (dominance and satisfaction-precedence) therefore create contrasting hierarchical structure. Dominance relations link segments at different hierarchical levels while satisfaction-precedence relations link segments at the same hierarchical level.

The Grosz & Sidner model, RST and SDRT are all capturing ways in which discourse is segmented, how the segments are related, and how the whole is hierarchically structured. The more cues we can draw on in the speech signal, the better we can understand what that structure is and how speakers and listeners communicate it to each other. When there are no overt cues to discourse structure, listeners must draw on more general reasoning about how the sentences are likely to fit together. This was the case with (1.1) above, where a plausible interpretation involves the banging of his head causing John's falling over, even though this causal information was not explicitly asserted. And while most work on cues to discourse structure has focused on lexical cues, there is a body of research that has also

identified systematic correlates between aspects of discourses' structure and prosodic measures of pitch, pause duration, intensity and speech rate (den Ouden, Noordman, & Terken, 2009; Hirschberg & Grosz, 1992; Lehiste, 1975). This indicates that the prosody of speech can carry cues to the structure of discourse. A fuller understanding of the production and perception of discourse prosody will illuminate a non-lexical way that interlocutors communicate discourse structure to one another.

Prosodic disambiguation of discourse

A number of studies have identified prosodic correlates of discourse in speech production (den Ouden, et al., 2009; Hirschberg & Grosz, 1992; Wichmann, 2000), but relatively little work has been done on the role of prosody on the perception of discourse structure. This work has generally used indirect measures of linguistic perception like naturalness judgments (Smith, 2004) or judgments about a sentence's location in the discourse (Lehiste, 1982), e.g. is it paragraph-final or not. Only two studies have tested whether discourse prosody can specifically affect the interpretation of linguistic expressions (Mayer, Jasinskaja, & Kölsch, 2006; Silverman, 1987).

Mayer et al. (2006) ran three experiments to test whether synthesized manipulations of prosody can bias interpretation of ambiguous pronouns whose resolution indicates listeners' likely interpretation of the overall discourse. They give the example of the discourse in (1.4):

(1.4)

a. Lena was happy after the tennis tournament.
b. The silver medal was a great achievement.
c. The coach congratulated her after the award ceremony.
d. For the next tournament, however, she hopes for the first place.

In this discourse, one referent "Lena" is introduced in (1.4a) and a second referent "the coach" is introduced in (1.4c). Then, a pronoun "she" appears in (1.4d) that could corefer with either "Lena" or "the coach."

Mayer et al. (2006) exploit both discourse structure and discourse recency to account for biases in the resolution of this coreference ambiguity. The structure of the

discourse in (1.4) can be used to account for the accessibility of some antecedents instead of others, drawing on the discourse semantic concept of the *right frontier constraint* (RFC) (Asher & Lascarides, 2003; Polanyi, 1988). The idea of the RFC is that a new discourse segment can only attach to its immediately preceding discourse segment or one that dominates it, but not to previously subordinated segments. Furthermore, an anaphor in the current discourse segment can access an antecedent in the segment to which it attaches and any that dominate that segment, but not to previously subordinated segments. While some possible antecedents are deemed inaccessible depending on their place in the discourse structure, there can still be multiple possible antecedents accessible in discourse segments on this right frontier. A pronoun is ambiguous when it can access multiple antecedents on this right frontier.

In the discourse in (1.4), sentence (a) introduces Lena's happiness and sentences (b) and (c) together "present the cause of Lena's happiness" (p. 1). When a sentence β presents the causes of the contents of an earlier sentence α, then we can say that an Explanation relation holds between α and β, i.e. Explanation(α,β) (Reese, Denis, Asher, Baldridge, & Hunter, 2007). Because Explanation is a subordinating relation, sentences (b) and (c) are both subordinated to (a). And because they jointly explain (a), (b) and (c) are coordinated to each other. Therefore, before sentence (d) is uttered, sentences (a) and (c) are on the right frontier of (1.4).

When sentence (d) is uttered, it could attach into the larger discourse either at sentence (a) or sentence (c). This is a high vs. low attachment ambiguity because the structural ambiguity is in terms of whether (d) attaches to the dominating segment (a) or the subordinated segment (c). If (d) attaches to (a), then the only available antecedent is "Lena" in (a) because "the coach" in (c) is no longer on the right frontier. If (d) attaches to (c), antecedents in both (a) and (c), i.e. "Lena and "the coach," are both still accessible. Mayer et al. (2006) claim that in this low attachment scenario, listeners would prefer to resolve the pronoun as coreferring with "the coach" in (c) because it is the most *recent* antecedent (p. 1). Therefore, in cases of ambiguity, they argue that discourse recency will drive listeners' interpretation preferences.

Therefore, both discourse structure and discourse recency are exploited to motivate their hypothesis that prosody can bias the resolution of the ambiguous

pronoun. They hypothesize that one set of prosodic manipulations can bias towards high attachment, and as a result towards coreference with "Lena" in (a). Another set of prosodic manipulations is hypothesized to bias relatively more towards low attachment, in which case both antecedents are accessible; in this case, discourse recency will create a preference for the pronoun's resolution towards the most recent antecedent. Although Mayer et al. (2006) present their claims as being related to hierarchy, they do not actually need hierarchy to account for their data. The two possible antecedents of the ambiguous pronoun are fully distinguishable in terms of discourse recency, without drawing on hierarchy at all, because the high attachment antecedent is always less recent than the low attachment antecedent. For a pronoun in (d), any antecedent in (a) is always less recent than an antecedent in (c).

To motivate their claims about what prosody will bias towards which attachment site, Mayer et al. draw on the literature that has identified systematic correlates between structures of discourse and speakers' prosody. One of the most consistent findings has been that larger breaks in a discourse correlate with longer pauses, compressed pitch before and pitch reset after (for a fuller review of discourse prosody in production, see chapter 2). The discourses Mayer et al. (2006) test have a larger structural break before the fourth sentence in the high attachment interpretation and a smaller one in the low attachment interpretation. The question that Mayer et al. were testing was whether the prosodic correlates available in discourse production could bias listeners' interpretation of the structure of ambiguous discourse, which would then be visible in the resolution of an ambiguous pronoun.

In the first of three experiments, Mayer et al. (2006) manipulated pause duration and overall sentence pitch to bias towards one interpretation of their ambiguous discourses. When trying to bias towards the low attachment interpretation, the manipulations created a relatively small prosodic boundary before the fourth sentence. This was achieved by maintaining all inter-sentential pause durations equal (400ms), by making sentences 2 and 3 have normal pitch range and sentence 4 having compressed pitch range. When trying to bias towards the high attachment interpretation, the manipulations created a large prosodic boundary before the fourth sentence. This was achieved by doubling the duration of the pause before sentence 4

to 800ms, while leaving the other pauses at 400ms; pitch range was normal for sentence 2, compressed for sentence 3 and expanded for sentence 4. For all versions, pitch range was expanded for sentence 1. These manipulations were in line with cues to discourse structure identified in discourse prosody production studies. Experiments 2 and 3 were identical to Experiment 1 except that for Experiment 2 only pitch range was contrastive and for Experiment 3 only pause duration was contrastive.

Results for Experiment 1 in Mayer et al. (2006) showed a significant difference in the resolution of the ambiguous pronoun depending on the prosody the participants heard. While there was an overall preference for low attachment for either prosodic condition, high attachment responses were significantly more frequent in the high attachment prosody condition (38%) than in the low attachment prosody condition (28%). Experiments 2 and 3 continued to show more low attachment than high attachment interpretations, but no difference between the two prosodic conditions. It seems listeners needed the cues available in both pause duration and pitch range to be biased in their interpretation. In sum, when prosodic manipulations created a smaller break before (1.4d), listeners were more likely to interpret the ambiguous pronoun as coreferring with the more recent antecedent in (1.4c). When prosodic manipulations create a larger break before (1.4d), listeners are more likely to interpret the ambiguous pronoun as coreferring with the more distant antecedent in (1.4a).

A study similar to Mayer et al. (2006) was run by Silverman (1987), testing the ability of synthetically manipulated prosody to bias the interpretation of ambiguous discourse. In the two studies presented in chapter 6 of Silverman's dissertation, he used six ambiguous discourses, each of which could be disambiguated by the location of a paragraph boundary. The example he provides is the following:

(1.5) Example discourse ambiguity used in Silverman (1987).

Version 1:

"This building company offers several different schemes for double glazing.
The cheapest is acrylic sheeting. You pay by the square metre, plus the mounting clips. Installation is extra.
The most expensive systems are the "slimline" and "royal" schemes. Prices include sealed glass units, and draught-proof frames. All materials are delivered free within Cambridge.
For details of any scheme, please contact your local building store."

Version 2:

"This building company offers several different schemes for double glazing.
The cheapest is acrylic sheeting. You pay by the square metre, plus the mounting clips. Installation is extra.
The most expensive systems are the "slimline" and "royal" schemes. Prices include sealed glass units, and draught-proof frames.
All materials are delivered free within Cambridge. For details of any scheme, please contact your local building store."

The ambiguity arises in the sentence "All materials are delivered free within Cambridge," and the domain over which the quantified phrase "all materials" ranges. In version 1, that phrase is part of the paragraph about the most expensive systems, and hence the "all materials" is referring only to the most expensive systems. This is what Mayer et al. would call a low attachment. In version 2, the phrase "all materials" begins a new paragraph, and so applies to all schemes, both the cheapest and most expensive; Mayer et al. would call this a high attachment.

Because the location of a paragraph boundary disambiguates these two interpretations, Silverman can test whether prosodic cues to the location of a paragraph boundary can bias the interpretation of the phrase "all materials." The prediction is that a large prosodic boundary can indicate the location of the paragraph boundary, either before or after the target sentence, and thus bias interpretation of "all materials" as referring to either all systems (high-attachment) or just the expensive systems (low-attachment).

Silverman's spoken discourse stimuli were created with computer-generated speech synthesis, as opposed to prosodic manipulations of human productions. Silverman acknowledges the existence of "problems associated with poor segmental

quality in the synthetic speech" (1987, p. 6.20). To address this, Silverman presented participants with written transcripts of each discourse, with paragraphing removed. This way the lexical material would be unambiguous but the paragraph structure would be ambiguous. In the experimental setting, participants heard computer-generated versions of ambiguous discourses while reading along with written transcripts of each discourse.

Silverman created two prosodic conditions, one that cued a paragraph boundary for low attachment and one that cued a paragraph boundary for high attachment. In the first experiment, he cued the location of a paragraph boundary by manipulating three prosodic features: final lowering of f0, initial raising of f0, and pause duration. In the second experiment, the pause durations were held constant but the pitch manipulations were the same as before. Both experiments found participants' interpretations of the ambiguous discourses were significantly affected by the prosodic manipulations. With all three prosodic features manipulated, participants chose the predicted interpretation (averaged across both the low and high attachment conditions) 84.2% of the time, compared to 71.7% when pause duration was held constant.

Prosodic disambiguation of sentences

It will be useful to contextualize the work on prosodic disambiguation of discourse in the larger literature that has focused on prosodic disambiguation of syntactic ambiguities. Research on prosody-based disambiguation goes back to Lieberman (1967), who made a claim about which kinds of sentences prosody can disambiguate and which it cannot. Lieberman argues that only sentences with differing deep and surface structures can be prosodically disambiguated, providing the following examples:

(1.6)

a. I will move on Saturday
b. They decorated the girl with the flowers
c. Vanderburg reports open forum
d. They kept the car in the garage
e. I fed her dog biscuits

What is meant then by having different surface structures is that the different meanings have different bracketings of their constituents. As demonstrated in (1.7), the boundaries between word groupings occur in different locations.

(1.7)

a. [I will move] [on Saturday] VS [I will move on] [Saturday]
b. [They decorated] [the girl with the flowers] VS [They decorated the girl] [with the flowers]
c. [Vanderburg reports] [open forum] VS [Vanderburg] [reports open forum]
d. [They kept] [the car in the garage] VS [They kept the car] [in the garage]
e. [I fed] [her dog biscuits] VS [I fed her] [dog biscuits]

For each of these structures, the prosody can distinguish the two possible bracketings. While theories of syntax have changed dramatically since 1967, the bracketing contrasts above still accurately reflect the fact that the organization of the structural elements of these sentences is different.

By contrast, Lieberman argues that linguistic structures with different deep structures and identical surface structures cannot be disambiguated by prosody, providing the following example:

(6.2) Flying planes can be dangerous
(6.3) [Flying planes] can be dangerous VS [Flying planes] can be dangerous

This sentence has a syntactic constituent “flying planes” that can be interpreted in two ways, yielding two meanings for the sentence. On the one hand, the sentence can mean that the activity of flying planes can be dangerous. On the other hand, it can mean that planes that are flying are dangerous. The important point for our purpose is that this distinction in meaning cannot be reduced to a difference in

bracketing, since the relevant phrase remains a constituent under both readings. While more modern theories of syntax may no longer use the same terminology of deep and surface structure, the constituent boundaries in (6.2) would still have "flying planes" forming a constituent to the exclusion of the rest.

This account provided by Lieberman (1967) depends on the claim that while all linguistic ambiguities have, by definition, more than one underlying meaning, only some of those ambiguities have differing bracketing structures of their syntactic constituents. In order to provide a more general account without relying on the theory-specific terminology of surface and deep structures, I will use the term *meaning* to refer to the underlying semantic contrast and *bracketing* to refer to the grouping of syntactic constituents. The schematizations below capture this account provided by Lieberman (1967) of which structures can be prosodically disambiguated (Figure 1.1) and which cannot Figure 1.2.

Figure 1.1: Linguistic ambiguities that can be prosodically disambiguated

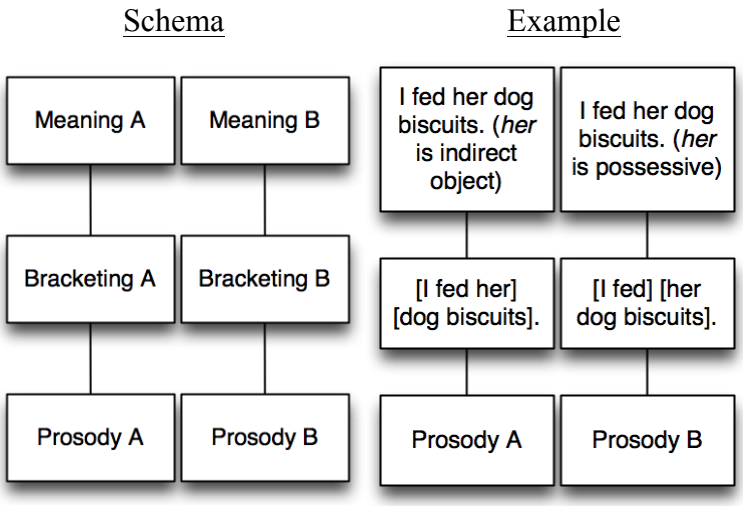
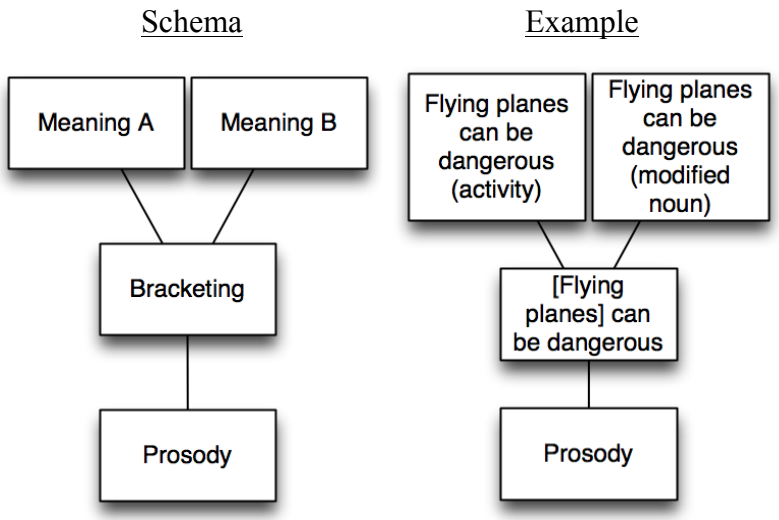


Figure 1.2: Linguistic ambiguities that cannot be prosodically disambiguated:



In Figure 1.1, two meanings map onto two distinct bracketings that then can be distinguished with contrasting prosody. In Figure 1.2, two meanings only have one bracketing, and so cannot be distinguished prosodically. For this reason, sentences that fit the schema in Figure 1.1 can be disambiguated prosodically while those that fit Figure 1.2 cannot.

Lieberman motivates his claims with illustrative, presumably representative examples grounded in his intuitions, but he does not provide experimental evidence to support them. A range of subsequent research has mostly found empirical support for the Lieberman account by testing it experimentally (e.g. Lehiste (1973), Lehiste et al. 1976, Price et al. 1991). Lehiste (1973) created a set of 15 linguistic ambiguities. Ten of those ambiguities had contrasting bracketings corresponding to each of the ambiguity's two meanings. The other five ambiguities could not be distinguished by constituency bracketing. Lehiste had four readers (two linguists, two non-linguists) read each ambiguous sentence three times. First, they produced each sentence without being told of the ambiguity, being asked subsequently which interpretation of the sentence they had. Then they were informed of the ambiguity, and were guided to say each sentence once for each meaning. These productions were then presented to 30

participants (15 linguists, 15 non-linguists) who were asked what each production of each sentence meant.

Lehiste found that all ambiguities with contrasting bracketings could be disambiguated, and all but one without contrasting bracketing could not. The one exception was the sentence "German teachers visit Greensboro," which was successfully communicated. This sentence could be referring to teachers who were from Germany, or teachers of German. Lehiste speculates that it may have been intonation that helped in this case, while the other successfully disambiguated sentences were distinguished primarily with duration contrasts. This raises the idea that pitch and pause duration may operate independently with respect to disambiguation effects.

While Lehiste's (1973) results generally support Lieberman's (1967) claims about prosodic disambiguation, they could not speak conclusively about which prosodic features are driving the disambiguation effect. With Lehiste (1973) having identified duration as a strong correlate of the two meanings, Lehiste, Olive, & Streeter (1976) sought to test whether experimentally isolating duration could show it to be a cause of the disambiguation effect. The materials in Lehiste et al. (1976) consisted of ten sentences from the original Lehiste (1973) study produced by one speaker. Seven of these ambiguities had contrasting bracketings while three did not. They then synthetically manipulated the duration of words or phrases, but not pause durations. They found that listeners could reliably retrieve the predicted meaning for all of those structures with bracketing contrasts while they could not for those sentences without bracketing contrasts. They included the sentence "German teachers visit Greensboro" in this study, which notably was not successfully disambiguated. As a result, the successful communication of this sentence in Lehiste (1973) was either due to prosodic features other than the durational contrasts tested by Lehiste et al. (1976), or the original finding was somehow unreliable.

These studies by Lehiste and colleagues successfully established an empirical basis for the general account originally provided by Lieberman (1967). But thus far, the contrast was between the whole categories of sentences with bracketing contrasts and those without. There is a great deal of variation among ambiguous sentences that

have bracketing contrasts, however, in terms of what kinds of structural contrasts actually distinguish the two meanings. Price, Ostendorf, Shattuck-Hufnagel, & Fong (1991) tested how, and how well, seven different kinds of ambiguous sentences could be prosodically disambiguated. While all of their ambiguities involved sentences with contrasting bracketings, and thus prosody was expected to be able to disambiguate them, it was unclear what the success rate for disambiguation would be for each structure or with what kinds of prosody each structure would be disambiguated.

The seven structures Price et al. (1991) tested are listed in the leftmost column of Table 1.1:

Table 1.1 Results of perceptual experiments in Price et al. (1991)

	Meaning 1	% correct	Meaning 2	% correct
Parenthetical clauses versus nonparenthetical subordinate clauses	Mary knows many languages, you know.	77	Mary knows many languages you know.	96*
Appositions versus attached noun (or prepositional) phrases	The neighbors who usually read, the Daleys, were amused.	92*	The neighbors who usually read the dailies were amused.	91*
Main clauses linked by coordinating conjunctions versus a main clause and a subordinate clause	Mary was amazed and Dewey was angry.	88*	Mary was amazed Ann Dewey was angry.	54
Tag questions versus attached noun phrases	Dave will never know why he's enraged, will he?	95*	Dave will never know why he's enraged Willy.	81
Far versus near attachment of final phrase	Raoul murdered the man with a gun. [Raoul used a gun to murder]	78	Raoul murdered the man with a gun. [the murdered man had a gun]	63
Left versus right attachment of middle phrase	When you learn gradually, you worry more.	94*	When you learn, gradually you worry more.	95*
Particles vs. prepositions	Then men won over their enemies. [i.e. they won them over]	82*	Then men won over their enemies. [i.e. they beat them]	81*

I will return to two of these structures later in the chapter for their similarity with discourse-level ambiguities.

In their experiment, Price et al. had four FM radio announcers read each sentence following disambiguating contextual material. Each speaker's productions were then played for between 12 and 17 listeners. Each listener heard one version of each of the 35 sentences in one listening session, and then returned for a second listening session to hear the production they had not yet heard for each of the 35 sentences. In each session, the listeners had an answer sheet containing contexts that would disambiguate to one or the other meaning, and were asked to "mark the context that they thought best matched what they heard" (p. 2960).

Price et al. (1991) found that participants were overall above chance at retrieving the context of the original production (84% correct). In their statistical analysis, the authors compared each meaning's accuracy rate to chance. So, for example, they would test whether productions of right attachments in right vs. left attachment ambiguities were significantly more likely be identified as right attachments, and then separately they would test whether identification of left attachments was significantly above chance. This means each ambiguous structure is analyzed twice, once for each meaning. The "% correct" column to the right of each meaning in Table 1.1 indicates at what percentage listeners were able to correctly retrieve the original context of the speaker's production. An asterisk marks whether that result is more than one standard deviation from the chance rate of 50%. There were 285 observations in each class. They found that of the 14 total meanings (7 structural ambiguities x 2 meanings), only three were close to chance while the rest showed a prosodic disambiguation effect. Participants' accuracy rates ranged from around chance for some structures all the way up to 96% accuracy for the most successfully communicated structure. Accuracy was also affected by which speaker was heard, indicating some speakers' productions were more effective at disambiguating than others.

Price et al. (1991) demonstrate that a wide range of syntactic ambiguities can be prosodically disambiguated, some more successfully than others. Having just

examined seven structurally distinct syntactic ambiguities and prosody's ability to disambiguate them, Price et al. (1991) nevertheless emphasize in their discussion the importance of examining a "larger number of syntactic structures" to better understand the relationship between prosody and syntax. Carlson (2009) reiterates this point in her review article when she writes that "it is likely that we will find effects of prosody in new sentence structures" (2009, 1197). Thus, understanding the relationship between prosody and linguistic structure benefits from examining prosody's behavior with respect to many different kinds of linguistic structures.

This current state of research suggests there are many open and worthwhile questions about how discourse ambiguities relate to what we already know about prosodic disambiguation of syntax. Research on prosodic disambiguation of discourse could illuminate how scalable certain kinds of prosodic meanings are. It could answer whether the contexts in which speakers produce disambiguating prosody are similar or different for syntactic and discourse-level structures. And this research may also help us understand the larger and more general question of how discourse-level and sentence-level structures are similar and how they are different.

Prosodic disambiguation of sentences vs. discourse

There are similarities and differences in how analogous structures at the sentence and discourse levels are prosodically disambiguated. One of the ambiguous structures examined in Price et al. (1991) is what they call a near/far ambiguity, which is structurally analogous to the high/low ambiguity of Silverman (1987) and Mayer et al. (2006). The far attachment is equivalent to a high attachment because attaching further back in the sentence would correspond to attaching higher up in a syntactic tree. And near attachment is equivalent to low attachment as both occur lower in the overall structure. In fact, Wagner & Watson (2010), in their review of the current state of work on prosody, discuss these ambiguities as high and low attachment ambiguities. For these reasons, I will refer to far/high attachment as high attachment, and near/low as low attachment.

Price et al. (1991) found that productions of the high attachment versions of their sentences tended to have a large prosodic break before the ambiguously attached

phrase, while low attachment versions had a relatively small break. To exemplify, productions with a break like in (6.3a) would be interpreted more often to have high attachment than productions like in (6.3b).

(1.8)

a. Raoul murdered the man [BIG BREAK] with a gun.
b. Raoul murdered the man [small break] with a gun.

That is, larger prosodic breaks before an ambiguously attached phrase are more likely to be interpreted as high attachments. This pattern is equivalent to the findings of Silverman (1987) and Mayer et al. (2006). Both studies were cuing a large boundary before an ambiguous sentence, and larger boundaries biased interpretations towards more high attachments. This suggests that the size of a prosodic boundary can bias interpretation of high/low ambiguities at both the sentence and discourse levels.

But the nature of how to prosodically indicate a large break may vary between sentences and discourses. Initially, Price et al. (1991) analyzed their productions in terms of phonological categories, including break size. They then analyzed the phonetic correlates of these phonological categories. They found durational cues to be the strongest phonetic markers of break size, finding longer segment and syllable durations before larger breaks. They add that "though intonation is an important cue, duration and pauses alone provide enough information to automatically label break indices with a high correlation…" (p. 2965), citing Ostendorf, Price, Bear, & Wightman (1990).

The results from Mayer et al. (2006) and Silverman (1987) suggest that intonation may play a more important role at the level of discourse. Mayer et al. found that pause and pitch manipulations together biased interpretation but either alone had no effect. And Silverman found a stronger effect when pause and pitch information were both available, but the pitch manipulations alone still had a significant effect. Perhaps then analogous structures at the level of discourse depend more on intonational cues than at the level of the sentence. One potential reason for this could be that durational variation could be noisier in discourse, where sentence

boundaries may receive a large pause independent of any particular attachment. I will return to this issue of how prosodic disambiguation relates to syntactic and discourse ambiguities in chapter 6.

Outline of Dissertation

As contextualized above, this dissertation explores the interface between prosody and discourse structure in production and perception. Chapter 2 presents the results of a production study where participants read aloud a newspaper article that was annotated for discourse structure. The study in chapter 2 tests for prosodic correlates of the size of a discourse boundary and the local hierarchical contrast of coordination vs. subordination between discourse segments (CoordSubord). It also makes a novel contribution by testing for an interaction between boundary size and CoordSubord, i.e. whether the effect on prosody of one structural feature depends on the value of the other.

Chapters 3 and 4 present the results of perception studies that test for an effect of synthesized manipulations of prosody on the interpretation of ambiguous discourse. Unlike discourse prosody in production, we know relatively little about effects of discourse prosody on discourse interpretation. Two studies have found manipulations of pause duration and pitch can bias the interpretation of ambiguous phrases (Mayer, et al., 2006; Silverman, 1987). These results are encouraging, showing that prosody can affect listeners' linguistic judgments of discourse-dependent linguistic phenomena. What these studies do not conclusively demonstrate is that prosody can disambiguate hierarchically structured discourse. I argue below that the ambiguities used by both Mayer et al. and Silverman, while described by the authors in hierarchical terms, can also be accounted for as near vs. far ambiguities. By contrast, the ambiguities used in the studies in chapters 3 and 4 of this dissertation cannot be explained as near/far ambiguities, instead requiring a hierarchical explanation. For this reason, these chapters test whether prosody can disambiguate *hierarchical* discourse structure.

Chapter 5 presents a discussion of the results of the perception studies in chapters 3 and 4. In Chapter 5, I examine the meaning of the one prosodic

manipulation that appeared to drive the overall interpretation effect, a terminal pitch rise vs. fall contrast. This discussion reviews work on listing intonation, and how listing intonation is implicated in the results. I argue that my results suggest that one possible meaning for rising terminal pitch is discourse coordination. I then discuss an example that Pierrehumbert & Hirschberg (1990) use to make the apparently conflicting claim that high terminal pitch indicates elaboration, a subordinating relation. I reanalyze their example, bringing their data in line with my results.

In chapter 6, I reflect on the dissertation as a whole. Comparing discourse prosody results in both production and perception, I argue that there may be a mismatch between what we are measuring in production studies and what listeners actually use in perception. Given the results of the dissertation's production and perception studies, I return to the issue of how prosody functions in the disambiguation of both sentences and discourse. The chapter continues by discussing language processing research that has focused less on what structures prosody *can* disambiguate but rather on the contextual effects on when speakers and listeners *do* prosodically disambiguate. While this literature has thus far focused on prosodic disambiguation of syntax, many of the questions they discuss can again be asked with respect to the disambiguation of discourse. Finally, I discuss potential future research projects.

Chapter 2

Prosodic Correlates of Discourse Boundaries and Hierarchy in Discourse Production

As mentioned in chapter 1, there is a body of research on discourse production that has identified systematic correlates between aspects of discourses' structure and prosodic measures of pitch, pause duration, intensity and speech rate (den Ouden, et al., 2009; Hirschberg & Grosz, 1992; Lehiste, 1975). This indicates that the prosody of speech can carry cues to the structure of discourse. One common feature of discourse with which prosody is correlated in this work is the size of a discourse boundary. This work uses diverse criteria to identify boundaries of different size. These criteria include orthographic markers of paragraph boundaries (Lehiste, 1975, 1982) and intuitive analyses of breaks in the discourse, either by the experimenter (Yule, 1980) or the participants themselves (Swerts, 1997). Other work creates measures of boundary size using a specific theory of discourse structure, e.g. den Ouden et al. (2009) use Rhetorical Structure Theory (Mann & Thompson, 1988) and Hirschberg & Grosz (1992) use the Grosz & Sidnel model (Grosz & Sidner, 1986). These studies use different terms to describe similar phenomena, but for consistency I will use the term boundary size. Compared with boundary size, less is known about prosodic correlates of local hierarchical relationships in discourse like coordination and subordination, though see den Ouden et al. (2009). And very little is known about their interaction, i.e. how the effects of boundary size and coordination/subordination may depend on each other.

The prosodic measures most commonly found to correlate with discourse have been pause duration and pitch maxima, though others have been explored as well.

Pause durations have tended to be longer at larger discourse boundaries (den Ouden, et al., 2009; Lehiste, 1982). Pitch maxima, characterized variously as pitch range (Hirschberg & Grosz, 1992), pitch reset (Auran & Hirst, 2004), and high onset pitch (Couper-Kuhlen, 2001; Yule, 1980), tend to be higher following larger boundaries in the discourse. Den Ouden et al. (2009) found a correlation between the nucleus/satellite contrast and articulation rate, but no correlation with pause duration or maximum pitch. And while they have gotten less attention, discourse has been found to correlate with other prosodic measures like amplitude (Herman, 2000; Hirschberg & Grosz, 1992) and rhythm (Müller, 1996).

In this chapter, I present the results of a discourse prosody production study. The study tests for prosodic correlates of boundary size and coordination/subordination, as well as their interaction. In addition to some of the more traditional prosodic measures, it includes measures of how far through a discourse segment pitch and intensity maxima occur. These measures can help illuminate if discourse structure also correlates with pitch or intensity peaks along a temporal dimension. In addition, this study presents the results of two correlation analyses, one correlating the predictor variables and the other correlating the prosodic measures. These correlation analyses reveal how independent the variables are, informing which ones to exclude and how to interpret those that remain.

The spoken data were elicited by having participants read aloud a newspaper article, and then correlating prosodic features of those productions with features of the article's discourse structure. As discussed by Smith (2004, p. 249), a benefit of using read speech instead of spontaneous speech is that the discourse annotation is not based on prosodic information (for more discussion of this circularity concern, see Swerts, 1997). But because read speech sometimes differs from spontaneous speech (Laan, 1997), further research would be needed to see if results from this study extend to more spontaneous forms of speech production.

Results from this study can inform how prosody correlates with discourse structure in speech production and set the stage for follow-up perception studies. The results could also inform and improve the development of speech synthesis and

recognition systems by better accounting for the prosodic variation those systems need to take into account.

Discourse Structure

The discourse structure variables in this study are derived from a discourse representation constructed using *Segmented Discourse Representation Theory* (SDRT) (Asher & Lascarides, 2003). As mentioned above, SDRT identifies the structure of a discourse by dividing it into segments, inferring rhetorical relations that hold between those segments (e.g. ELABORATION) and assembling them into a hierarchical structure. To set the stage for the explanation of how these variables were created, it will be helpful first to exemplify how SDRT performs each of these processes. We can work through these processes with the excerpt in (2.1), drawn from the newspaper article used in this study. The article and all SDRT annotations come from the DISCOR annotated corpus (Reese, et al., 2007), a research project that used SDRT to determine the discourse structure of natural language texts and identify dependencies between anaphors and their antecedents. The goal of the current study is not to test the value of an SDRT representation against other theories' representations, but to take the SDRT analysis as a good discourse representation from which to test relationships between prosody and discourse structure.

(2.1) The White House will try to assuage at least some opponents' concerns as Congress undertakes to reconcile the Senate bill with a much different House measure. Justice Department officials, who were criticized for not visibly exerting influence over the Senate bill last year, will play a more overt role in removing or modifying the more extreme provisions this year. Deputy Attorney General Philip Heymann plans to testify at House crime legislation hearings, and Mr. Clinton himself held out the carrot of help to endangered youth in his speech to Congress.

The first step in analyzing the discourse structure of (2.1) is segmentation. Sentence boundaries were all treated as segment boundaries, and sub-sentential portions were treated as discourse segments if they served "a discernible discourse

function" (Reese, et al., 2007:3)[1]. For ease of representation, the resulting segments are each assigned a number corresponding to their sequential position in the text:

(2.2) [27 The White House will try to assuage at least some opponents' concerns] [28 as Congress undertakes to reconcile the Senate bill with a much different House measure.] [29 Justice Department officials, who were criticized for not visibly exerting influence over the Senate bill last year, will play a more overt role in removing or modifying the more extreme provisions this year.] [30 Deputy Attorney General Philip Heymann plans to testify at House crime legislation hearings,] [31 and Mr. Clinton himself held out the carrot of help to endangered youth in his speech to Congress.]

The brackets indicate segment boundaries and the numbers are a shorthand way to refer to the discourse's segments.

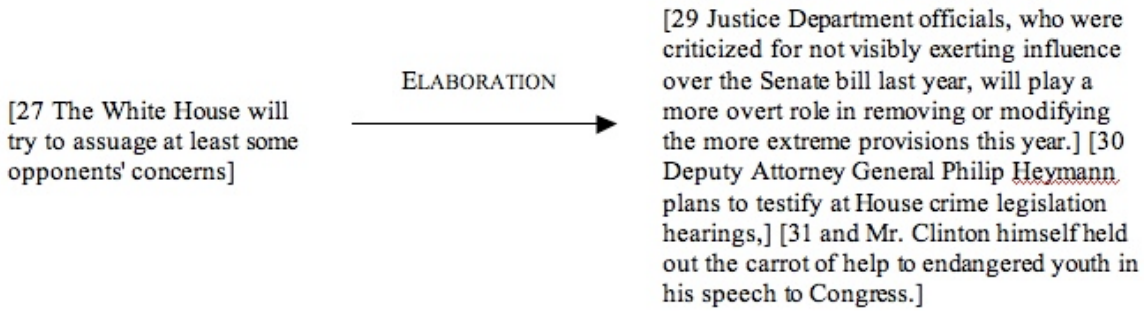
After segmentation, rhetorical relations are identified that are inferred to hold between those segments. Written as RELATION(ARG1,ARG2), a relation is said to hold between its two arguments. The DISCOR annotators identified the following relations in the excerpt:

(2.3) Elaboration(27,[29,30,31])
BACKGROUND(27,28)
CONTINUATION(29,30)
CONTINUATION(30,31)

Here, ELABORATION(27,[29,30,31]) captures a rhetorical relation of elaboration where the text corresponding to segment 27 is elaborated by the text corresponding to segments 29, 30 and 31:

[1] Full details of the SDRT annotations are available in the DISCOR annotation manual (Reese, et al., 2007).

Figure 2.1: Schematic representation of an ELABORATION relation



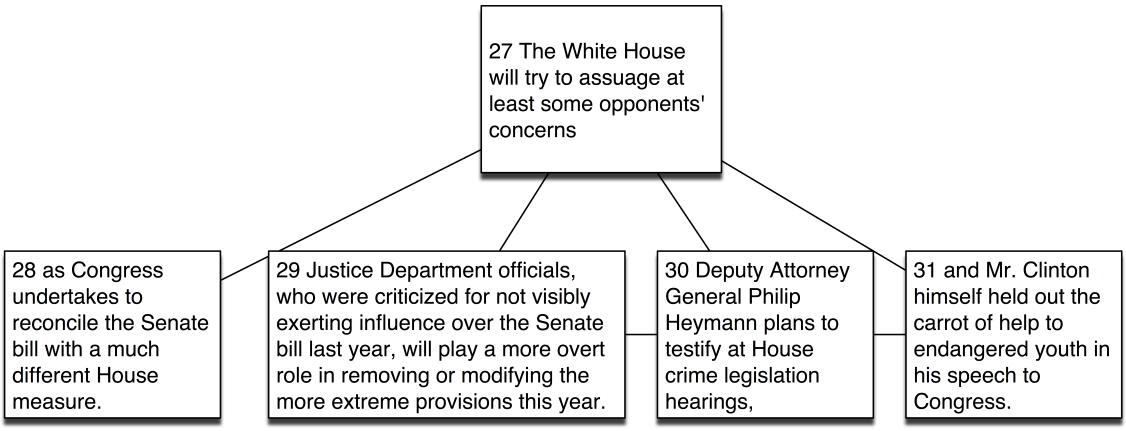
The DISCOR annotation manual explains that "ELABORATION(α,β) holds when β provides further information about the eventuality introduced in α" (Reese, et al., 2007:7). In this example, the first argument of the elaboration relation introduces the proposition of the White House assuaging opponents' concerns, about which the second argument provides further information in the form of the Justice Department's more overt role (segment 29), Heymann's testimony (segment 30) and Clinton's reaching out (segment 31). In this example, the elaboration relation's second argument is composed of three discourse segments while the other argument is composed of a single discourse segment. SDRT arguments can be simple (made up of a single segment) or complex (made up of multiple segments). Along the same lines, the relations BACKGROUND(27,28), CONTINUATION(29,30) and CONTINUATION(30,31) indicate one background and two continuation relations between their first and second arguments, respectively. A discourse is said to be coherent if rhetorical relations connect all of its segments.

In SDRT, hierarchical structure is captured by categorizing all relations as either coordinating or subordinating. The information captured in the coordinating/subordinating contrast is the 'granularity' or level of detail being given in the discourse" (Asher & Lascarides, 2003:8). The second argument of a subordinating relation provides more detail than the first argument, while the second argument of a coordinating relation provides a similar level of detail as the first argument. The relation ELABORATION, for example, is a subordinating relation, meaning the second argument provides more detailed information than the first, and as a result is a level below the first. CONTINUATION is a coordinating relation,

meaning the second argument provides a similar level of detail as the first and are at the same level. This hierarchical view contrasts with conceptions of discourse as a set of propositions or possible worlds, as well as with "the dynamic semantic view of text information as a sequence of information updates" (Asher & Vieu, 2005:591). SDRT's hierarchical structure also achieves empirical gains in its ability to account for phenomena like anaphoric reference and temporal structure in ways a non-hierarchical theory cannot (Asher & Lascarides, 2003).

In the graphical representation in Figure 2.2, vertical lines are used to indicate subordinating relations and horizontal lines to indicate coordinating relations. The arguments of those relations are represented as the boxes at the end of the lines.

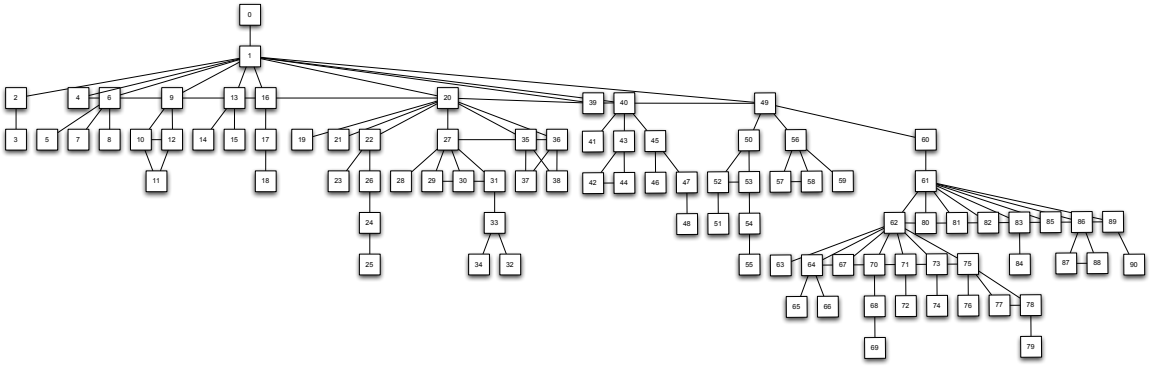
Figure 2.2: Graphical representation of SDRT analysis of segments 27-31



This graph shows segments 29, 30 and 31 subordinated to, i.e. a level below, segment 27 while coordinated to each other. In terms of propositional content, segment 27 introduces the proposition of the White House assuaging opponents' concerns, and segments 29-31 elaborate on that proposition. Segment 28 gives background information about 27.

If a similar analysis is applied to the whole article, we get the graph in Figure 2.3.

Figure 2.3: Graphical representation of SDRT analysis of entire article



The graph in Figure 2.3 captures the segmentation, discourse relation and hierarchical organization information that constitutes an SDRT discourse structure representation.

Methods

The production data in this study were elicited by having participants read, analyze and then read aloud a newspaper article. The article's structure, modeled with SDRT, was converted into two predictor variables corresponding to boundary size (Bsize) and the contrast between coordinating and subordinating relations (CoordSubord). Prosodic measures were taken from the readings and tested for correlations with those predictor variables.

Participants

Ten students from the University of Michigan participated in this study in exchange for ten dollars. Eight speakers were female and two were male. All reported English as their native language and English as their major field of study. English majors were chosen because they were expected to be particularly capable of identifying the discourse structure of a news article, a necessary first step to test the larger question of how prosody correlates with discourse structure. With this population, non-significant results are less likely to be due to speakers not identifying the discourse structure and more likely to be due to how prosody correlates with discourse structure. And because the goal of this study is to gain the greatest insight

into how prosody correlates with discourse, not to identify the average person's discourse prosody, this non-random selection best fulfilled this goal.

Materials

Participants were asked to read aloud the 1994 Wall Street Journal newspaper article titled *Blacks' increasing vocal opposition to violence is matched by strong opposition to crime bill* (Davidson, 1994)[2]. The article comes from the DISCOR corpus of news articles annotated within SDRT (Reese, et al., 2007). The article addresses new crime legislation proposed during Bill Clinton's presidency, the reaction to it among black leaders and the various political factions involved. This article was chosen because it was sufficiently long and diverse in terms of features necessary to test the research questions. Having all participants read the same article, instead of each one reading a different article, meant that variability between speakers could not be due to idiosyncratic differences between texts. This study's discourse structure variables were derived from the article's discourse structure as characterized in the SDRT annotations from DISCOR.

Design

This study's overarching goal is to determine how information about a discourse segment's position in the discourse can help predict that segment's prosody. It will be tested by asking whether speakers indicate with their prosody (a) the size of a boundary between discourse segments (Bsize), (b) whether a segment is coordinated or subordinated to the most recent segment to which it is attached (CoordSubord), and (c) whether the effect of either (a) or (b) is mitigated by the other. These sub-questions will be addressed using predictor variables for Bsize and CoordSubord that were converted from the above SDRT representation and then testing them for significant correlations with prosodic measures.

---

[2] The full text of the article as presented to participants is in Appendix A.

Boundary Size (Bsize)

The Bsize variable captures the amount of structure intervening between sequential segments of a discourse, e.g. segment 47 and 48. The Bsize variable's values are calculated as the number of nodes in the discourse structure intervening between two sequential segments. In practice, this involves identifying the shortest path between two segments and counting how many other segments must be traveled through to reach the next one.

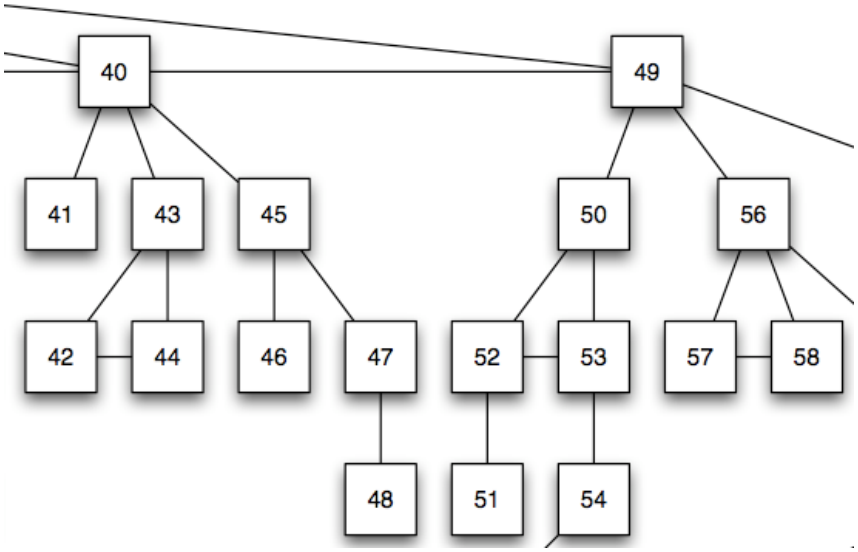
The excerpt in (2.4) below can help exemplify how the values for Bsize were calculated.

(2.4) Excerpt from news article used in study, with discourse segments numbered in sequential order

40. But the mainstream civil-rights leadership generally avoided the rhetoric of "law and order,"
41. regarding it as a code for keeping blacks back.
42. Law and order didn't mean justice,
43. Mr. Jackson used to say,
44. but "just us."
45. In the past, many were hesitant to speak about crime in public
46. because "the larger community would talk about 'lock them up and throw the key away' and hide behind black leaders in doing it,"
47. explains Rep. Craig Washington,
48. the Houston Democrat who led the caucus hearing.
49. Now there is escalating discourse within the black community about what it can and must do to stop crime.
50. Just after the new year, Mr. Jackson held the first of several conferences focusing on just that.
51. "The premier civil-rights issue of this day is youth violence in general and black-on-black violence in particular,"
52. he has said.

The text in (2.4) has a discourse structure of the form shown in Figure 2.4. In Figure 2.4, the pair 47-48 has no intervening segments, and so has a boundary size of 0. By contrast, the pair 48-49 has three intervening nodes (segments 47, 45 and 40) and so has a boundary size of 3.

Figure 2.4: Graphical representation of SDRT analysis of segments 40-58



Boundary size was calculated in the same way for all sequential pairs, resulting in the following distribution:

Table 2.1: Distribution of discourse segment frequencies and average number of words at each level of boundary size

Boundary Size	Level 0	Level 1	Level 2	Level 3	Level 4
Discourse segments (n=90)	54 (60%)	24 (27%)	7 (8%)	4 (4%)	1 (1%)
Average # words	11.47	16.25	15.86	21.25	4

As the table makes clear, there are decreasing numbers of segments as boundary size increases. While a more even distribution would be preferable for the statistical analysis, it is the nature of this discourse, and perhaps all discourses, that boundaries between adjacent segments are more often small than large. In addition, the analysis requires enough observations at each level for the statistical model to be able to compare them. Because there was only one level 4 boundary, level 4 was merged with level 3. Combining the data from levels 3 and 4 led to sufficient data at every level for the statistical model to have the power necessary to test for Bsize as a predictor of prosody.

Coordinating vs. Subordinating Relations (CoordSubord)

The CoordSubord variable is designed to test whether local hierarchical relationships between discourse segments result in different prosody in speech production. The variable captures three different ways a discourse segment could be connected to the larger discourse structure. The Coord part of CoordSubord is named for discourse segments whose most recent attachment is via a coordinating relation, e.g. segment 58 in the graph in Figure 2.4. The Subord part of CoordSubord is named for discourse segments whose most recent connection is via a subordinating relation, e.g. segment 57 in the graph in Figure 2.4. The variable only considers the most recent discourse relation for each discourse segment, regardless of what connections that discourse segment may have to earlier relations in the discourse. The decision to code based on the most recent relation was due to the overall distribution of relations. All discourse segments except the first were subordinated to at least one other segment, while only 27 segments were coordinated. And all of those coordinated segments were more recently coordinated than subordinated. Therefore, the contrast between coordination and subordination appears in terms of which relation is most recent.

There is a third group of discourse segments that are not connected to any earlier discourse segments but only to those that come later in the discourse. These discourse segments are produced before it is clear to which segments they are rhetorically related. Of the 90 total discourse segments in the text, 56 were coded as subordinated, 27 as coordinated, and 7 were related to an upcoming segment. The data for the third level of CoordSubord were excluded from the analyses; this was done to better isolate the purpose of the variable, namely to compare coordinated vs. subordinated discourse segments. And because this resulted in a total of 83 out of 90 segments remaining, there was still enough data to address the question about effects of coordination vs. subordination on prosody.

Prosodic Measures

This study focused on a broad set of prosodic measures, including pause duration, pitch, intensity, and speech rate (Table 2.2).

Table 2.2: The set of acoustic features extracted from each discourse segment. For the pitch and intensity peak location measures, the unit captures what proportion through a discourse segment the peak occurred.

Acoustic feature	Units	Description
Pause duration	milliseconds	Duration of silence preceding discourse segment
Pitch maximum (f0max)	Hz	Maximum F0 across entire discourse segment
Pitch minimum (f0min)	Hz	Minimum F0 across entire discourse segment
Mean pitch	Hz	Mean F0 across entire discourse segment
Initial pitch	Hz	Mean F0 for initial 5% of discourse segment
Pitch peak location	0-1	How far through the discourse segment the highest pitch point occurs
Max intensity	dB	Maximum decibel level in discourse segment
Mean intensity	dB	Mean decibel level in discourse segment
Min intensity	dB	Minimum decibel level in discourse segment
Intensity peak location	0-1	How far through the discourse segment the highest intensity point occurs
Speech rate	Words/duration	Words divided by duration of discourse segment

The measures were extracted automatically using a script in the acoustic analysis software program Praat (Boersma & Weenink, 2009). For women, the pitch window was set at 100-500Hz, while for men the window was 75-300Hz. For the automatic measurement, it was necessary to adjust Praat's default pitch settings to be more conservative about what it accepts as pitch in order to reduce errors. For f0max, it sufficed to raise the voicing threshold from the default 0.45 to 0.6, as performed elsewhere (Ljolje, 2002). The resulting output was reliable enough that the few remaining errors could be spotted and fixed by hand. The f0min measurements, however, required different and more conservative pitch settings. F0min can be chaotic to measure because speakers sometimes enter creaky voice as they descend in pitch. When using the f0max pitch settings, f0min had a binomial distribution, with a cluster of measurements around 100 Hz all in creaky voice. In order to filter out these creaky voice measurements, the Praat settings were made more conservative by

increasing the voicing threshold, the octave cost and the voicing/voiceless cost[3], resulting in a more normal distribution. Remaining outliers were checked individually.

The measures for pitch peak location and intensity peak location are measured as how far through a discourse segment the peak is produced. If the peak occurred at the very beginning, the measure would equal zero; if the peak occurred at the very end, the measure would equal 1. And if the peak occurred 20% of the way through the discourse segment, the measure would equal 0.2. The goal of these peak location measures was to explore variation along the temporal dimension, i.e. where in the segment prosodic phenomena occurred. Given that high onset pitch has been found to occur after large boundaries in discourse (Auran & Hirst, 2004; Wichmann, 2000), pitch peaks were expected to occur earlier following larger discourse boundaries and on coordinated discourse segments. Though little is known about its behavior, intensity peak location was included to be able to compare results for pitch and intensity.

Excluded from the analysis were discourse segments with disfluent speech production. Previous research has described disfluencies as "fillers like uh or um, unfilled pauses, repeated words, repaired words, or even disfluent-sounding prosody" (Arnold, 2008:508). Because the focus of this study is on prosodic production, this study treats lexically anomalous production as disfluent, but not "disfluent-sounding prosody" like intuitively unexpected lengthening or awkward pauses. Disfluency was defined as when a speaker repeated a word, said a word that was not present in the text, did not utter a word that was present, or had some extra-verbal interruption like coughing or sneezing. 153 out of 910 total segments (17%) were excluded from the analysis due to disfluency.

Each speaker read the text of the newspaper article aloud twice. For nine of the ten speakers, prosody measures were taken from their second reading. For the tenth speaker, the first reading was used. The second reading was chosen for most speakers because it was thought they would have gotten more familiar with the text

[3] For details, see Appendix C.

and the task, and so have produced more fluent speech. The one speaker whose first reading was used was flustered when asked to read it again. I analyzed her first reading because on the first reading she appeared less distracted from the task.

Procedure

The recordings were done in the noise-controlled sound lab at the University of Michigan, using Praat and an AKG C 4000 B microphone. Participants first read the article silently to themselves, then paraphrased it out loud, and finally read the entire article out loud twice. Participants were directed to read the text aloud in such a way as to most clearly communicate the article to a listener. All of the article's paragraphing was removed, but sentence-level punctuation was left in. There were no subheadings. As a result, participants had no information about paragraph structure, and so overt paragraphs themselves could not account for prosodic variation.

The motivation for having participants paraphrase the article out loud before reading it aloud was both to get participants to think about the text in a structural way and to have a record of what they saw as the text's most important points. The paraphrases provide a check on whether the SDRT representations are capturing features of the text that participants found important. A comparison between the paraphrases and SDRT appears in the Discussion section. Participants were encouraged to study the article for as long as necessary prior to reading aloud in order to understand its structure as well as possible.

Results

Potential effects of discourse structure on prosody were tested by fitting a Linear Mixed Model to the data. Each model used contained the predictor variable(s) of interest (Bsize, CoordSubord), control variables, and the dependent measures of prosody. The control variables were included to help rule out explanations other than discourse structure for the prosodic variation. One potential confound may be that speakers change their prosody over the five to ten minutes participants took to read the text, perhaps due to factors like fatigue or wandering attention. A variable was

included in the model that indicated how far along in the discourse the segment was uttered to try to control for these potential location effects (*Number*, for discourse segment number). Another confound could be whether material in the text was in quotes or not. A variable was added that indicated whether the discourse segment was wholly, partially, or not at all in quotes (*Quot*). Of the text's 90 discourse segments, 71 had no quoted material, 9 had some, and 11 had all quoted material. Additionally, some discourse segments began sentences while others began in the middle of sentences; a sentence-initiality variable was added to capture this information (*Sentinit*). Of the text's 90 discourse segments, 46 are sentence-initial and 44 are non-initial. And finally, the length of a discourse segment may affect how extreme a prosodic measure becomes; to capture this information, one variable was included that indicated the number of words in the discourse segment (*Words*) and another that indicated the duration of the segment in seconds (*Duration*).

Before analyzing the results of discourse structure predicting prosody, it will be useful to analyze correlations among the predictor variables (predictors of interest as well as controls) and then analyze correlations among the prosodic outcomes. The correlation analyses can show which variables pattern together, revealing if some are redundant and can be excluded. The correlations can also help in the interpretation of the variables that remain, by showing how independent each variable is from the others.

In Table 2.3, all predictor variables are laid out in a matrix of Pearson correlations. The higher the correlation values, the more closely the variables pattern together. As the value gets closer to zero, the variables are more independent from each other. And as the correlation gets closer to negative one, the more the variables pattern in opposite directions.

Table 2.3: Correlation Matrix for Predictor and Control Variables

	Sentence-Initiality	Boundary Size	Coord/Subord	Quoted	Overall Embeddedness	Discourse Segment Number	Words	Segment Duration
Sentence-Initiality	1	0.471	-0.086	0.011	0.186	-0.038	0.263	0.248
Boundary Size	0.471	1	0.022	0.017	0.249	-0.042	0.261	0.194
Coord/Subord	-0.086	0.022	1	0.085	-0.167	-0.073	-0.03	-0.009
Quoted	0.011	0.017	0.085	1	-0.012	0.005	0.018	-0.02
Overall Embeddedness	0.186	0.249	-0.167	-0.012	1	-0.723	0.003	-0.11
Discourse Segment Number	-0.038	-0.042	-0.073	0.005	-0.723	1	0.022	0.097
Words	0.263	0.261	-0.03	0.018	0.003	0.022	1	0.904
Segment Duration	0.248	0.194	-0.009	-0.02	-0.11	0.097	0.904	1

It is clear the variables *Words* and *Duration* are highly correlated with each other (.904) and independent of the other variables. These two variables seem to be capturing the same information. This makes sense because the number of words in a segment is likely to affect how long it takes to say that segment. Because of the high correlation, one of either *Words* or *Duration* should be removed. And because *Duration* is less correlated with *Bsize* than words, *Duration* will be left in the model and *Words* will be removed. A weaker correlation shows up between *Bsize* and *Sentinit* (0.471), indicating that segments after larger boundaries are more likely to be sentence-initial. *CoordSubord* is largely independent of the other variables, and notably has almost no correlation with *Bsize*.

Correlation results for the prosodic measures are laid out in Table 2.4.

Table 2.4: Correlation Matrix for Prosodic Measures

	Pause Duration	Pitch Maximum	Pitch Minimum	Mean Pitch	Initial Pitch	Pitch Peak Location	Maximum Intensity	Mean Intensity	Minimum Intensity	Intensity Peak Location	Speech Rate (words/sec)
Pause Duration	1	0.212	-0.111	0.05	0.2	-0.262	0.262	0.099	-0.177	0.258	-0.052
Pitch Maximum	0.212	1	0.654	0.88	0.846	-0.134	0.276	0.2	0.187	0.233	-0.082
Pitch Minimum	-0.111	0.654	1	0.876	0.675	-0.003	0.047	0.146	0.383	0.09	-0.063
Mean Pitch	0.05	0.88	0.876	1	0.843	-0.071	0.181	0.226	0.34	0.16	-0.072
Initial Pitch	0.2	0.846	0.675	0.843	1	-0.191	0.212	0.167	0.26	0.232	-0.07
Pitch Peak Location	-0.262	-0.134	-0.003	-0.071	-0.191	1	-0.216	-0.163	0.042	-0.211	-0.102
Maximum Intensity	0.262	0.276	0.047	0.181	0.212	-0.216	1	0.707	-0.105	0.116	-0.116
Mean Intensity	0.099	0.2	0.146	0.226	0.167	-0.163	0.707	1	0.317	0.082	0.188
Minimum Intensity	-0.177	0.187	0.383	0.34	0.26	0.042	-0.105	0.317	1	0.051	0.309
Intensity Peak Location	0.258	0.233	0.09	0.16	0.232	-0.211	0.116	0.082	0.051	1	0.143
Speech Rate (words/sec)	-0.052	-0.082	-0.063	-0.072	-0.07	-0.102	-0.116	0.188	0.309	0.143	1

The largest cluster of high correlation values are among the measures for maximum, minimum, mean and initial pitch. Because they pattern together, only one measure is needed to capture this effect. While any of the correlated pitch measures could be used, pitch maximum was retained and pitch mean, pitch minimum and initial pitch

were excluded. This decision was made because pitch maximum is a common measure in other discourse prosody research (den Ouden, et al., 2009; Hirschberg & Grosz, 1992). Also, maximum and minimum intensity are both correlated with mean intensity, though not with each other. Therefore maximum and minimum intensity were retained but mean intensity was removed. The remaining measures are largely uncorrelated.

The final set of predictor, control and dependent variables are listed in Table 2.5. There is one control variable in this table that was not included in the correlation analyses This variable captures the prosody in the previous discourse segment, testing how much a prosodic measure's value in a current discourse segment is predictable from that same measure in the prior segment. For example, saying the previous segment loudly might lead to the subsequent segment being louder. As a result, a speaker's maximum intensity in one segment may be highly related to their maximum intensity in the prior segment. This *ProsPrev* variable can help separate variation in a prosodic measure that is due to the previous segment's prosody and not to the discourse structure. The *ProsPrev* variable was not included in the correlation analysis because it varies from dependent variable to dependent variable.

Table 2.5 Full list predictor, control and dependent variables used in the Linear Mixed Model

Predictor and Control Variables	**Dependent Variables (Prosody)**
Boundary Size (Bsize)	Pause Duration
Coordination vs. Subordination (CoordSubord)	Pitch Maximum (f0max)
Sentence-Initiality (Sentinit)	Pitch Peak Location
Segment Duration (Duration)	Maximum Intensity
Quoted Material (Quot)	Minimum Intensity
Discourse Segment Number (Number), i.e. how far through the discourse the segment occurs.	Intensity Peak Location
Previous segment's value for same prosodic measure (ProsPrev).	Speech Rate

The ability of each predictor variable to predict each dependent variable was tested for significance with a Linear Mixed Model (LMM). Because each subject is providing many data points, the observations are not fully independent, an assumption in statistical models like ANOVA. We may be better able to predict a

prosodic outcome by taking into account who produced it. We can take these subject effects into account by including a random intercept for subjects in a linear mixed model.

Linear mixed models offer a range of benefits over other repeated measures models like repeated measures ANOVA. Quené & van den Bergh have run two studies demonstrating the benefits of mixed models over ANOVA, first with normally distributed data (2004) and then with binary data (2008). In both cases, mixed models serve as better fits of the data. Mixed models benefit from being able to accommodate missing data and unequal cell sizes, two concerns in this data set. Another benefit is the ability to avoid making false assumptions about the independence of the observations by taking into account repeated measures. Mixed models also afford higher statistical power and so are more able to accurately identify effects in the data. All statistical modeling was performed with SPSS.

Boundary Size

To identify overall patterns of boundary size on prosodic outcomes, a linear mixed model was fitted that contained Bsize but not CoordSubord. This model tells us what effect a change in boundary size has on each prosodic outcome. Boundary size was entered as a continuous variable.

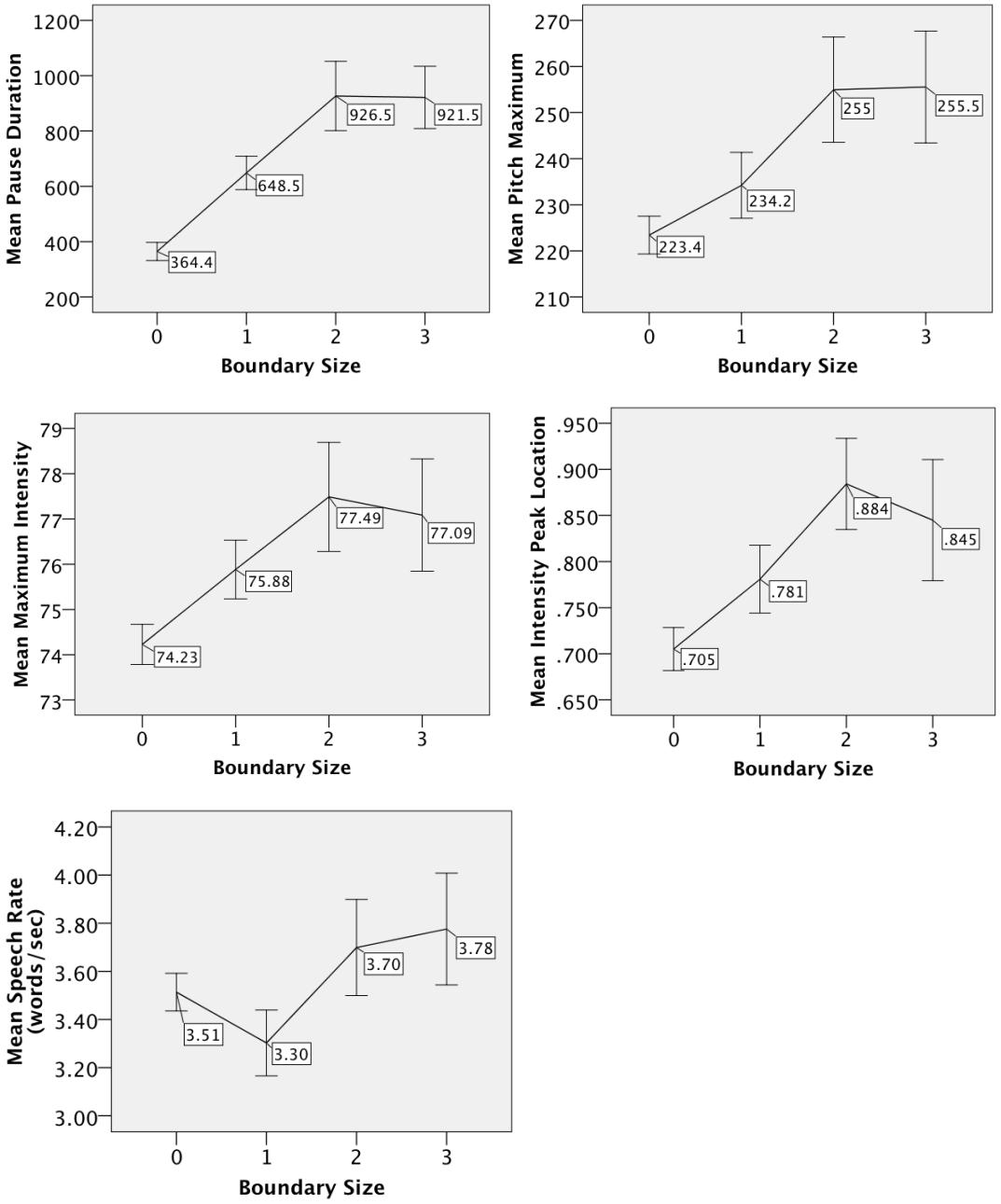
Table 2.6: Results for boundary size as a predictor of prosody, collapsing across CoordSubord. The intercept indicates the model's predicted value for each prosodic measure when boundary size is 0. The coefficient indicates the model's predicted slope, i.e. the amount of change in that prosodic measure for every level increase in boundary size. *=p<.05, **=p>.01.

Boundary Size	Intercept	Coefficient	F-statistic	p-value
Pause	665.659	78.490	25.806	0.000**
Maximum Pitch	215.205	4.392	11.449	0.001**
Pitch Peak Location	0.176	-0.022	2.432	0.119
Maximum Intensity	70.820	0.612	18.243	0.000**
Minimum Intensity	33.139	-0.080	0.140	0.709
Intensity Peak Location	0.722	0.052	17.313	0.000**
Speech Rate	3.948	0.124	10.734	0.001**

In Table 2.6, we see the results for each prosodic measure by row. The intercept indicates the model's predicted value for each prosodic measure when boundary size is 0. The coefficient indicates the model's predicted slope, i.e. the amount of change in that prosodic measure for every level increase in boundary size.

Results for boundary size indicate that pause duration, max pitch, max intensity, and speech rate all increase as a discourse segment's preceding boundary increases in size. Furthermore, a discourse segment's intensity peak occurs later in the segment as the preceding boundary gets larger. These effects are all highly significant ($p<.01$). The graphs in Figure 2.5 plot each prosodic measure on the y-axis for each level of boundary size on the x-axis.

Figure 2.5: Line graphs with boundary size on x-axis and pause duration, f0max, max intensity, intensity peak location and speech rate on the y-axis. Error bars indicate 95% confidence intervals.



All prosodic measures other than speech rate show an increase from level 0 to 1 to 2, with a plateau between 2 and 3. Speech rate drops from level 0 to 1 before rising to 2 and 3. Results for the control variables are presented in Table 2.7.

Table 2.7 Results for all independent variables in the model with boundary size as the only predictor variable of interest. Prosodic measures are in the left column, and predictor variables are along the top row. *=p<.05, **=p>.01.

		Intercept	sentinit	number	quot	duration	prosprev	bsize
Pause Duration	F	88.521	295.646	0.914	1.058	20.965	21.427	25.806
	p	0.000**	0.000**	0.340	0.348	0.000**	0.000**	0.000**
Maximum Pitch	F	215.540	72.485	6.065	1.706	87.167	0.454	11.449
	p	0.000**	0.000**	0.014*	0.182	0.000**	0.501	0.001**
Pitch Peak Location	F	61.690	11.079	0.732	1.680	1.525	8.093	2.432
	p	0.000**	0.001**	0.393	0.187	0.217	0.005**	0.119
Maximum Intensity	F	644.139	49.337	10.406	9.283	71.205	4.176	18.243
	p	0.000**	0.000**	0.001**	0.000**	0.000**	0.041*	0.000**
Minimum Intensity	F	237.309	5.090	0.181	1.357	177.497	22.302	0.140
	p	0.000**	0.024*	0.670	0.258	0.000**	0.000**	0.709
Intensity Peak Location	F	281.138	3.777	0.278	1.761	2.444	2.715	17.313
	p	0.000**	0.052	0.598	0.173	0.118	0.100	0.000**
Speech Rate	F	399.602	1.025	28.866	7.810	38.547	4.826	10.734
	p	0.000**	0.312	0.000**	0.000**	0.000**	0.028*	0.001**

There are many significant effects of the control variables on the prosodic outcomes, demonstrating the importance of including them in the model.

Sentence-initiality (*Sentinit*) was a strong predictor of all measures except intensity peak location and speech rate. So, while a segment's preceding pause duration was dramatically affected by whether that segment was sentence-initial or not, that segment's speech rate was not.

The position of the discourse segment in the overall discourse (*Number*) was a significant predictor for max pitch, max intensity and speech rate, but not for pause duration, pitch peak location or intensity peak location. This suggests that over time speakers changed how high, loud and fast they spoke, but did not change pause durations or the relative position of the pitch and intensity extremes. Whether a discourse segment contained quoted material predicted maximum intensity and speech rate.

A discourse segment's duration in seconds (*Duration*) was a significant predictor of all prosodic measures except the relative pitch and intensity peak measures. This suggests that where in a discourse segment a pitch or intensity peak occurs is not dependent on how long it takes to say the segment.

And finally, the prosody of the previous segment (*ProsPrev*) significantly predicted pause duration, pitch peak location, max and min intensity, and speech rate. It seems for these measures, speakers may get into periods of using the prosody in one way that spans multiple discourse segments and is independent of the discourse structure. For example, speakers may get into periods of longer or shorter pauses.

In sum, measures of preceding pause duration, pitch maximum and intensity maximum all increased as preceding boundary size increased. Moreover, intensity peaks occurred later in segment after larger boundaries. And while Figure 2.5 indicates speech rate drops from boundary size 0 to 1 before increasing at levels 2 and 3, overall it was the case that speech rate increased following larger boundaries.

CoordSubord

To identify overall patterns of CoordSubord on prosodic outcomes, a model was fitted that contained CoordSubord as a predictor but not boundary size. This model tells us what effect a change in a segment being subordinated or coordinated has on each prosodic outcome. CoordSubord was entered as a binary categorical variable.

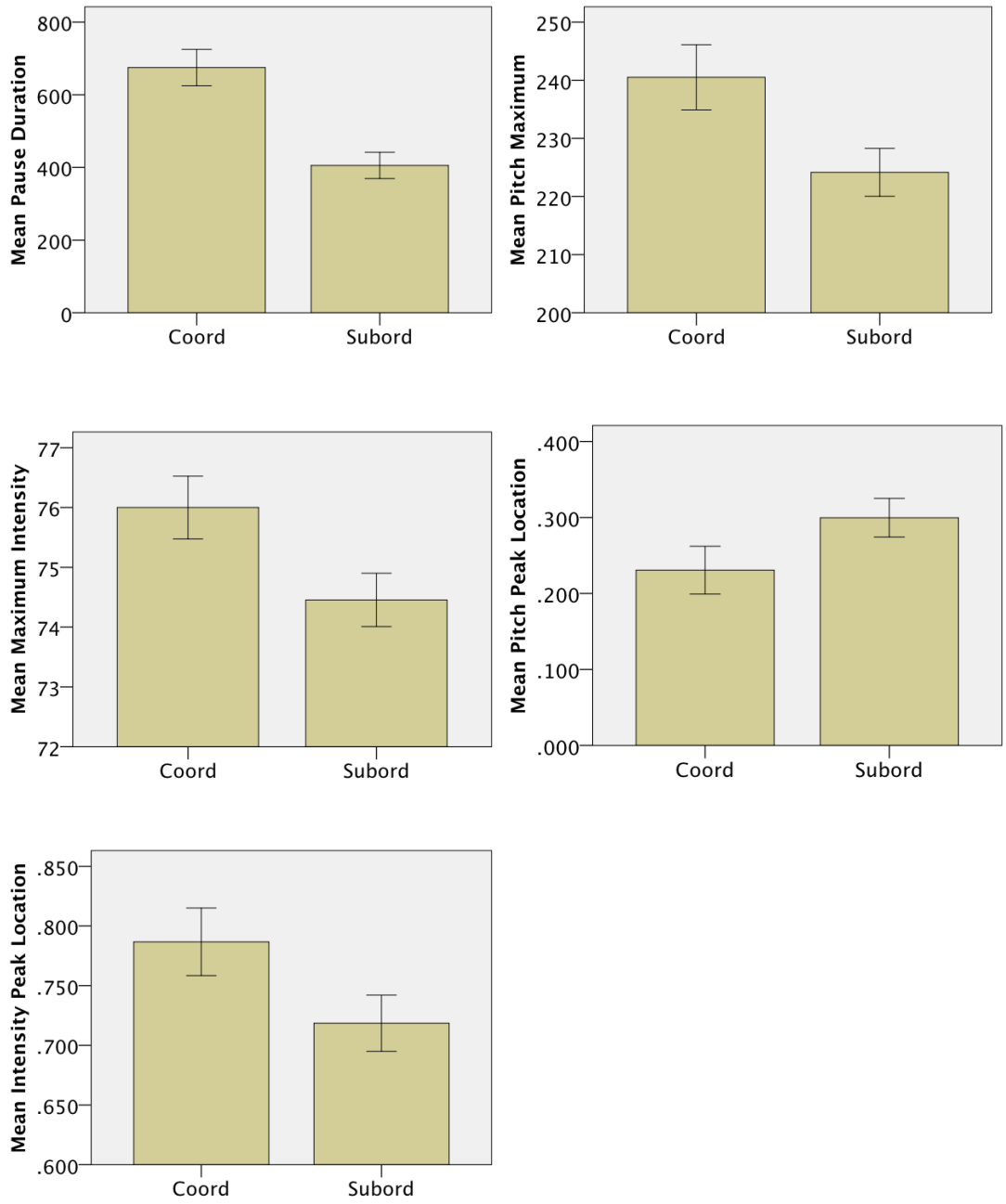
Table 2.8: Results for boundary size as a predictor of prosody, collapsing across CoordSubord. The intercept indicates the model's predicted value for each prosodic measure when a segment is subordinated (the reference value). The coefficient indicates the model's predicted slope, i.e. the amount of change in that prosodic measure by being coordinated instead of subordinated. *=p<.05, **=p>.01.

CoordSubord	Intercept	Coefficient	F-statistic	p-value
Pause	695.116	123.877	25.544	0.000**
Maximum Pitch	221.491	10.496	25.604	0.000**
Pitch Peak Location	0.181	-0.054	5.900	0.015*
Max Intensity	70.869	1.194	25.827	0.000**
Minimum Intensity	33.170	-0.148	0.173	0.678
Intensity Peak Location	0.758	0.054	6.996	0.008**
Speech Rate	4.107	0.065	1.063	0.303

In Table 2.8, we see the results for each prosodic measure by row. The intercept indicates the model's predicted value for each prosodic measure when a discourse segment is subordinated. The coefficient indicates the model's predicted slope, i.e. the amount of change in that prosodic measure when a segment, instead of being subordinated, is coordinated.

Results for CoordSubord indicate that a discourse segment's preceding pause duration, max pitch and max intensity all increase when a discourse segment is coordinated instead of subordinated. Furthermore, a coordinated discourse segment's pitch peak occurs earlier while its intensity peak occurs later relative to subordinated discourse segments. The graphs in Figure 2.6 plot each prosodic measure on the y-axis for both Coord and Subord on the x-axis.

Figure 2.6: Bar graphs with CoordSubord on x-axis and pause duration, f0max, max intensity and speech rate on the y-axis.



Results for the control variables are presented in Table 2.9.

Table 2.9: Results for all independent variables in the model with CoordSubord as the only predictor variable of interest. Prosodic measures are in the left column, and predictor variables are along the top row. *=p<.05, **=p>.01.

		Intercept	sentinit	number	quot	duration	prosprev	CS
Pause Duration	F	122.865	351.059	1.777	0.666	28.632	35.860	25.544
	p	0.000**	0.000**	0.183	0.514	0.000**	0.000**	0.000**
Maximum Pitch	F	236.632	81.453	4.340	2.495	103.376	0.042	25.604
	p	0.000**	0.000**	0.038*	0.083	0.000**	0.838	0.000**
Pitch Peak Location	F	54.040	12.188	1.094	2.160	2.432	6.885	5.900
	p	0.000**	0.001**	0.296	0.116	0.119	0.009**	0.015*
Maximum Intensity	F	662.713	60.715	12.946	11.942	85.234	4.251	25.827
	p	0.000**	0.000**	0.000**	0.000**	0.000**	0.040*	0.000**
Minimum Intensity	F	237.196	5.439	0.203	1.415	176.765	21.780	0.173
	p	0.000**	0.020*	0.652	0.244	0.000**	0.000**	0.678
Intensity Peak Location	F	338.275	9.579	0.527	2.099	4.107	1.171	6.996
	p	0.000**	0.002**	0.468	0.123	0.043*	0.280	0.008**
Speech Rate	F	428.699	0.015	29.782	8.378	35.582	2.835	1.063
	p	0.000**	0.903	0.000**	0.000**	0.000**	0.093	0.303

Like in the model with boundary size, there are many significant effects of the control variables, demonstrating the importance of having them in the model. Only two results are different when CoordSubord is in the model instead of Bsize. First, sentence-initiality becomes a strong predictor of intensity peak location. This is perhaps not that surprising if we recall that sentence-initiality and Bsize are somewhat correlated (r=0.471). When Bsize is not accounting for some of the variation in a prosodic outcome, sentence-initiality fills some of that absence. And second, the speech rate of the previous segment no longer predicts speech rate of the current segment. This suggests that part of why a previous segment's speech rate was predictive of the current segment's speech rate is due to whether those segments are linked to the larger discourse via coordination or subordination. So even though CoordSubord does not predict speech rate, it seems to have an effect on other factors.

We will explore the complexity of the relationship between CoordSubord and speech rate in more detail in the next section.

Interaction of Boundary Size and CoordSubord

In addition to modeling Bsize and CoordSubord independently as predictors of prosodic outcomes, we want to see if the effect of one variable depends on the value of the other. For example, the CoordSubord contrast may only be relevant at some levels of Bsize. We can test this by modeling both predictors together in the same model, including each as a main effect as well as their interaction. If the interaction is significant, then the slope for coordinated segments across the levels of boundary size differs from the slope for subordinated segments across the levels of boundary size. That is, a significant interaction would tell us that the effect of a segment being coordinated vs. subordinated would depend on the size of the preceding boundary.

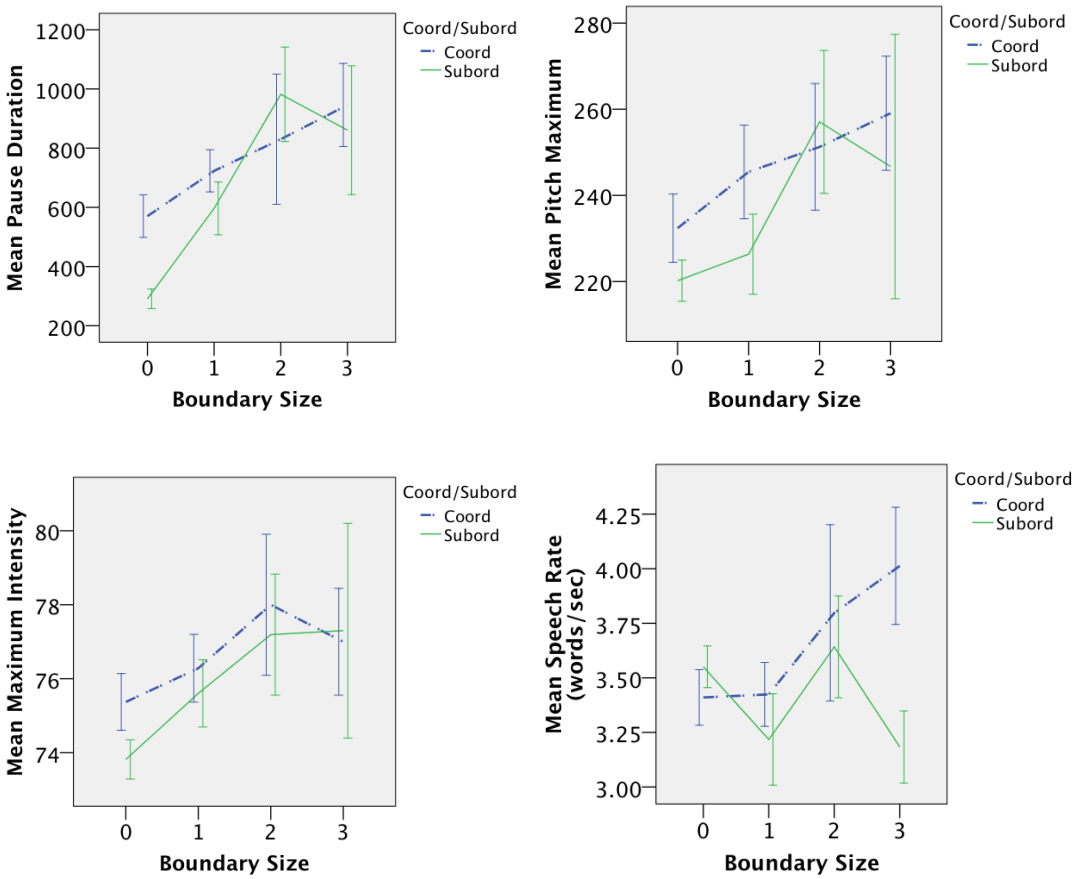
A linear mixed model was fitted to the data with Bsize, CoordSubord and their interaction as predictors, along with the controls listed in Table 2.5, for each prosodic outcome. Results are shown in Table 2.10.

Table 2.10: Table of results for Linear Mixed Model with Bsize, CoordSubord and the interaction Bsize*CoordSubord. The CoordSubord result indicates whether prosodic outcomes are significantly different between Coord and Subord when Bsize=0. The Bsize result indicates whether the prosodic outcomes for Subord (the reference value of CoordSubord) change as Bsize changes. *=p<.05, **=p>.01.

		Bsize	CoordSubord	Interaction
Pause	F	25.698	30.857	8.795
	p	0.000**	0.000**	0.003**
	Coefficient	119.936	165.961	-84.734
Maximum Pitch	F	9.110	25.147	4.111
	p	0.003**	0.000**	0.043*
	Coefficient	6.347	12.782	-4.908
Pitch Peak Location	F	1.641	2.571	0.185
	p	0.201	0.109	0.667
	Coefficient	-0.012	-0.044	-0.011
Max Intensity	F	16.171	27.501	5.788
	p	0.000**	0.000**	0.016*
	Coefficient	0.899	1.507	-0.657
Minimum Intensity	F	0.118	0.165	0.029
	p	0.731	0.685	0.865
	Coefficient	-0.110	-0.180	0.071
Intensity Peak Location	F	14.996	2.689	0.098
	p	0.000**	0.101	0.754
	Coefficient	0.045	0.041	0.008
Speech Rate	F	7.500	7.533	33.086
	p	0.006**	0.006**	0.000**
	Coefficient	-0.104	-0.206	0.413

Table 2.10 shows a significant interaction between Bsize and CoordSubord for pause duration, max pitch, max intensity, and speech rate. This means that for each of these prosodic measures, the effect of CoordSubord depends on Bsize. The nature of this interaction is visible in the graphs in Figure 2.7.

Figure 2.7: Line graphs with boundary size on x-axis, a dashed blue line for Coord and a solid green line for Subord. The graphs are tiled by prosodic measure, with the relevant scale for each on their y-axis (pause duration, f0max, max intensity and speech rate).



These four graphs plot boundary size along the x-axis, with the relevant prosodic measure on the y-axis. The blue dashed line corresponds to Coord and the solid green line corresponds to Subord. For three of the four measures (pause duration, max pitch and max intensity), we see separation between Coord and Subord when BSize=0. In these three cases, Coord has a higher value (longer pause duration, higher max pitch and max intensity). As boundary size gets bigger, the Coord and Subord lines get closer, meaning that the differences between Coord and Subord get smaller. Speech rate behaves differently. For speech rate, Subord is higher at Bsize=0, the lines cross and separation increases with Coord higher at higher levels of boundary size.

For all four of these prosodic measures, results in Table 2.10 show us that the difference between Coord and Subord is significant when Bsize=0 (the reference

value for Bsize). We can test whether the difference between Coord and Subord is significant at the other levels of Bsize by making each level the reference value for Bsize (Table 2.11).

Table 2.11: Results testing whether the prosodic outcomes are significantly different between Coord and Subord measurements at each level of boundary size. *=p<.05, **=p>.01.

CoordSubord	at Bsize=0	at Bsize=1	at Bsize=2	at Bsize=3
Pause Duration	F=30.857; p<.001**	F=9.431; p=.002**	F=.006; p=.940	F=.006; p=.940
Max Pitch	F=25.147; p<.001**	F=11.954; p=.001**	F=.564; p=.453	F=.100; p=.752
Max Intensity	F=27.501; p<.001**	F=11.058; p=.001**	F=.190; p=.663	F=.448; p=.503
Speech Rate	F=7.533; p=.006**	F=9.184; p=.003**	F=27.563; p<.001**	F=31.805; p<.001**

These results show that the difference between coordinated and subordinated segments is significant for pause duration, max pitch and max intensity when Bsize=0 or 1, but not when Bsize>1. By contrast, speech rate shows a significant contrast between Coord and Subord at every level of boundary size. In Figure 2.7, we see that speech rate increases for Coord segments and decreases for Subord segments as Bsize increases. Moreover, the two lines cross. At Bsize=0, Coord segments are spoken significantly slower than Subord segments. But at all levels of Bsize>0, Coord segments are spoken significantly *faster* than Subord segments. We saw there was a main effect of Bsize on speech rate (see Table 2.6) but no main effect of CoordSubord on speech rate (see Table 2.8). While one may have interpreted the lack of main effect of CoordSubord on speech rate as meaning CoordSubord didn't matter for speech rate, the interaction results show it just matters in a more complex way. By analyzing the interaction effect, we see that the impact of boundary size on speech rate is mediated by whether a discourse segment is coordinated or subordinated.

Discussion

This study has found evidence of prosodic correlates of both boundary size (Bsize) and coordination vs. subordination (CoordSubord). Moreover, it identified significant interactions between Bsize and CoordSubord, showing that prosody's

relationship to CoordSubord is integrally related to its relationship to Bsize and vice versa. The interaction of Bsize and CoordSubord was significant for pause duration, max pitch, max intensity, and speech rate. This means that for these measures the effect of Bsize or CoordSubord depends on the value of the other.

The results for Bsize show increasing values for pause duration, max pitch and max intensity as Bsize increased. This is in line with existing research on prosodic correlates of discourse structure, which Smith (2004) summarizes as suggesting "greater prominence at the beginning of a discourse or immediately after a major boundary" (p. 250). This greater prominence is indicated in this study by longer pauses and higher pitch and intensity. Speech rate is more complicated, showing an overall trend of increasing speech rate with increasing boundary size. But the picture is actually more complex, where speech rate actually drops from level 0 to 1, and then increases substantially for levels 2 and 3. Furthermore, there is a significant interaction between Bsize and Coord in the prediction of speech rate. Subordinated segments actually show a mild slowing in speech rate as Bsize increases, while Coord segments get produced faster. And finally, intensity peaks occur later in a discourse segment as Bsize increases. This suggests that later intensity peaks, by patterning with the other prosodic measures, may work in tandem with pause duration, max pitch and max intensity to indicate greater prominence.

Results for CoordSubord showed higher values of pause duration, max pitch and max intensity for coordinated segments than subordinated segments. This provides prosodic evidence that coordination is in a more prominent position, because the same measures that conveyed prominence for Bsize do so for CoordSubord as well. But Bsize and CoordSubord are not correlated with each other ($r=0.022$, see Table 2.3), so this prosodic prominence is conveying different information about the discourse structure. It also makes sense to think of coordination as more prominent than subordination, as by definition coordinated segments are hierarchically higher than subordinated segments, all else being equal.

More surprising are the results for the two proportional measures for pitch and intensity peak locations, which showed that coordinated segments had earlier pitch peaks but later intensity peaks. Given that pitch and intensity peak values pattern

together for Bsize, it is remarkable that they pattern in opposite directions in terms of how far through the discourse segment those peaks occur. The result for earlier pitch peaks is consistent with research on pitch reset and claims that high onset pitch occurs at topic onsets (Auran, 2007; Yule, 1980). Less is known about the behavior of intensity peaks, raising questions about how these two measures are able to operate independently and in opposite directions.

We also know relatively little about prosodic correlates of coordination and subordination in discourse, though den Ouden et al. (2009) test something similar in their study of prosodic correlates of the RST distinction between nuclei and satellites. In RST, all discourse segments are classed as either nuclei or satellites, where satellites are those segments that are less important and can more easily be removed without disturbing the larger coherence of the discourse. Danlos (2010) compares RST and SDRT in terms of their theoretical underpinnings and ability to account for the felicity and infelicity of discourses. He concludes that the two theories roughly rely on the same set of discourse relations and give them the "same type," i.e. coordinating/subordinating or nucleus/satellite. He seems to be treating the two binaries coordinating/subordinating and nucleus/satellite as comparable and in some sense equivalent. It is noteworthy then that den Ouden et al. (2009) found different results for prosodic correlates of the nucleus/satellite distinction than this study found for CoordSubord. Den Ouden et al. found no correlation between pause duration or max pitch with nuclei and satellites, but did find that nuclei were produced with a slower articulation rate than satellites. In contrast, this study found coordinated segments were produced with longer preceding pause durations and higher maximum pitch, but no difference in speech rate. There are a few possible interpretations for these contrasting results. First, den Ouden et al. (2009) used Dutch texts and Dutch participants while this study was conducted entirely in American English. Perhaps the languages themselves can account for the different prosodic correlates. Second, it could be due to a difference in coding, where the way the nucleus/satellite distinction in den Ouden et al. (2009) and CoordSubord in this study were coded differed. Third, it is possible that RST's nuclearity and SDRT's coordination/subordination contrast are not actually equivalent. One piece of evidence for nuclearity and CoordSubord

being different is in their relation to measures of boundary size. In den Ouden et al. (2009), the measures for nuclearity and boundary size, which they call “Hierarchy”, are highly correlated ($p<.001$) (p. 125). In the study described in this paper, Bsize and CoordSubord are not correlated ($r=.022$). This suggests the nuclearity variable in den Ouden et al. (2009) is capturing much of the same information as boundary size, while CoordSubord in this study reflects different features of the discourse than boundary size.

And while the independent variables Bsize and CoordSubord are not correlated, results show a significant interaction between them as predictive of pause duration, max pitch, max intensity, and speech rate. For pause duration, pitch max and intensity max, there is a significant difference between coordinated and subordinated discourse segments at levels 0 and 1 of Bsize, but this CoordSubord effect disappears when Bsize is 2 or 3. This indicates that when the boundary between segments is smaller, information about whether the new segment is coordinated or subordinated matters. But when boundary size increases, the coordination/subordination contrast is no longer significant.

The interaction between Bsize and CoordSubord is significant for speech rate but in a different way. For speech rate, CoordSubord is a significant predictor at all levels of Bsize. When Bsize=0, subordinated segments are produced significantly faster. But when Bsize is 1 or larger, coordinated segments are produced faster. Furthermore, the speech rate of subordinated segments falls mildly as Bsize increases. By contrast, coordinated segments show a steady increase in speech rate as Bsize increases. This suggests that much of the effect of speech rate occurs in the coordinated segments. So, why would coordinated segments be spoken faster as Bsize increases, but subordinated segments would not? The difference may be due to differences in novelty. A new coordinated segment is creating a new space in the discourse, while a new subordinated segment is providing more information about something already under discussion. If the assumption here holds that newer information is produced with faster speech rate, then this relative novelty contrast could account for the difference. It is also possible this effect is related to the speech being monologic, read speech. In this experiment, speakers did not have a listener

present with whom to interact and for whom to adjust their speech. It is possible that speakers behaved more with respect to their own needs than if listeners were present. They also were tasked with reading a text aloud verbatim without speech errors. In this task, the speech planning process involved reading instead of planning in one's own head. Perhaps the reliance on text for linguistic material affected speaking rate. It is unclear why novelty, monologue or reading aloud would lead to faster speech after larger boundaries, but these factors may be involved in an explanation. A separate explanation would say that listeners expect a default speech rate for default interpretations. Since large boundaries are less common than smaller ones, it is a relatively marked context. Perhaps this study's speakers were using marked prosody, i.e. faster speech, as a way of conveying this marked discourse context.

This study also examined measures which capture temporal information about how fast a speaker gets to pitch and intensity peaks. These measures reveal different patterns for pitch and intensity peaks: coordinated segments have earlier pitch peaks and later intensity peaks relative to subordinated segments. Intensity peaks also were later in a segment as Bsize increased. These results demonstrate there is potentially meaningful prosodic variation along this temporal dimension. Therefore, a fuller account of discourse prosody will need to take into account the location of prosodic peaks in addition to the values of those peaks.

As the above discussion shows, there are a number of significant correlations between discourse structure and speakers' prosody in the context of this study. Especially notable is how similar the correlations are between Bsize and the prosodic measures of pause duration, max pitch and max intensity. This raises the question of whether all three measures independently correlate with discourse structure, or whether there is some underlying prosodic category that gets fed forward to the phonetic realization of pause duration, pitch and intensity. If we posit a direct relationship, then we miss the potential generalization across the prosodic measures. Instead, we could posit an underlying category that mediates between discourse and the acoustic measures. Instead of speakers directly connecting discourse to the acoustics, they would have a representation of something like discourse prosodic emphasis; cf. Smith's (2004) term "greater prominence" to refer to similar patterns

across prosodic measures (p. 250). In this case, one way discourse structure would interface with prosody is by generating more or less prosodic emphasis, which would itself then get spelled out in terms of phonetic measures like pause duration, pitch and intensity.

But if there is an underlying category behind the overt manifestations of pause duration, pitch and intensity, it raises questions about why there is still so much variability from measure to measure. It also raises the question of what motivation there could be for providing redundant cues to the structure of discourse. One possible explanation is that this variability provides necessary flexibility for discourse prosody to convey information about discourse structure. There are many factors that can affect the realization of pause duration, pitch and intensity, and can have an effect in different ways for the different prosodic measures. While overall the prosodic correlates of discourse structure may appear to be redundant, individual productions could exploit only some subset of the three prosodic measures. For example, in cases where pause duration is determined by other factors, speakers can still draw on pitch or intensity. What may from a macro-perspective seem redundant, in more individual instances could be important flexibility. Hirschberg & Grosz (1992) make a similar conclusion when they write that "different configurations of intonational features may be employed to convey the same discourse information in different contexts. For while our aggregate statistics show certain trends, not every token exhibits all these differences" (p. 446). Furthermore, redundancy of cues can also reinforce meanings that could otherwise be difficult to convey. In fact, redundant cues in production appear to facilitate the perception of discourse prosody (Mayer, et al., 2006; Silverman, 1987).

It is also worth mentioning that this study tested for correlation, not causation. Subsequent research could try to determine whether discourse structure *causes* prosodic correlates. For example, if speakers are presented with a single discourse that could be interpreted in two ways corresponding to two different discourse structures, would speakers produce the different structures with different prosody? And could listeners successfully communicate to listeners which interpretation they intended? Holding the lexical and syntactic information constant while varying the

discourse structure would help isolate the discourse structure as the cause of the prosodic correlates.

Paraphrase Analysis

In discourse prosody research, a common concern has been how to get a good representation of the discourse's structure independent of any prosody. When using spoken data as the basis of the structural analysis, there is a risk of circularity where prosodic information motivates the structure that is then correlated with prosodic measures. Scholars have tried to solve this problem by focusing on discourses with relatively uncontroversial structures, like BBC news broadcasts (Wichmann, 2000), or using naïve participants to mark discourse boundaries (Swerts, 1997). Others have used a specific discourse theory, like the Grosz & Sidner model (Grosz & Hirschberg, 1992) or Rhetorical Structure Theory (den Ouden, et al., 2009), taking the theory to provide a good approximation of how participants were representing the structure of discourse. Similarly, this study used a specific discourse theory (SDRT) and took the annotations to be a good approximation of how participants were representing the discourse's structure. Unlike previous studies, this one had participants paraphrase the discourse before reading it aloud, which could provide one way to check whether participants' sense of key points corresponded well to key points in the SDRT representation. If the main topics of the paraphrases line up well with the main topics of the SDRT analysis, then we have evidence to support the claim that SDRT is capturing something about how participants are representing the discourse.

To explore whether such a correspondence existed, I first listened to each participant's paraphrase of the article and noted the topics mentioned. Then, I examined the SDRT representation to see what topics are in those discourse segments after the largest boundaries (level 3). I am using the term *topic* here to capture something like discourse topic, i.e. what the content is about. For a full list of topics and results of this analysis, see Appendix D.

Results show that nearly all participants mentioned topics 1 and 2, with fewer mentioning subsequent topics. The first topic, the overall topic for the article, was not captured by the boundary size measure. This is not surprising given that the overall

topic of the article was mentioned at the very beginning, before enough has been said for there to be a large boundary. The second topic, addressing specific concerns with a crime bill, was also mentioned by nearly all participants. Subsequent topics received a minority of mentions, apparently with decreasing mentions as the topics were later in the article. This suggests paraphrasing may highlight topics introduced earlier in a discourse, with subsequent topics deemed less integral. Furthermore, each paraphrase was short, lasting between 45 and 90 seconds. In this time, participants only covered 2-4 topics, clearly choosing to emphasize the first two. These paraphrases suggest participants understood the discourse, and that the boundary size measure grounded in the SDRT representation captures some relevant aspects of how participants understood the discourse.

Conclusion

The significance of this study's findings are constrained in two ways. First, the generalizability of these findings is limited by the use of read speech. Because read speech has been shown to differ from spontaneous speech (Laan, 1997), we cannot assume that this study's findings would necessarily show up in non-read speech. But even in this constrained context of read speech, it does show that speakers can produce speech in such a way as to carry discourse structural information. The skill of the reader has also been found to be an important dimension for variation in read speech. As noted by Esser (1988) and Wichmann (2000), amateur and professional readers differ in how they read aloud, with professional readers tending to more consistently use prosodic features. This study used amateur readers and was still able to identify prosodic correlates of discourse structure.

Second, by using only SDRT to represent the discourse structure with which prosodic measures were correlated, comparisons cannot be made between the representations of SDRT and other theories. Were one to annotate the structure of a single discourse using multiple theories, then prosodic correlates could reveal which theory had the strongest correlations and permit comparisons between theories (e.g. Den Ouden (2004) using RST and the Grosz & Sidner model). This information could

help identify which theories have the strongest prosodic correlates and potentially adjudicate between them.

Having demonstrated some ways prosodic measures correlate with discourse structure, this study does motivate follow-up work that could test whether listeners exploit that information in their perception. I see two kinds of issues discourse prosody perception studies could address: disambiguation and facilitation. If identical lexical material has multiple possible discourse structures, could prosody bias interpretation towards one or another? If stimuli are created where there is a mismatch between the prosody and the discourse structure, would listeners show processing difficulties? A study in Auran (2007) suggests mismatching prosody can induce processing difficulties in the interpretation of French discourse. Also, if one discourse is produced with distinct discourse prosody and another without, would listeners rate the speech with the discourse prosody as easier to understand or as more effective? Would comprehension or retention increase? It is possible one aspect of what makes good speakers easier to understand is their use of discourse prosody.

Finally, the findings discussed in this paper may lend themselves to practical applications, e.g. speech synthesis and speech training. Because speech synthesis systems currently tend to suffer from unnatural-sounding prosody, perhaps the correlates identified here could help inform ways to improve them. And if discourse prosody is found to facilitate comprehension and assessments of speaker effectiveness, teaching people to use discourse prosody could help them become more effective speakers.

Chapter 3

Prosodic Effects on the Interpretation of Discourse Ambiguities Using a Set of Synthesized Prosodic Manipulations (Psychology Subject Pool)

In this chapter, I present an experiment testing whether a set of prosodic manipulations can bias the interpretation of an ambiguous discourse. The ambiguity depends on how the sentences of the discourse are related to each other. All discourses in this study were three sentences long, where sentences 2 and 3 attach to sentence 1 via either a coordinating or a subordinating relation. For example, the discourse in (3.1) could be interpreted such that the narrator read about housing prices and watched a cool documentary while sitting in on the history class (the Subord interpretation) or separate from the history class (the Coord interpretation).

(3.1) I sat in on a history class. I read about housing prices. And I watched a cool documentary.

The Coord interpretation of the discourse indicates the listener thinks the three events described by the three sentences happened independently, while the Subord interpretation indicates sentences 2 and 3 provided more detail about the event described in sentence 1. And for these discourses, all lexical and syntactic material is held constant, meaning the ambiguity is at the level of discourse and how discourses are structured. And in contrast to the studies in Mayer et al. (2006) and Silverman (1987), outlined in chapter 1, the ambiguity cannot be reduced to a near/far contrast. Instead, the contrast is necessarily hierarchical. As a result, this study is able to test prosody's ability to disambiguate hierarchical discourse, controlling for the confounding factor of discourse recency.

Methods

Participants

Forty students from the University of Michigan Psychology Subject Pool participated in this study in exchange for course credit. All reported being native speakers of American English. Ages ranged from 17 to 21 with a mean of 18.43. 27 of the 40 participants were female, 13 male. 14 (35%) reported knowing a second language. As for education level, 20 reported having completed high school while 20 others reported having some undergraduate education. Given that all were currently taking Intro to Psychology, those who actually reported only a high school education really had had at least 3 weeks of college.

Materials

A total of 102 discourses were generated, each discourse being ambiguous between the Coord and Subord interpretations described above. These 102 discourses were included in a norming study to test for general preferences for each discourse's interpretation as Coord or Subord. This norming study was meant to ensure that the discourses used in the studies were practically, not just logically, ambiguous. That is, the goal was to create a set of discourses where it was reasonable, not just possible, to interpret them as either Coord or Subord. A more complete description of the method and results of this norming study are available in Appendix E. The norming study resulted in a continuum of discourses from preferred interpretations for Coord, to equibias, to Subord. The discourses were ranked from most to least equibiased, i.e. from most to least ambiguous. The 52 most ambiguous discourses in the norming study were selected as the discourses for this study. The 48 most ambiguous discourses were used as target stimuli, with the remaining four serving as training.

All spoken materials were recorded in the sound-attenuated booth in the University of Michigan Linguistics Department's Sound Lab. All individual sentences in all discourses were separated and entered into a list, resulting in a list of

52 x 3 = 156 sentences. Each sentence was then placed into a carrier context like in (3.2).

(3.2) I am going to read a sentence. I read about housing prices. I just read a sentence.
I am going to read a sentence. I sat in on a history class. I just read a sentence.
I am going to read a sentence. I went for a run. I just read a sentence.

After randomizing the order of presentation, a 30-year-old, female, native speaker of American English was recorded reading each sentence out loud one at a time in its carrier context. This reader was instructed to say the sentences as naturally as possible. Productions that had missing words, extra words, or extra-verbal interruptions like coughing and sneezing were coded as disfluent. These disfluent productions were re-recorded afterwards until all productions were fluent. In some cases, the speaker independently chose to re-record a sentence; in these cases, the final production was used.

Then, each target sentence was spliced out from these readings. These sentences' prosody was manipulated in Praat (Boersma & Weenink, 2009) in the following ways. First, all files were normed for intensity in order to prevent unintended intensity variation. There may be variation in intensity due to how far the speaker was from the microphone or other reasons, but the goal is for all the sentences to be as similarly and neutrally produced as possible. And because intensity is one of the prosodic features manipulated, it would be helpful to have all the sentences start with the same intensity; this way the manipulations don't have as much noise from the randomness of the sentences' original production. The average intensity across all of the original productions was 57.2 dB. So, to reduce the total amount of artificial manipulation in the norming, the mean intensity for all files was normalized to 57.2 dB. It is these intensity-normalized files that are used for the subsequent manipulations.

The specific prosodic manipulations described below have two motivations. First, the literature on prosody and discourse production indicates that a range of prosodic measures tend to correlate with types of discourse structure. For example, larger discourse boundaries tend to have longer pauses and higher subsequent pitch

(den Ouden 2009, Yule 1980), and hierarchically higher discourse segments tend to have higher mean pitch and longer preceding pauses (Hirschberg & Grosz 1992, Tyler under revision). As discussed in the introduction, hierarchically higher discourse segments generally convey broader, more general and more important information than hierarchically lower discourse segments. Specific definitions of discourse hierarchy vary from theory to theory, but they all capture intuitions about the relative level of detail of parts of the discourse. A second motivation for the specific prosodic manipulations comes from a study pursued at the University of North Carolina, which can speak directly to these discourse ambiguities and their production and perception (Tyler, Kahn, & Arnold, 2011). For the production component of the study, participants were presented with the ambiguous discourse texts, were asked to explain the two possible meanings and then to read them verbatim to make clear to a listener which meaning they intended to convey. These productions showed systematic correlates with prosodic measures like pause and sentence duration. When the productions were presented to listeners, those listeners were unable to identify which meaning the speaker intended to convey. But within the overall null effect, there was one speaker whose productions listeners were able to correctly identify 75% of the time. Inspection of the prosody of this speaker's productions revealed contrasts in terminal pitch contours and pause durations. The prosodic contrasts performed by this one especially successful speaker motivated the manipulations discussed below.

Sentence-final pitch contours for S1 and S2, but not S3, were manipulated. For each contour manipulation, it was important to have a consistent temporal window. Because the final contour generally began at the last stressed syllable, the window for manipulation was from the last stressed syllable to the end. In a pitch manipulation object in Praat, all pitch points from the last stressed syllable to the end of the file were selected, and all but the first and last of these pitch points were deleted. Then, the last pitch point was multiplied by a factor of the pitch of the last stressed syllable's pitch point, depending on whether the sentence was a first or second sentence in the discourse and whether it was a Coord or a Subord manipulation. For Coord manipulations, S1's final pitch point was multiplied by 1.6,

and S2's final pitch point was multiplied by 1.3. For Subord manipulations, S1's final pitch point was multiplied by .75, and S2's final pitch point was multiplied by 1.1. These multiples were motivated by the productions of the successful speaker in the UNC study (Tyler, et al., 2011), but also somewhat adjusted to sound more consistent and natural. This resulted in a discourse's final pitch contours for S1 and S2 looking like the following:

Figure 3.1: Sentence-final pitch contour manipulations for sentences 1 and 2 in the Coord condition

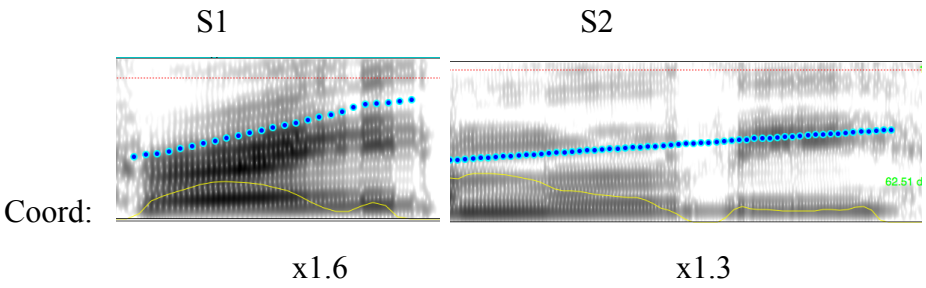
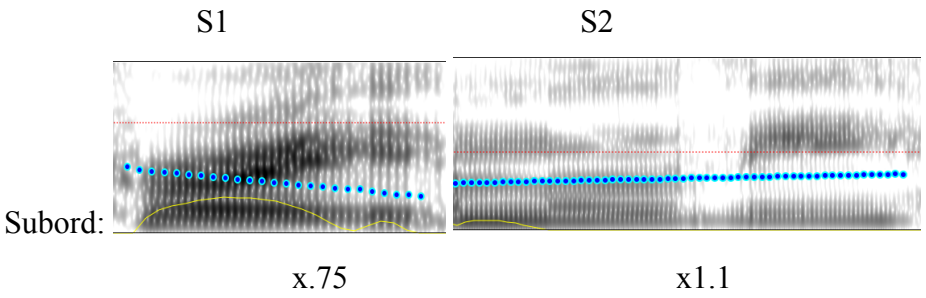


Figure 3.2: Sentence-final pitch contour manipulations for sentences 1 and 2 in the Subord condition



These final contours are actually linear slopes, and as a result do not have the non-linear movements in the original productions. But, this was a consistent way of constructing the final contour manipulation.

After assigning a new pitch contour to all S1s and S2s, the mean pitch and intensity of S2 and S3 were multiplied by 1.1 for the Coord condition and .9 for the Subord condition. For pitch, this was achieved with a Praat script that multiplied all pitch frequencies in the Manipulation object. For intensity, a Praat script using the *scale intensity* function reassigned mean intensity from the overall average of 57.2 to 57.2*1.1=62.92 for Coord or 57.2*.9=51.48 for Subord.

After all these manipulations on the sentences were complete, the files were re-examined and any silences at the edges were trimmed. This way the inserted pauses account for all of the silence between sentences. Then, the sentences were concatenated with pauses between them. For the Coord condition, the first pause P1 between S1 and S2 was 920ms and P2 between S2 and S3 was 400ms. For the Subord condition, the first pause P1 between S1 and S2 was 400ms and P2 between S2 and S3 was 20ms. Like the pitch manipulations, these pause durations were motivated by the productions of the successful speaker in the UNC study (Tyler, et al., 2011), but also somewhat adjusted to sound more consistent and natural. This resulted in the following structure for each discourse, by prosodic condition:

(3.3)	Coord:	S1	P1(920ms)	S2	P2(400ms)	S3
(3.4)	Subord:	S1	P1(400ms)	S2	P2(20ms)	S3

It was these final concatenated sound files that were presented to participants, and it was these prosodic contrasts that correspond to the predictor *prosody* in the statistical model.

Design

The question addressed in this experiment is whether prosody influences discourse interpretation, and therefore one predictor variable is the prosodic manipulation. Subjects hear one of two conditions, either with prosody manipulated to encourage coordinating or subordinating interpretations. Yes/no questions directly querying the interpretation were used, acknowledging that subjects may be more likely to answer yes to such a question not because of the meaning of the discourse but because of a bias towards answering yes. To control for this effect, question type was varied such that a yes answer alternatively indicated a coordinating or a subordinating interpretation. For example, they might hear the discourse in (3.1) while being presented one of the following two interpretation questions:

(3.5) Coord bias: Did Sally mean that she read about housing prices and watched a cool documentary separate from history class?
(3.6) Subord bias: Did Sally mean that she read about housing prices and watched a cool documentary in history class?

For the Coord-biasing question, a yes answer indicates a Coord interpretation and a no indicates a Subord interpretation. For the Subord-biasing question, a yes answer indicates a Subord interpretation and a no indicates a Coord interpretation. The predicted answer would be a Coord interpretation after hearing Coord prosody, and a Subord interpretation after hearing Subord prosody.

The design was 2x2, crossing prosody and question type. The 48 target discourses were separated into four blocks of 12 discourses, with the blocks always presented in the same order. Each block always contained the same discourses but was randomized within. This allowed for comparison between blocks of 12 to see if presentation order had an impact on prosody's effect on interpretation. From piloting, it appeared that participants may not initially use prosody in their interpretation but with repetition they begin to do so. This blocking was included to check this potentiality. Within each block of 12, there were 3 discourses in each cell of the 2x2 design crossing prosody and question-bias. Four groups of participants were created, with each group seeing 12 discourses in each cell of the 2x2 design. This way, each participant saw an equal number of each prosodic condition and each question type. The groups were counterbalanced so each discourse was presented an equal number of times. Each participant group and presentation quarter was also assigned a balance of discourses with a range of ambiguity, from those where both the Coord and Subord meanings were nearly equally accessible to those where either Coord or Subord was preferred more than the other. There were no fillers. While fillers with more blatant prosodic contrasts could have been included, this may have prevented participants from paying attention to the more subtle contrasts in the prosodic manipulations of interest. Fillers with different kinds of structural contrasts could also have been included, but this would have made it more difficult for listeners to get used to the discourses and the elicitation questions. The discourse ambiguities are difficult enough to process, giving listeners a chance to get comfortable with the relevant

meaning contrast and the elicitation questions should reduce noise in the data due to processing difficulties.

Preceding the 48 target discourses were 4 training discourses, one in each cell of the 2x2 design. All participants saw the same training discourses, in the same order, with the same questions and same prosody. For participants, the first four discourses were indistinguishable from the remaining 48. The training discourses, which were not included in the final analysis, provided a chance for participants to get some basic familiarity with the task before their data counted.

Procedure

Participants listened to the discourses and answered all associated questions in a Qualtrics survey (Qualtrics Labs Inc., 2009). They were told they were going to hear a series of stories told by a woman named Sally, and that those stories could be interpreted multiple ways. Their task would be to answer questions about how they interpreted the stories. They were instructed to adjust the volume to comfortable levels, and that they could listen to each discourse as many times as desired. For each discourse, listeners answered three questions. First, they saw a page with an audio play button and an interpretation question that queried whether they got the Coord or Subord interpretation of the discourse. The interpretation question was a yes/no question that asked *Did Sally mean that [Coord Interpretation]* or *Did Sally mean that [Subord Interpretation]*.

After answering the interpretation question, they advanced a screen and were asked how confident they were in their interpretation on a 1-100 scale. And finally, they answered a factual comprehension question about the discourse they just heard to check whether they were paying attention. Participants saw only one question on the screen at a time, could not advance without answering the question, and could not go back and change previous answers.

For example, on the first screen they might hear the discourse in (3.1) while being presented one of the interpretation questions in (3.5) or (3.6). Then, they would be asked:

(3.7) How confident are you in that choice?

And finally, they would answer a factual comprehension question like the following:

(3.8) Did Sally pick up some beer?

After answering all three questions, they would advance to the next discourse and continue.

Predictions

Prosody is predicted to bias interpretation, with listeners providing more Coord interpretations when they hear Coord prosody than when they hear Subord prosody. Conversely, listeners are predicted to provide more Subord interpretations when they hear Subord prosody than when they hear Coord prosody.

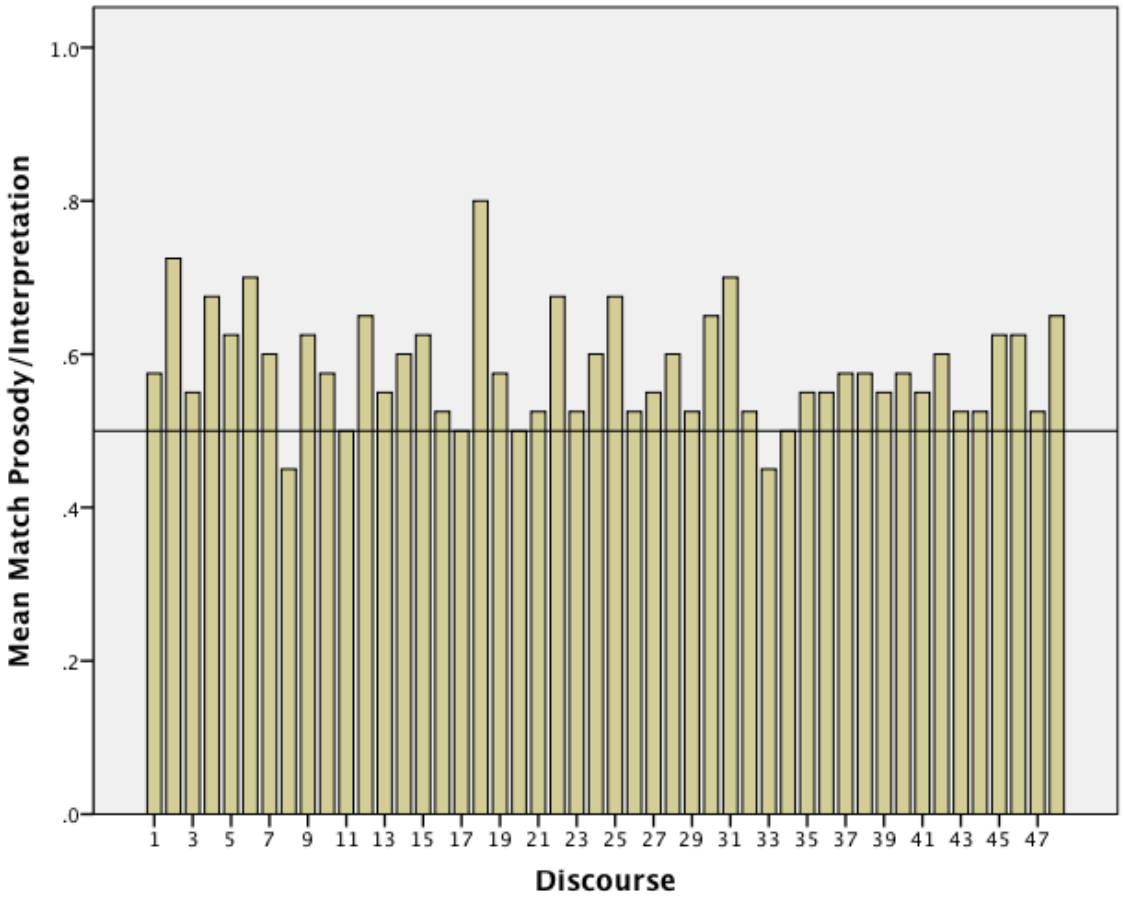
Results

The comprehension questions after each discourse were intended to help separate participants who were paying attention from those who were inattentively racing through the survey. Because I wanted participants who were paying attention, those who got a large number of comprehension check questions wrong would be excluded from the analysis. Results showed that all forty participants got at least 36 comprehension questions correct, suggesting they were all attentively participating. Therefore, the data from all forty participants were included in the analysis.

The core question in this study is whether prosody biases discourse interpretation. This was tested first by checking if there was a significant chance that listeners' interpretations would match the prosody, where a match would be a Coord interpretation after hearing Coord prosody or a Subord interpretation after hearing Subord prosody. The statistical model used was a Generalized Linear Mixed Model with a binary logit link function; it had a random effect for each subject and a binary

dependent variable of match vs. mismatch between prosody and interpretation. Prosody could have been tested as a predictor of interpretation, though this result is identical to testing relative likelihood of match vs. mismatch. And using match has the benefit of making subsequent analyses simpler. By collapsing all cases of matching prosody and interpretation into their own outcome (match), subsequent analyses using other predictors are getting more directly at the core question of whether those predictors affect the likelihood of a listener's interpretation matching the prosody they heard, regardless of which prosody they heard. This avoids having to do analyses for each other predictor variable (e.g. discourse ambiguity, question bias, demographic factors) with both Coord and Subord prosody independently. For a discussion of the benefits of mixed effects models with binary outcomes relative to other repeated measures models, see Quené & van den Bergh (2008) and Jaeger (2008). When a random effect was also included for item (i.e. discourse), the model could not detect any variance from item to item (see Figure 3.3 for a graph of match rate by item). There was non-convergence and the validity of the model fit was uncertain. Therefore, the random effect for items was removed. The standard deviation of match rate from item to item was 0.072.

Figure 3.3: This graph plots mean match rate (prosody/interpretation) on the y-axis and discourses (the items) on the x-axis. The graph shows the variability from item to item in how likely listeners were to make the interpretation predicted by the prosodic manipulation. A horizontal line at 0.5 match rate is included for reference.



The issue of how to deal with non-convergence of random effects in mixed models has been the subject of recent work in statistical methods. In general, research in psycholinguistics has argued for the inclusion of random effects for subjects and items as a way of controlling for variation between subjects and between items (Clark, 1973; Jaeger, 2008). But sometimes the inclusion of these variables leads to non-convergence, meaning the model cannot converge on best estimates for these variables and the resulting model fit is uncertain. What leads to non-convergence and how to deal with it in models with random effects has been discussed recently in an article by Barr, Levy, Scheepers, & Tily (under review). Barr et al. mention that "although the issue seems not to have been studied systematically, it is our impression that fitting maximal LMEMs is less often successful for categorical data than for

continuous data" (p. 33). This suggests the binary outcomes in this chapter's study may be more likely to result in non-convergence than continuous outcomes. Other common causes of non-convergence can be having empty cells, an insufficient number of total observations or variables measured on a small scale (Allison, 2004; Kahn, 2012). There are a number of strategies for how to deal with non-convergence, including identifying data problems, increasing the number of iterations and multiplying the dependent variable by a large number (Allison, 2004; Barr, et al., under review; Kahn, 2012), but even with such strategies the model may not converge. In this study, it is unlikely the non-convergence is a result of insufficient data because there are many observations (48 observations per participant, with 40 participants). The non-convergence is also unlikely to be due to empty cells, as all discourses received at least some matching and some mismatching interpretations. Therefore, solutions for non-convergence due to sample size or empty cells were not employed.

In such a case, Barr et al. say "the next step is to seek out the next most complex model that does converge" (p. 34). Similarly, Kahn recommends removing effects in the model "until the most-complex model possible successfully converges" (2012). And Allison (2004) writes "the most widely used method … is simply to delete from the model any variables whose coefficients did not converge" (p. 248). In this study, the non-convergence occurred only when a variable for random item effects was included. Even when the random subjects effect was removed, the model with the random items effect still did not converge. It seems the non-convergence is a result of the random item effect independent of the random subject effect. For these reasons, the variable for random item effects was excluded from the statistical model, resulting in a model that successfully converged.

Results showed chance of match was significant ($t=3.931$, $p<.001$), with a positive coefficient, meaning prosody and interpretation were significantly more likely to be matched than not. Table 3.1 shows the raw count of matches and mismatches for both Coord and Subord prosody.

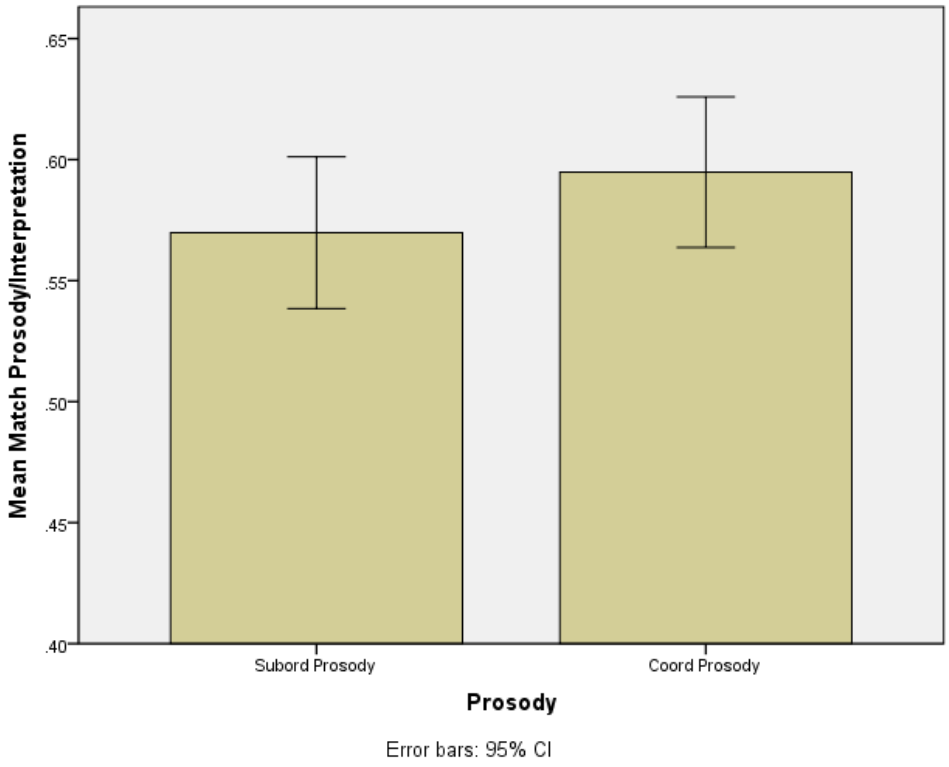
Table 3.1: Frequency table of prosody (by row) crossed with match (by column).

	Match	Mismatch	Total

Coord Prosody	571	389	960
Subord Prosody	547	413	960
Total	1118	802	1920

The graph in Figure 3.4 plots prosody on the x-axis and match between prosody and interpretation on the y-axis. Match rates for both prosodic conditions are above chance. The overlapping error bars also suggest the two prosodic conditions do not have different match rates. This was tested statistically by adding prosody into the model as a predictor of match. Results showed no significant difference between the two conditions of prosody (F=1.274, p=.259), indicating that neither Coord nor Subord prosody was significantly better than the other at getting a matching interpretation.

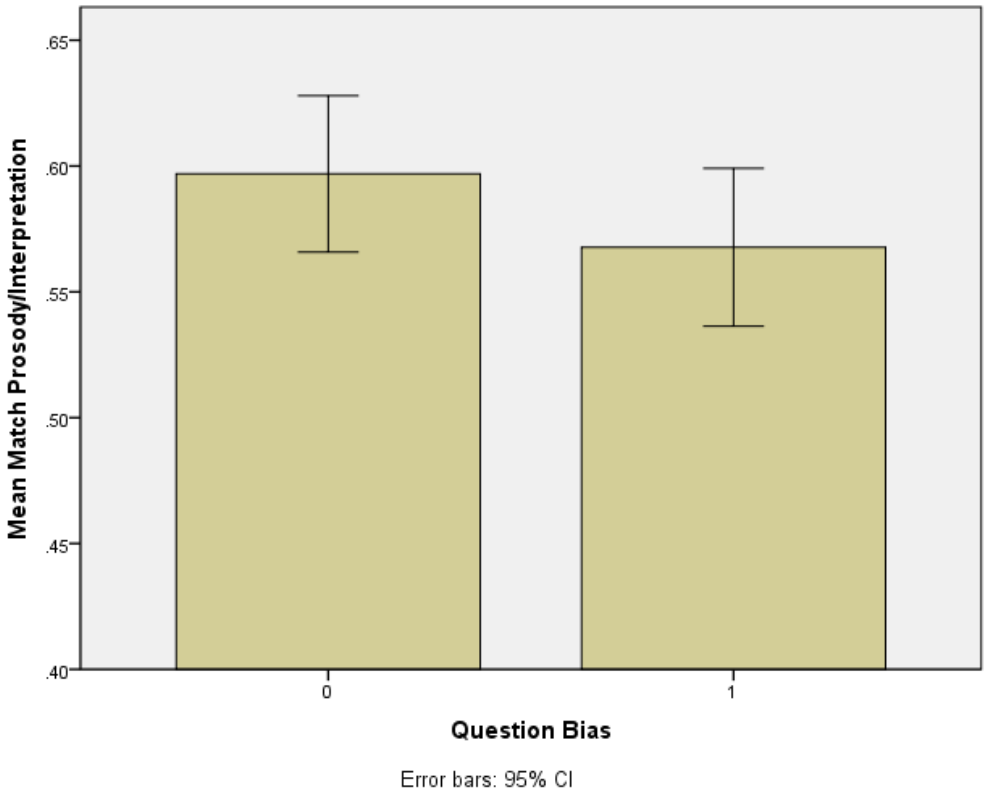
Figure 3.4: This graph plots the two prosodic conditions on the x-axis and match rate on the y-axis (1=match, 0=mismatch), with 95% confidence intervals.



Participants' interpretations of the ambiguous discourses were elicited with two kinds of questions, one where yes indicates a Coord interpretation and another

where yes indicates a Subord interpretation. These two questions were included to try to control for a potential confound of participants tending to respond yes more often than no, all else equal. When question bias was included in the model, results showed no effect of question bias on likelihood of match (F=1.734, p=.188). Therefore, the bias of the question did not affect prosody's ability to bias discourse interpretation. The graph below shows match on the y-axis with question bias on the x-axis.

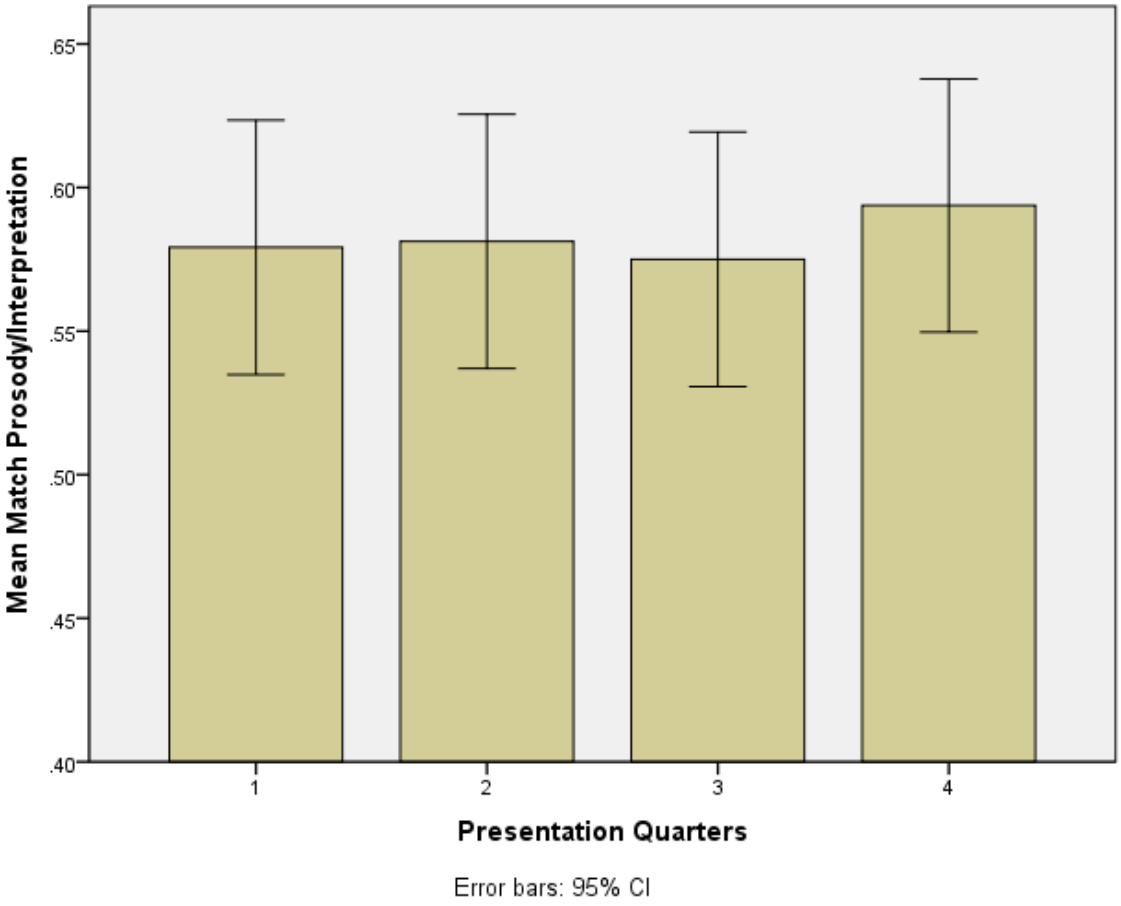
Figure 3.5: This graph plots the two conditions for question bias on the x-axis and mean match on the y-axis (1=match, 0=mismatch), with 95% confidence intervals.



Another question about prosody's effect on discourse interpretation is whether participants are using prosody consistently across the whole study, only in the beginning or only at the end. This question about how prosody's effect on interpretation could change over time was explored by comparing the four quarters of the experiment. Every subject saw the same 12 discourses in each quarter. To test for changes in performance over time, a continuous variable for presentation order was included in the model. This variable was found not to be significant (F=.143, p=.705),

suggesting that when in the study a discourse was confronted did not effect the likelihood of match between prosody and interpretation. Therefore, psychology subject pool subjects in a laboratory setting appear to use prosody consistently over time, neither showing a learning effect nor a fatigue effect. The graph below shows match between prosody and interpretation on the y-axis with presentation quarters on the x-axis.

Figure 3.6: This graph plots the four quarters in which discourses were presented on the x-axis and mean match on the y-axis (1=match, 0=mismatch), with 95% confidence intervals.

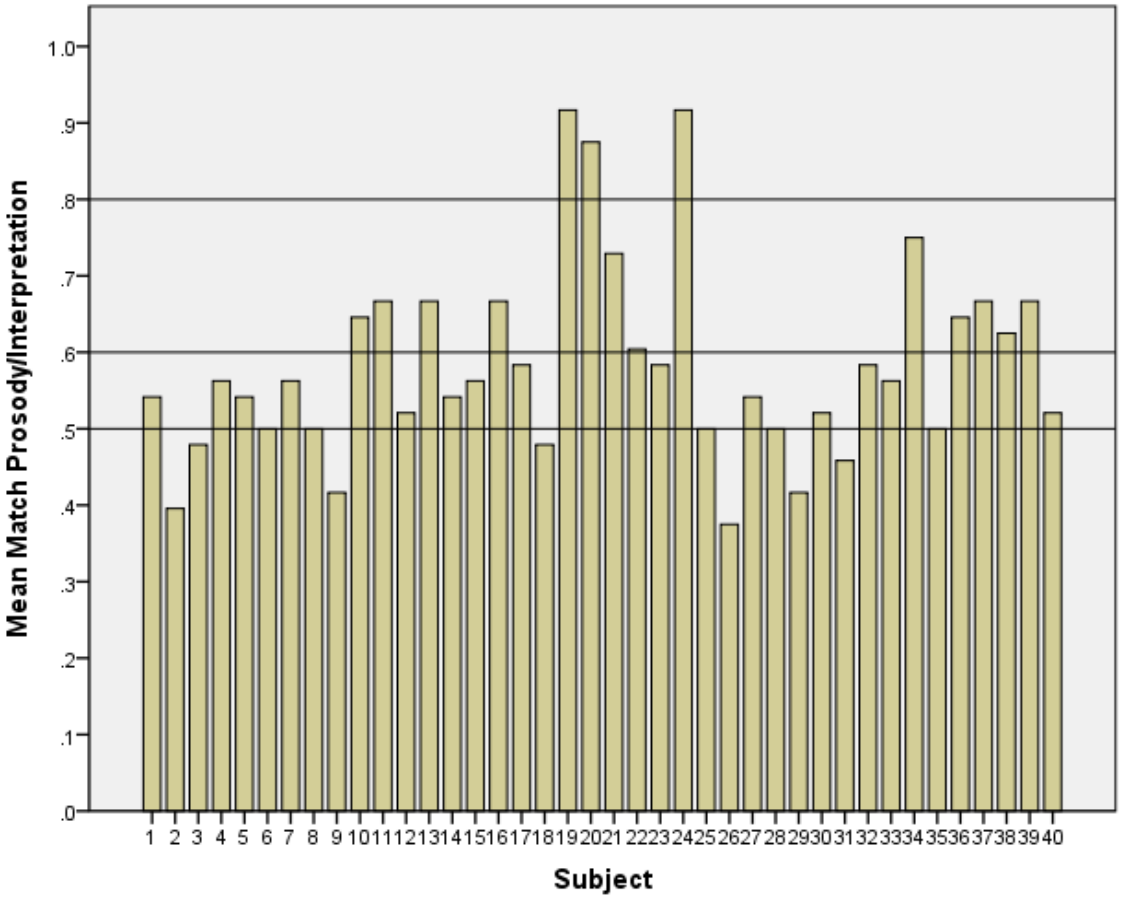


It is also possible that the ability of prosody to disambiguate discourse depends on the practical ambiguity of the discourses themselves; the lexical material of one discourse could bias so much towards one interpretation that prosody would have no effect, while when multiple meanings are more equally accessible a factor like prosody could have an impact. As described in Appendix E, 102 discourses were normed for how available the two desired interpretations were, and the 52 most

ambiguous discourses were selected (48 target discourses and four practice discourses). Within these 48 target discourses, there was variation from one discourse to the next in terms of underlying preferences for one interpretation or the other. A covariate was included in the model that measured the absolute value of the difference between the number of people who chose Coord and Subord interpretations. The scale of this variable was from zero (equibiased) to 26 (most biased), with higher values indicating greater bias towards either Coord or Subord. The degree of a discourse's underlying ambiguity was not found to affect participants' ability to match discourse interpretation with the prosody ($F=.034$, $p=.854$). This suggests that the degree of ambiguity did not affect the interpretation of this set of discourses.

Results so far have been discussed for the participant population as a whole, but there is substantial variation from subject to subject. The graph in Figure 3.7 plots on the y-axis each subject's match rate, with higher numbers indicating greater match. The study's 40 subjects are lined up along the x-axis.

Figure 3.7: This graph plots each subject on the x-axis and mean match on the y-axis (1=match, 0=mismatch). Horizontal lines at 0.5, 0.6 and 0.8 match rate are included for reference.



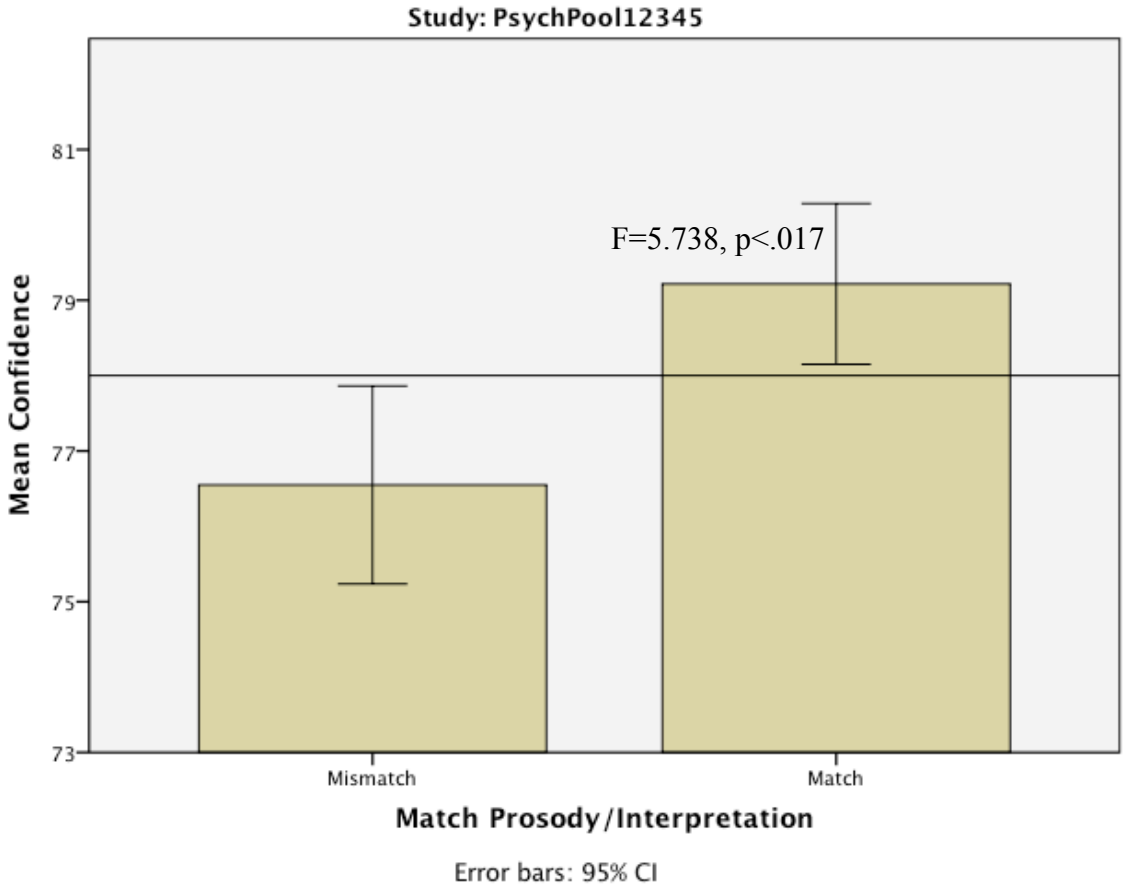
As visible in the graph, most subjects cluster around or above 0.5. Three participants show much greater match, suggesting some participants are more attuned than others to the prosodic contrasts in this study. These differences cannot be accounted for by the basic demographic information, including education level, gender, age, and mono- vs. multi-lingual status, that was elicited at the end of each survey. When these measures were included as main effects in the statistical model, none were significantly predictive of match.

Confidence

In addition to their discourse interpretation judgment, participants were also asked how confident they were in that judgment. A Linear Mixed Model was run with random effects for both subject and item, and confidence as a continuous dependent

variable. Results show match between prosody and interpretation was predictive of confidence (F=5.738, p=.017), indicating that one was more confident in one's judgment when that judgment used the prosody as predicted. The graph below plots confidence on the y-axis and match on the x-axis, showing greater confidence when making a matching interpretation.

Figure 3.8: This graph plots the two conditions for match on the x-axis and confidence on the y-axis, with 95% confidence intervals. A horizontal line at 78 is included for reference.



Discussion

The results of this study indicate that the five prosodic manipulations collectively bias the interpretation of ambiguous discourse. Furthermore, both sets of manipulations have a significant impact on interpretation, meaning the overall effect

is not the result of only one of the two sets of manipulations. And the effect for the Coord prosody is not significantly different from the effect for the Subord prosody, meaning not only are both sets of manipulations contributing but they are contributing equally.

A couple limitations to this study remain. One concern is the quality of the sound manipulations, and whether that has an impact on these findings. It is unclear how much listeners could pick up on the sound files being manipulated as opposed to natural, and whether that might affect listeners' behavior. Similarly, it is possible that the text of some discourses seemed more natural, i.e. more likely for a real person to say, than others. And while this discourse text naturalness was not normed, it could potentially have an effect on how listeners interpret the discourses and how they integrate cues like prosody.

It is also important to emphasize that this study only tested the effects of these prosodic manipulations on discourse interpretation, and cannot speak to what other manipulations might do. Perhaps other prosodic manipulations that were not tested here could have a similar effect. That is, these prosodic features may not be the only means by which prosody could affect interpretation of these discourses.

Another limitation of this chapter's study is the use of a set of five prosodic manipulations, obscuring the contribution of each one. Prior research shows that prosodic effects on discourse interpretation seem to be stronger when more cues are used in tandem. Silverman (1987) showed an 84% disambiguation success rate with three cues (two pitch and one pause), but with the pause duration contrast neutralized the success rate dropped to 71% (p. 6.27). Mayer et al. (2006) similarly showed a drop in disambiguation success when fewer prosodic cues were included. Their initial study had both a pitch and pause manipulation, showing a significant effect of prosody on interpretation; but when the same study was re-run with either just pause or just pitch, the effect disappeared. Silverman explains that "this is hardly surprising: the more redundantly the prosodic structure is encoded in the acoustic signal, the more likely it is that listeners will be able to recover it and use it during speech perception" (p. 6.27). He argues the higher success rate is because each listener is likely to be better able to recover the prosodic structure. An alternative explanation is

that there is actually inter-subject variability in which cues they pay attention to. It is possible the drop in success rate when fewer prosodic contrasts are included is due to some listeners no longer having the cues that were relevant for them. For example, it is possible some of the participants in Silverman's studies paid attention most to pause duration. When the pause duration contrast was neutralized, these participants would no longer be able to use pause duration in their interpretation. It is likely these two explanations are both right, as all participants are likely to be able to perceive meaningful contrasts in different prosodic measures to some degree, but the degree each participant draws on each measure is also likely to vary.

It is quite possible that the effect I have found with this set of five prosodic manipulations will diminish or disappear when fewer prosodic cues are included. Nevertheless, a fuller understanding of prosody's effect on discourse interpretation needs a better account for the contribution of each prosodic measure to the overall effect. Is one prosodic measure driving it and the others don't matter much? Are none significant alone but when combined they have an effect? Is it purely a cumulative effect, where each manipulation has some weight and there is a threshold where listeners start to use prosody consistently? For the type of ambiguity used in this study, does pitch contour, pause duration or mean pitch/intensity matter most? The next chapter will address questions like these by changing which prosodic cues are available to listeners and gauging their use of prosody to disambiguate discourse.

The next chapter will re-run the study discussed in this chapter as well as run follow-up studies contrasting subsets of prosodic manipulations. Instead of using participants from the University of Michigan Psychology Subject Pool, participants will be drawn from Amazon's Mechanical Turk. Using Mechanical Turk offers the practical benefits of being able to run many subjects quickly and inexpensively relative to bringing people into the lab. The participants recruited through Mechanical Turk also tend to be more diverse in terms of demographic factors like age and education level. If the effect in this chapter is replicated with Mechanical Turk participants, then the studies testing subsets of prosodic manipulations can be run using Mechanical Turk instead of bringing people into the laboratory. This would

facilitate and expedite data collection. In the discussion at the end the next chapter, these studies will be compared with Mayer et al. 2006) and Silverman (1987).

Chapter 4

Isolating the Synthesized Prosodic Manipulations Influencing the Interpretation of Discourse Ambiguities (Amazon Mechanical Turk)

The previous chapter presented a study that explored the ability of synthetic manipulations of prosody to bias the interpretation of ambiguous discourse. The prosodic contrast had an overall effect on interpretation, with discourses in the Coord prosody condition resulting in more Coord interpretations than those in the Subord prosody condition. But because there were five total prosodic manipulations that constituted the prosodic contrast, it was unclear which one or ones contributed to the effect. This chapter presents the results of a series of follow-up studies that test various combinations of prosodic contrasts, which will help isolate which ones are driving the effect on discourse interpretation.

Running a series of follow-up studies runs into the practical problem, when using participants from the Psychology Subject Pool, of having to bring many people into the laboratory. This can be time-consuming and easily exhaust the participants available. One alternative source of participants is the online labor marketplace Amazon Mechanical Turk, a crowdsourcing platform where requestors post tasks that workers can complete for a set fee. Mechanical Turk, an increasingly popular source of participants for behavioral research (Kittur, Chi, & Suh, 2008), has the benefits of a more diverse subject pool, fast data collection and inexpensive rates. If the effects found in the previous chapter with participants on Mechanical Turk can be replicated, then Mechanical Turk would be shown to be a reliable source of participants for this research. This would facilitate running follow-up studies.

This chapter presents the results of five studies using participants from Mechanical Turk. One study is identical to the one presented in the previous chapter, which will serve to demonstrate that the effect found with participants from the Psychology subject pool is replicable with participants from Mechanical Turk. There are four other studies that differ only in the set of prosodic manipulations that constitute the contrast between the Coord and Subord prosody conditions. These studies help narrow down which prosodic contrast is driving the effect on discourse interpretation.

Amazon Mechanical Turk

Amazon's Mechanical Turk service is an online labor marketplace where anyone (called "requestors") can post tasks ("HITs", or Human Intelligence Tasks) that others (called "workers") can log on and complete for a set fee. This "crowdsourcing" platform has been growing in popularity as a source of participants in behavioral research, offering the benefits of participant diversity, speed and cost (for more discussion of the benefits of Mechanical Turk for research, see Kittur, et al., 2008). It has also begun to be used in linguistics research specifically (Gibson, Piantadosi, & Fedorenko, 2011; Sprouse, 2011a; Sprouse, 2011b; Sprouse & Almeida, to appear).

But as it is still a relatively new source of data, there are a number of potential concerns about its usefulness for research purposes. Fortunately, most of these concerns seem to either be unsubstantiated by empirical research or survivable by clever design (Paolacci, Chandler, & Ipeirotis, 2010). For example, one concern may be that the Mechanical Turk population is not representative and thus results are not generalizable. But experiments run with internet samples tend to be diverse and achieve similar results compared to traditional samples (Gosling, Vazire, Srivastava, & John, 2004). And because you can restrict participation to subjects from a particular country or to those who meet other specified criteria, you still have control over what population you would like to focus on. There does appear, however, to be a recent shift from Mechanical Turk subjects being largely from the United States to an

increasing number of participants from India (Paolacci, et al., 2010; Ross, Irani, Silberman, Zaldivar, & Tomlinson, 2010).

Another concern is the validity of the results. Because Mechanical Turk studies involve subjects participating anonymously on the internet, those subjects may not pay as close attention to the task. This concern appears not to be a problem, as many studies using traditional sources of subjects have been replicated with Mechanical Turk subjects (Akkaya, Conrad, Wiebe, & Mihalcea, 2010; Paolacci, et al., 2010; Snow, O'Connor, Jurafsky, & Ng, 2008). And while there are real concerns about participants "gaming" the system (Downs, Holbrook, Sheng, & Cranor, 2010), there are ways to filter out these participants through filter trials or comprehension check questions to assess participant effort and attention to task. Sprouse (2011b) found, aside from having to exclude slightly more participants, Mechanical Turk data are almost indistinguishable from data collected in more traditional laboratory experiments.

One advantage of Mechanical Turk is that it avoids potential experimenter bias that can affect laboratory studies (Paolacci, et al., 2010). When a subject comes to a lab for an experiment and interacts with the experimenter, there is variability in that interaction from participant to participant. With Mechanical Turk, participants interact with the same online interface and thus are not affected by variation in experimenter behavior.

It seems then that Mechanical Turk can be a reliable, efficient and productive source of participants for behavioral research. In fact, as long as care is given to the study's design (Kittur, et al., 2008), this is the exact conclusion of a number of studies specifically assessing Mechanical Turk's reliability for research (Akkaya, et al., 2010; Paolacci, et al., 2010; Snow, et al., 2008; Sprouse, 2011b).

Methodology

The studies presented in this chapter are identical to the study in the previous chapter, with two exceptions. First, instead of using the University of Michigan Psychology Subject Pool, participants are drawn from Amazon Mechanical Turk,

with its concomitant differences in payment, setting and other factors. Second, the prosodic contrast is constructed of different sets of prosodic manipulations, except for the one study that is a replication and so contains the same prosodic manipulations.

Participants

All participants in these studies participated via Amazon's Mechanical Turk service in exchange for two dollars. Each subject's IP address was compared against a list of all IP addresses of participants from all versions of these studies. If an IP address appeared for more than one subject's data, then the first survey they took (as identified by a timestamp) was kept but all subsequent data were excluded. Though it's possible an IP address could randomly be the same, because they are dynamically assigned, Berinsky, Huber, & Lenz (2010) found repeated IP addresses were rare and so not a big problem. Of a total of 323 surveys taken in the five studies in this chapter via Amazon's Mechanical Turk service, nine of those surveys were taken by participants who were not participating for the first time. The data in these nine surveys were excluded. One survey was taken by a participant who reported not being a native speaker of American English, while all other subjects reported being native speakers of American English. The non-native speaker's data were also excluded. In the end, 313 total surveys were included in the analysis.

Materials

The prosodic contrast in the previous chapter was a combination of five different prosodic manipulations (for details on how they were synthesized, see chapter 3):

(4.1) The five prosodic manipulations in the perception studies.

1: Terminal pitch on S1
2: Terminal pitch on S2
3: Pause duration between S1 and S2
4: Pause duration between S2 and S3
5: Mean pitch and intensity on S2 and S3

For ease of reference, I will refer to each study as a compound of its participant pool and the set of prosodic manipulations it contained. For example, the previous chapter's study is PsychPool12345 because it used participants from the Psychology Subject Pool and contained all five prosodic manipulations. This chapter presents the results of five new studies that use participants from Mechanical Turk: MTurk12345, MTurk12, MTurk1, MTurk2, and MTurk345. So, for example, MTurk12 contains a prosodic contrast that differs only in S1 and S2 terminal pitch, while MTurk12345 replicates PsychPool12345 but with Mechanical Turk participants.

MTurk345 contrasted along the dimensions 3,4 and 5 listed above. This involved contrasting pause durations and mean pitch/intensity of S2 and S3 while holding terminal pitch on S1 and S2 constant. Holding the pitch contour constant raises the thorny methodological issue of which terminal pitch to hold it constant to. In an effort to not be particularly like either manipulated version of terminal pitch, the pitch from the last stressed syllable to the end was flattened. The flattening was achieved by forcing the f0 contour from the last stressed syllable to the end to stay at the same Hz value. Both the Coord and Subord conditions had this same flat terminal pitch.

MTurk2 contrasted only in the terminal pitch on S2. Therefore, S1 terminal pitch, pause durations, and mean pitch/intensity were all held constant. All pause durations were set at 400ms, the original mean pitch and intensity were left unmanipulated, and S1 terminal pitch was flat. MTurk12 contrasted in terminal pitch on both S1 and S2. All pause durations were set at 400ms, and the original mean pitch and intensity were left unmanipulated. And finally, MTurk1 contrasted only in the terminal pitch on S1, with S2 terminal pitch, pause durations, and mean pitch/intensity held constant.

Design

Participants answered three questions for each of the 48 target discourse discourses. First, they provided their interpretation of the meaning of the discourse. Second, they indicated their confidence in their answer (1-100). And finally, they answered a simple comprehension check question intended to filter out those

participants who were not paying attention. For details about the questions, see chapter 3.

The design was 2x2, crossing prosody and question type. The order of presentation was controlled by separating the 48 target discourses into four blocks of 12 discourses, with the blocks always presented in the same order. Each block always contained the same discourses but was randomized within. This allowed for comparison between blocks of 12 to see if presentation order had an impact on prosody's effect on interpretation. From piloting, it appeared that participants may not initially use prosody in their interpretation but with repetition they begin to. This blocking was included to check this potentiality. Within each block of 12, there were 3 discourses in each cell of the 2x2 design crossing prosody and question-bias. Four groups of participants were created, with each group seeing 12 discourses in each cell of the 2x2 design. This way, each participant saw an equal number of each prosodic condition and each question type. The groups were counterbalanced so each discourse was presented an equal number of times. Each participant group and presentation quarter was also assigned a balance of discourses with a range of ambiguity, from those where both the Coord and Subord meanings were nearly equally accessible to those where either Coord or Subord was preferred more than the other. There were no fillers.

Preceding the 48 target discourses were 4 training discourses, one in each cell of the 2x2 design. All participants saw the same training discourses, in the same order, with the same questions and same prosody. For participants, the first four discourses were indistinguishable from the remaining 48. The training discourses, which were not included in the final analysis, provided a chance for participants to get some basic familiarity with the task before their data counted.

Procedure

The participants from Mechanical Turk took the same survey and received the same instructions as the Psychology Subject Pool participants in the previous chapter. While it is unknown in what exact context they took the survey, they were instructed to be in a quiet environment and have good headphones.

Predictions

Prosody is predicted to bias interpretation, with listeners providing more Coord interpretations when they hear Coord prosody than when they hear Subord prosody. Conversely, listeners are predicted to provide more Subord interpretations when they hear Subord prosody than when they hear Coord prosody.

Results

The comprehension questions after each discourse were intended to help exclude participants who were not paying attention. Because it was preferred to have participants who were paying attention, those who got a large number of comprehension check questions wrong would be excluded from the analysis. Results showed that of the 314 participants in the data set, eleven subjects performed poorly on the comprehension questions (>20% incorrect). These eleven subjects' data were excluded from the analysis, resulting in a total of 303 participants in the final analysis.

Demographic data were collected from each participant at the end of the survey. These data are aggregated in Table 4.1.

Table 4.1: Demographic data for participants by row, with a column for each study.

	MTurk12345	MTurk12	MTurk1	MTurk2	MTurk345	PsychPool12345
Total participants	60	58	60	63	61	40
% male	43%	33%	33%	32%	30%	33%
Multilingual (#yes)	10	11	3	13	7	14
Age (mean) (min-max) (StdDeviation)	 35 18-59 11.03	 34 18-61 11.83	 33 18-62 10.44	 35 18-64 13.24	 36 19-71 11.95	 18 17-21 0.863
Education[4]	1: 1 2: 11 3: 18 4: 19 5: 4 6: 7	1: 1 2: 9 3: 17 4: 19 5: 5 6: 7	1: 3 2: 10 3: 21 4: 14 5: 4 6:8	1: 0 2: 10 3: 19 4: 20 5: 6 6: 9	1: 1 2: 4 3: 20 4: 22 5: 4 6: 10	1: 0 2: 20 3: 20 4: 0 5: 0 6: 0

Participants in the Mechanical Turk studies were distributed in terms of gender much like PsychPool12345, usually with approximately twice as many women than men. The Mechanical Turk studies had a much wider age range, averaging around 35 with a standard deviation greater than 10. They also had participants with a range of levels of educational attainment. Participants in PsychPool12345 were more homogenous, in terms of age as well as education.

The question of whether prosody biases discourse interpretation was tested first by checking if listeners' interpretations systematically matched the prosody they heard. The statistical model used was a Generalized Linear Mixed Model with a binary logit link function, run in SPSS 19; it had a random effect for each subject and a binary dependent variable of match vs. mismatch between prosody and interpretation. Match was defined as participants supplying Coord interpretations upon hearing Coord prosody or Subord interpretations upon hearing Subord prosody. When a random effect was also included for item (i.e. discourse), the model could not detect any variance from item to item for all but one study, i.e. the model did not converge. For the studies that did not reach convergence, this meant the validity of

[4] Legend for Education question: 1=Did not complete high school; 2=High school; 3=Some undergraduate education; 4=Undergraduate degree; 5=Some graduate education; 6=Graduate degree.

the model fit was uncertain and so the random effect for items was removed (see Chapter 3 for a discussion of non-convergence in random effects modeling). The one exception was MTurk1, which had no error when a random item effect was included. In Table 4.2, standard deviations between items are broken out by study. PsychPool12345 has the highest standard deviation but also a smaller sample size. Of the studies with Mechanical Turk participants, the study with the highest standard deviation between items was the one that reached convergence when a random item effect was included in the statistical model. This suggests that insufficient item variance accounted for the non-convergence in models with a random item effect for the other studies.

Table 4.2: Descriptive statistics for each study, with standard deviations from item to item.

Study	N	Minimum	Maximum	Mean	Item to Item Std. Deviation
MTurk12345	2880	.43	.70	.553	.052
MTurk12	2784	.45	.67	.550	.056
MTurk1	2880	.38	.68	.543	.070
MTurk2	3024	.37	.60	.495	.055
MTurk345	2928	.38	.64	.498	.066
PsychPool12345	1920	.45	.80	.585	.072

Additionally, all models were tested with just a random item effect and no random subject effect, and with these models the studies MTurk1 and MTurk345 converged. This shows that, with the exceptions of MTurk1 and MTurk345, the models can converge on subject variance but cannot converge on item variance, suggesting the random subject effect is a more important component of the model. As a result, when MTurk1 data is analyzed alone, a random item effect was included. In all other cases, there was no random item effect.

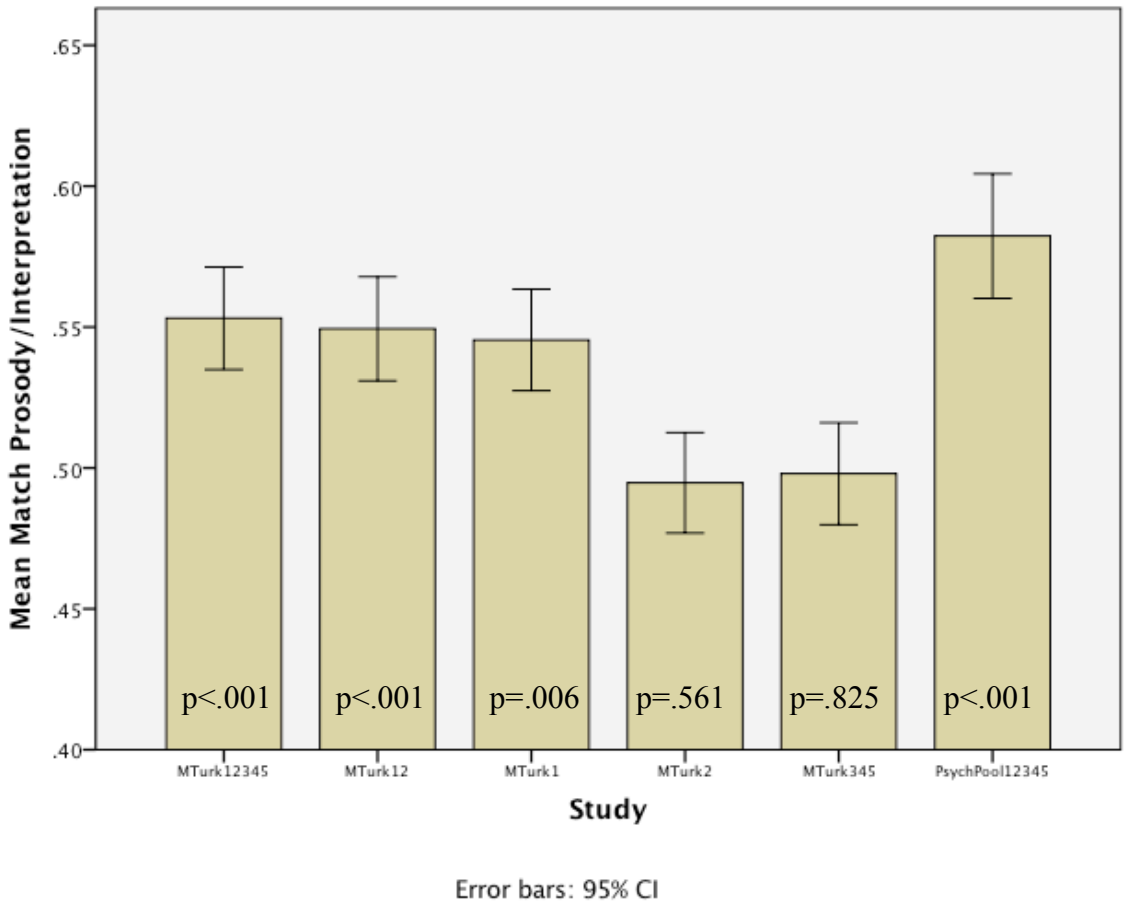
The results in Table 4.3 show three studies with Mechanical Turk participants are significantly more likely to match prosody and interpretation than to mismatch: MTurk12345, MTurk12 and MTurk1.

Table 4.3 Results testing whether likelihood of match was different from likelihood of mismatch, with a column for each study. Also included are frequencies for match and mismatch for both Coord and Subord prosody.

	MTurk12345	MTurk12	MTurk1	MTurk2	MTurk345	PsychPool12345
Match	t=4.415, p<.001	t=3.902, p<.001	t=2.743, p=.006	t=-.582, p=.561	t=-.222, p=.825	t=3.931, p<.001
Coord prosody (Match/ mismatch)	747/693	746/646	747/693	687/825	686/778	571/389
Subord prosody (Match/ mismatch)	846/594	784/608	818/622	809/703	772/692	547/413

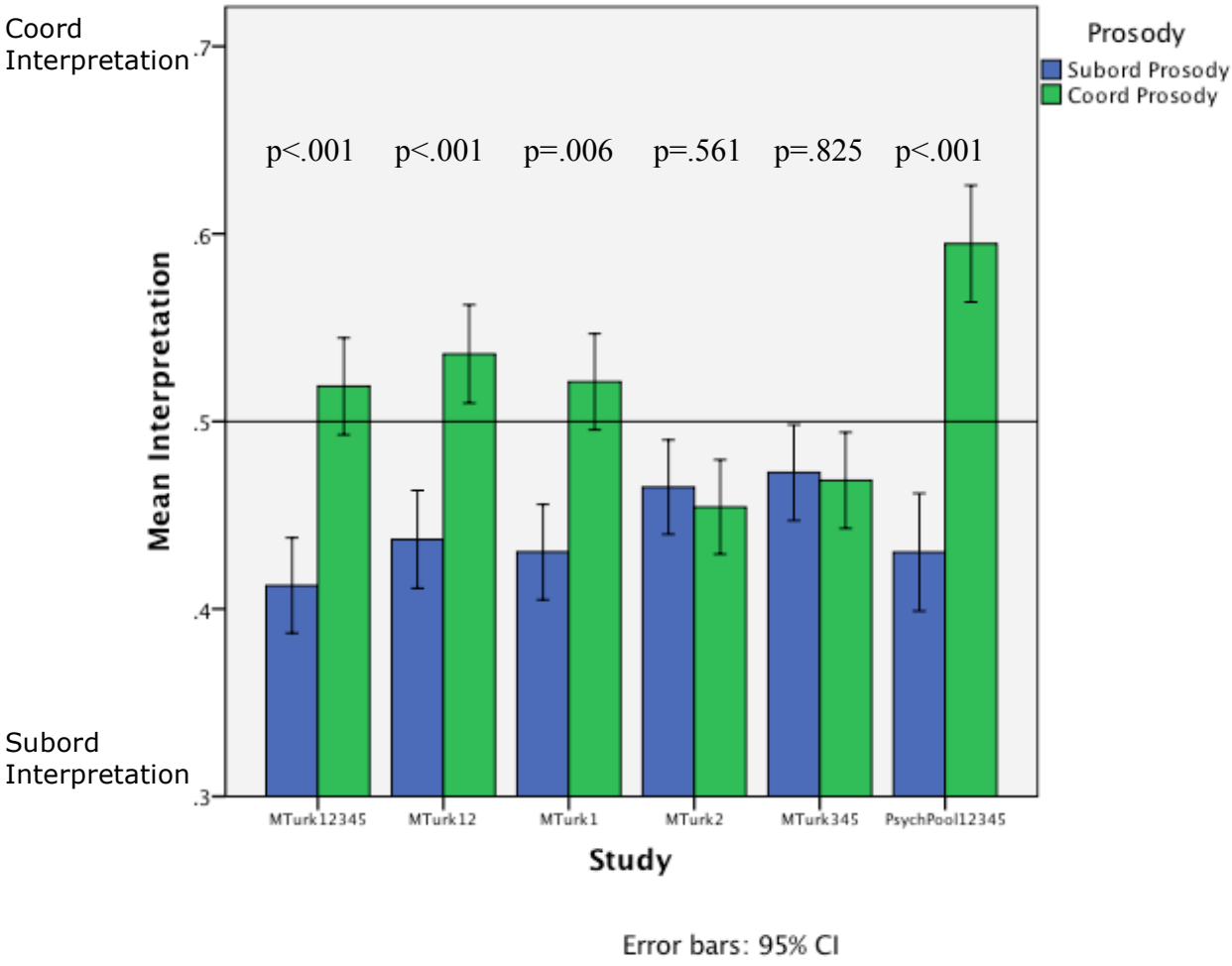
The other two studies showed no such effect: MTurk2 and MTurk345. Figure 4.1 plots the match rate (on the y-axis) for each study (on the x-axis), with .50 indicating the responses are at chance.

Figure 4.1: This graph plots each study on the x-axis and match rate on the y-axis (1=match, 0=mismatch), with 95% confidence intervals. Statistical results testing whether likelihood of match was different from likelihood of mismatch are overlaid on each study's column.



All studies that showed an effect of prosody on interpretation contained the S1-final pitch contour manipulation; the two studies that showed no effect did not contain this manipulation. This indicates it was this rise/fall contrast ending the first sentence that drove the interpretation effect. Figure 4.2 plots prosody against interpretation for each study.

Figure 4.2: This graph plots each study on the x-axis and mean interpretation on the y-axis (1=Coord, 0=Subord), with 95% confidence intervals. Statistical results testing whether likelihood of match was different from likelihood of mismatch are overlaid above each study's column. The right column for each study indicates results for Coord prosody, while the left column indicates results for Subord prosody.



Each cluster of two vertical bars corresponds to a study, with the right bar being Coord prosody and the left bar being Subord prosody. Mean interpretation is plotted on the y-axis, with more Coord interpretations making the bar higher and more

Subord interpretations making the bar lower. This graph shows the separation between the two prosodic conditions, with more judgments overall below the .50 chance line. This suggests somewhat of an overall bias towards Subord interpretations regardless of prosody, despite having run a norming study to identify the most ambiguous discourses (Appendix E).

While there was an overall effect showing match is more likely than mismatch in three Mechanical Turk studies, it is possible the Coord prosody and Subord prosody are contributing unequally to the overall effect. To test this, a predictor variable for Coord vs. Subord prosody was added into the model. If prosody is a significant predictor of match, then the Coord and Subord prosody conditions are contributing different amounts to the overall effect. Results, in Table 4.4, show that the studies MTurk12345, MTurk1, MTurk2 and MTurk345 showed a significant effect of prosody. All four of these studies had a negative coefficient for prosody, such that going from Subord prosody (coded as 0) to Coord prosody (coded as 1) indicated a significant reduction in match likelihood. Therefore, match rate is higher in the Subord prosody condition. However, this may be due to an overall preference for Subord interpretations.

Table 4.4: Results for prosody as a predictor of match, with a column for each study. This tests whether Coord or Subord prosody is more likely to result in match.

	MTurk12345	MTurk12	MTurk1	MTurk2	MTurk345	PsychPool12345
Prosody	Coeff: -.280 t=-3.718, p<.001	Coeff: -.111 t=-1.453, p=.146	Coeff: -.203 t=-2.685, p=.007	Coeff: -.323 t=-4.433, p<.000	Coeff: -.235 t=-3.176, p=.002	Coeff: .106 t=1.129, p=.259

One useful way of representing the sensitivity of participants to the informativity of the prosodic contrasts is with the measure d' ("d-prime") as discussed in Signal Detection Theory (SDT) (Heeger, 2003; Keating, 2004; Macmillan & Creelman, 2005; Swets, 1996; Tanner & Swets, 1954; Wickens, 2002; WISE, 2006). D' is a measure used to capture subjects' ability to discriminate between two conditions. If d' is zero, then there is no sensitivity to the difference between the two conditions. If d' is greater or less than zero, then participants are

sensitive to the differences between the two conditions. The primary purpose of using d' is to account for response bias. For example, if the analyses looked solely at Subord prosody to see if in that condition participants chose more Subord interpretations, then any effect could be due to an overall bias towards Subord interpretations and have nothing to do with the Subord prosody. Because these studies contrast interpretation (Coord vs. Subord) on both dimensions of prosody (Coord vs. Subord), this response bias is not a concern. Nevertheless, d' provides a simple way of graphically representing participants' ability to discriminate between the two prosodic conditions in each study, and it does so with a standardized scale. In this dissertation's studies, the two conditions are prosodic contrasts composed of different combinations of prosodic features. A positive d' value indicates that participants are more likely to match prosody and interpretation than mismatch, while a negative d' indicates mismatch is more likely.

Figure 4.3: This graph shows results for d' for each study, with results of a statistical test comparing likelihood of match vs. mismatch overlaid above each study.

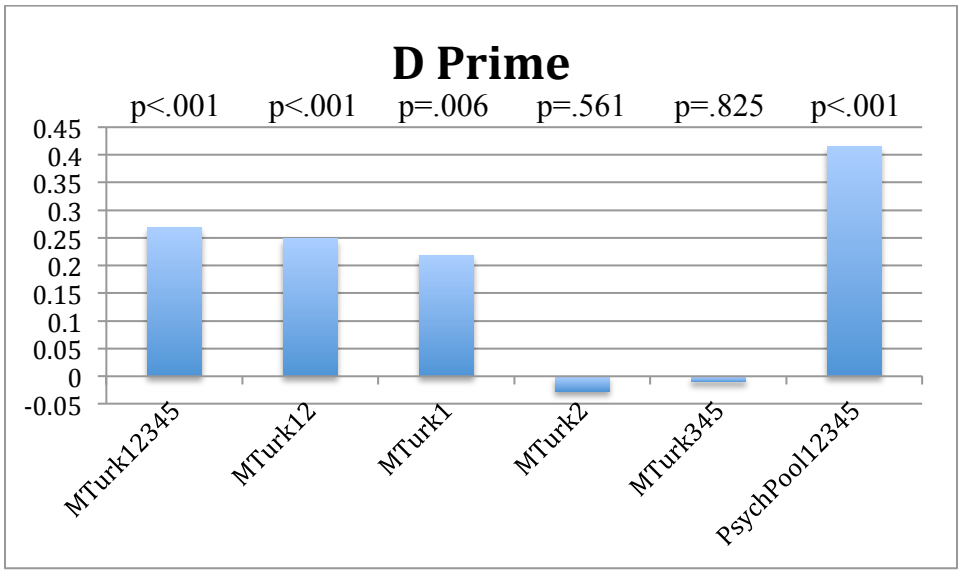


Figure 4.3 has the different studies on the x-axis with d' values on the y-axis. These d' values are raw values and do not measure a significant difference. The p-values from the statistical analyses have been overlaid to show which effects are significant. It is

clear that the studies with the S1-final pitch contrast are likely to result in more matches than mismatches, with a d' of between 0.2 and 0.45.

The bias of the interpretation elicitation question was also counterbalanced such that half of the questions participants saw would have a "yes" answer correspond to a Coord interpretation, and the other half would have a "yes" answer correspond to a Subord interpretation. Overall, there was a bias towards answering "yes," where across all studies 55% of answers were "yes" answers (t=-10.711, p<.001). But if these "yes" answers were equally distributed across both prosodic conditions, then this bias would not influence the effect of prosody on interpretation. When prosody is entered as a predictor of a yes response, the result is not significant (t=1.731, p=.084). Therefore, this bias towards yes responses does not differ across prosodic conditions. To test whether question bias affects match rate, a variable for question bias was included in the model.

Table 4.5: Results for question bias as a predictor of match, with a column for each study.

	MTurk12345	MTurk12	MTurk1	MTurk2	MTurk345	PsychPool12345
Question bias predicting match	Coeff: -.048 t=-.639, p=.523	Coeff: -.088 t=-1.147, p=.252	Coeff: -.106 t=-1.405, p=.160	Coeff: .016 t=.218, p=.827	Coeff: -.011 t=-.148, p=.882	Coeff: -.124 t=-1.317, p=.188

As is visible in Table 4.5, question bias does not predict match in any of the studies. This suggests prosody's effect on interpretation is not affected by question bias.

A separate concern was whether participants changed their behavior over the course of the experiment, perhaps improving with practice or deteriorating with fatigue. To test this question, a continuous variable was included in the model that coded whether a judgment was made in the first, second, third or fourth quarter of the experiment.

Table 4.6: Results of presentation quarter as a continuous predictor of match, with a column for each study.

	MTurk12345	MTurk12	MTurk1	MTurk2	MTurk345	PsychPool12345
Presentation quarters predicting match	t=-1.764, p=.078	t=-1.538, p=.124	t=1.39, p=.165	t=-.943, p=.346	t=-.165, p=.869	t=.379, p=.705

As seen in Table 4.6, presentation quarter was not predictive of match in any of the studies. This suggests that participant behavior with respect to prosody's effect on interpretation did not change over the course of the study.

It is also possible that the ability of prosody to disambiguate discourse depends on the practical ambiguity of the discourses themselves; the lexical material of one discourse could bias so much towards one interpretation that prosody would have no effect, while when multiple meanings are more equally accessible a factor like prosody could have an impact. As described in Appendix E, 102 discourses were normed for how available the two desired interpretations were, and the 48 most ambiguous discourses were selected. Within these 48 discourses, there was variation from one discourse to the next in terms of underlying preferences for one interpretation or the other. To test whether the underlying ambiguity of a discourse affected prosody's effect on interpretation, a covariate was included in the model that measured the absolute value of the difference between the number of people who chose Coord and Subord interpretations. This variable captures how equibiased the ambiguity is, regardless of whether the bias is towards Coord or Subord. The Generalized Linear Mixed model was run with this variable for underlying ambiguity as a covariate to see if the degree of a discourse's underlying ambiguity affects participants' ability to match discourse interpretation with the prosody. Results in Table 4.7 show an effect for studies MTurk12345 and MTurk12, but no others.

Table 4.7: Results of the underlying ambiguity of each discourse as a continuous predictor of match, with a column for each study.

	MTurk12345	MTurk12	MTurk1	MTurk2	MTurk345	PsychPool12345
Under-lying ambiguity predicting match	Coeff: -.010 t=-1.964, p=.050	Coeff: -.010 t=-2.039, p=.042	Coeff: .005 t=.903, p=.366	Coeff: -.006 t=-1.183, p=.237	Coeff: -.000 t=-.056, p=.956	Coeff: .001 t=.183, p=.855

If all three Mechanical Turk studies that showed an effect of prosody on interpretation are included at once, underlying ambiguity does not come out as significant (t=-1.737, p=.082). There is no reason *a priori* for why MTurk1 would show a different effect for underlying ambiguity than MTurk12345 and MTurk12. The fact that the effect for MTurk12345 and MTurk12 gets washed away when combined with MTurk1 suggests that those effects are weak, and that the degree of ambiguity had a small effect, if any at all, on the interpretation of this set of discourses.

There is also substantial variability in match rate from subject to subject. The graphs in Figure 4.4 plot match rate on the y-axis with subjects in ascending order of match rate on the x-axis, broken out by study. The horizontal line is at the .50 match rate, i.e. the chance level.

Figure 4.4: Results for each subject on the x-axis, tiled by study, with results of likelihood of match vs. mismatch overlaid on each study. The y-axis plots each subject's match rate, in ascending order. A horizontal line at 0.5 match rate is included for reference.

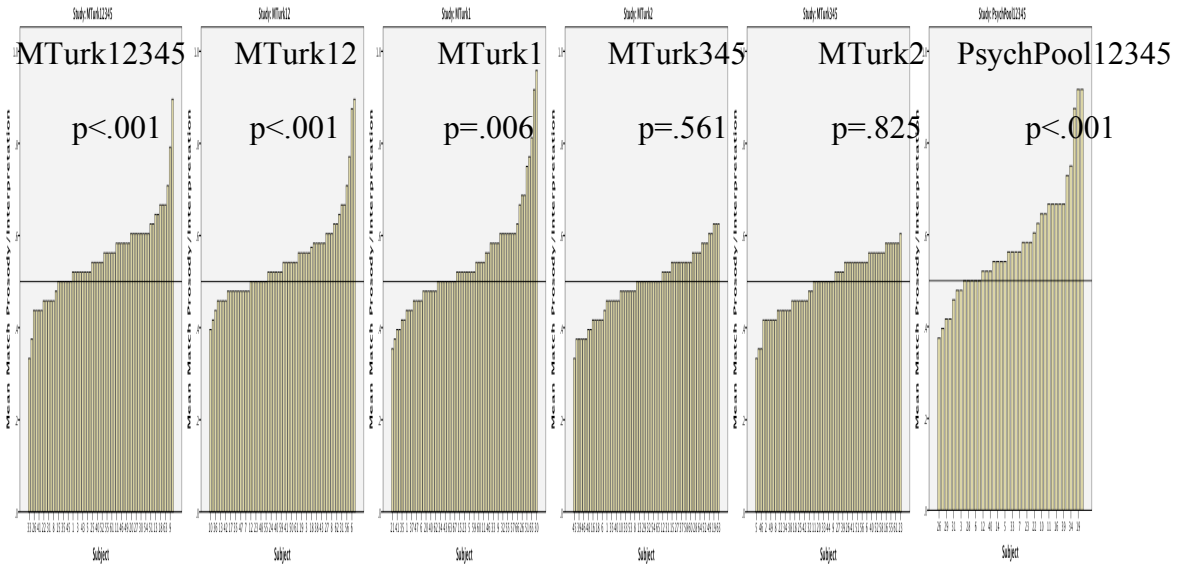


Table 4.8: Number of subjects with >80% match rate and >70% match rate, with a column for each study.

	MTurk 12345	MTurk 12	MTurk 1	MTurk 2	MTurk 345	PsychPool 12345
# subjects>80% match rate	1	2	3	0	0	3
# subjects>70% match rate	3	4	5	0	0	5

The significance values for match rate for the study as a whole are overlaid on each study. The major difference to note here is the few participants who performed dramatically better than the rest. These are visible on the far right of studies MTurk12345, MTurk12, MTurk1 and PsychPool12345, the four studies that showed a significant effect of prosody on interpretation. The overall effect may then be the result of only a few individuals. None of the subject demographic data collected predicted match rate, so what led these individuals to perform so much better is still unknown.

Confidence

In addition to providing their interpretation of each discourse, participants also provided their confidence in each judgment they made. Confidence judgments were

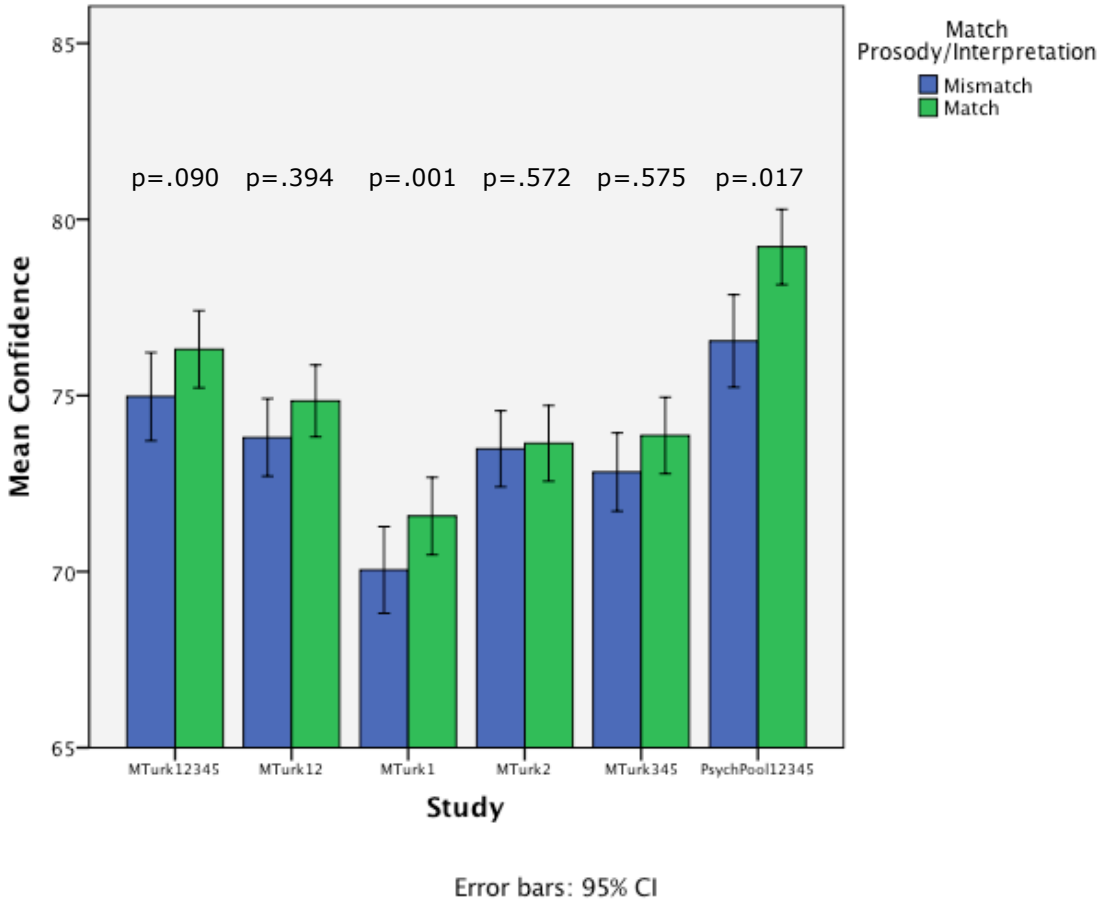
collected on a 1-100 scale and so constitute a continuous, not binary, outcome. As such, it was analyzed statistically using a Linear Mixed Model, not a Generalized Linear Mixed Model. This model contained a random subject effect and a random item effect. Table 4.9 shows the results of match as a predictor of confidence, i.e. whether participants were more confident in their judgments when their interpretation matched what was predicted from the prosody. Including subject as a grouping variable to account for correlated random effects, match was a significant predictor in two studies: MTurk1 and PsychPool12345. It is not surprising there was no effect of match on confidence for MTurk2 and MTurk345 as those studies had match likelihood at chance. If there is no indication they are using prosodic information in their interpretation, it is less likely their confidence would be affected by whether their interpretations matched the prosodic condition. By contrast, MTurk12345 and MTurk12 showed a higher than chance match rate, but no effect of match on confidence. While this is harder to account for, a combined data set including all studies that had greater than chance match rates (MTurk12345, MTurk12, MTurk1, PsychPool12345) showed a significant effect of match on confidence (F=16.831, p=000). This suggests match does predict confidence, but the effect may require enough data to have the statistical power to detect it.

Table 4.9: Results for match as a predictor of confidence, with a column for each study.

	MTurk12345	MTurk12	MTurk1	MTurk2	MTurk345	PsychPool12345
Match predicting confidence	F=3.114 p=.078	F=.814 p=.367	F=11.657 p=.001	F=.264 p=.608	F=.303 p=.582	F=5.738 p=.017

The graph in Figure 4.5 plots match on the x-axis and confidence on the y-axis, broken out by study. Within each study, the bar on the right is for where the interpretation matched the prosody, and the bar on the left is for when interpretation and prosody mismatched.

Figure 4.5: This graph plots confidence on the y-axis, with each study on the x-axis. Each study is split into a bar on the right for match results and a bar on the left for mismatch results. Error bars indicate 95% confidence intervals. Results for match as a predictor of confidence are overlaid above each study.



The graph shows confidence is higher in the match condition for all studies. Only in PsychPool12345 do the confidence intervals not overlap, though the difference is also significant in MTurk1. The graph in Figure 4.5 plots raw outcomes, ignoring any potential subject effects. The graph suggests that overall confidence may not pattern differently when interpretation matches prosody vs. when they mismatch. The statistical test includes a random subject effect, meaning that it can rule out some variability as being due to listener differences. It is perhaps for this reason that the statistical test, despite the overlap in 95% confidence intervals in Figure 4.5, is able to find a significant difference in confidence in the two match conditions for MTurk1.

Discussion

Like participants from the Psychology Subject Pool, Mechanical Turk participants also show an effect of prosody on interpretation, where they are more likely to match than mismatch. This effect was demonstrated in MTurk12345 with the same set of prosodic manipulations as PsychPool12345, thereby replicating the effect found using participants from the Psychology Subject Pool with Mechanical Turk participants. Three studies found an effect of prosody on interpretation (MTurk12345, MTurk12, MTurk1) while two did not (MTurk2, MTurk345). The rising vs. falling terminal pitch contour contrast was present in all studies that showed an effect of prosody on interpretation and in none that showed no effect. This distribution indicates the prosodic manipulation driving the overall effect was the pitch contour at the end of the first sentence of the discourses. The other manipulations had no independent impact on interpretation.

In the next chapter, I will examine the meaning of the S1-final rise that biased interpretation in more detail, but here I will comment briefly on the other manipulations. While manipulations 2, 3, 4 and 5 had no independent effect on interpretation, it is important to exercise caution in the interpretation of these null effects. The two pause duration contrasts did not affect interpretation, but this does not mean that no pause duration contrasts would. One possibility is that listeners can hear the pause duration contrast and are simply not assigning any meaning to it. Another possibility is that listeners cannot even hear the contrast. In this case, they may assign meanings to some pause duration contrasts in discourse interpretation, but not to contrasts they cannot hear.

The bias of the interpretation question and the quarter in which a discourse was presented were found not to affect the likelihood of match between prosody and interpretation. By contrast, the underlying ambiguity of a discourse, as revealed from the norming (see Appendix E), did predict match in MTurk12345 and MTurk12. The other Mechanical Turk study that had an effect of prosody on interpretation (MTurk1), however, did not show underlying ambiguity to predict match. When the data from all three studies are combined, the effect disappears, suggesting the effect is small, if present at all.

There is also variability from one participant to another, as seen most remarkably in the results of a few high performers in the studies where prosody biased interpretation. Unfortunately, none of the demographic data collected could explain why these individuals performed so much better, nor can I follow up with these individuals to find out what distinguished them. An explanation of this variability is therefore relegated to future research.

And finally, listeners were more confident in their decisions when they chose an interpretation that matched what was predicted by the prosody. At some level then, listeners are aware of their use of prosody inasmuch as they are more confident when using it as predicted.

I'll now compare the results from the studies in this and the previous chapter to the previous work on discourse prosody perception in Mayer et al. (2006) and Silverman (1987) (see chapter 1 for introductions to these two studies). While Mayer et al. (2006), Silverman (1987) and my studies all show the ability of prosody to bias the interpretation of ambiguous discourse, my studies are different from theirs in important ways. One difference lies in the kind of interpretation required in the experiments. Mayer et al. use a pronoun as a proxy for discourse interpretation, while two of the six discourses in Silverman (1987) rely on a universally quantified phrase with ambiguous domain restriction (e.g. "all materials"). In both cases, the judgment is about the meaning of a particular lexical item or phrase. In fact, Mayer et al. conclude by claiming the results of their Experiment 1 demonstrate prosody influences "the resolution of anaphoric pronouns" (p. 4) and Experiments 2 and 3 show listeners need both pause duration and pitch range parameters manipulated to be able to "disambiguate structurally ambiguous discourse" (p. 4). In the context of their experiments, Mayer et al. are treating pronominal anaphora resolution as equivalent and interchangeable with the disambiguation of structurally ambiguous discourse. This equivalence is especially problematic when, as discussed above, their data can be fully explained as a discourse recency effect and not one of discourse structure. Furthermore, there are many factors known to affect pronoun resolution; for examples of other biases on pronoun interpretation, see Kehler, Kertz, Rohde, & Elman (2008). Using pronouns as a proxy for discourse interpretation has the benefit of providing an

easily interpretable task for participants, but it raises the question of whether Mayer et al.'s results are strictly about prosody's effect on the disambiguation of discourse structure or also influenced by other factors that affect the interpretation of pronouns. A similar concern holds for the domain restriction of quantifier phrases like "all materials" used by Silverman, as quantifier domains are also affected by contextual factors (Fintel, 1994).

In my discourses, the participants are asked questions that reveal their interpretation of how the sentences fit together, i.e. did the events in sentences 2 and 3 happen during the event in sentence 1 or not. A benefit of this approach is that it more directly accesses the interpretation of the discourse and does not have to rely on the indirect means of the meaning of a pronoun or any lexical item. Discourse is structured out of whole discourse segments, and using questions that elicit interpretations that depend on whole segments may be a more direct means of identifying prosody's effect on discourse interpretation.

Another difference between my studies and those of Mayer et al. and Silverman is the kind of ambiguity we examine. While Mayer et al. and Silverman discuss the meaning contrast in terms of hierarchy, an alternate account that only draws on discourse recency is equally explanatory. They discuss their ambiguity as one of high vs. low attachment (Mayer et al.) or paragraph structure (Silverman), where there are groupings of discourse units and embedded within those units are more discourse units. For Mayer et al., the low attachment interpretations always involve attachment to the immediately preceding discourse segment, while high attachment interpretations always attach further back in the discourse. And for Silverman, low attachment involves restricting the domain of a quantifier phrase like "all materials" to either the immediately preceding material or more material further back in the discourse. In both cases, discourse recency can distinguish the meaning contrast equally as well as a hierarchical account. It is possible both Mayer et al. and Silverman may have only found that prosody can indicate how far back in a discourse to go to resolve the meaning of a phrase. Because both the hierarchical and recency accounts fully account for the contrasting meanings, either account is equally

explanatory. As a result, their results do not conclusively demonstrate prosody's effect on the interpretation of hierarchical discourse structure.

The discourses used for the studies in chapters 3 and 4 of this dissertation isolate the hierarchical contrast from a recency contrast, creating a minimal pair contrast between coordination and subordination. As such, they can be used to demonstrate an effect of prosody on the interpretation of the *hierarchical* structure of discourse. The contrast in meaning for those discourses is not in terms of *where* a new segment attaches but *how* it attaches, i.e. whether it is coordinated or subordinated. For example, in both interpretations of the discourse in (4.2), S2 (reading about housing prices) attaches to S1 (sitting in on a history class).

(4.2)

S1: I sat in on a history class.
S2: I read about housing prices.
S3: And I watched a cool documentary.

The ambiguity is whether the event of S2 was part of, i.e. elaborates, the event of S1 or is a separate, independent event. Unlike the ambiguities in Silverman and Mayer et al., the two interpretations cannot be explained with reference to near vs. far attachment because in both meanings S2 attaches to S1. In order to account for the meaning contrast, we need a hierarchical theory of discourse. And because prosody was able to bias the interpretation of discourses like (4.2), chapters 3 and 4 provide evidence that prosody can bias the interpretation not just of discourse generally, but specifically of the *hierarchical* structure of discourse.

This structural difference is relevant to note also because the interpretation effects in my studies and in Silverman and Mayer et al. are produced with different prosody. All of our studies manipulated pause duration and pitch, contrasting mainly in how pitch was manipulated. Mayer et al. manipulated the pitch range of whole sentences into normal, compressed, or expanded conditions, while Silverman cued discourse boundaries with final lowering before and initial raising after the boundary. And the purpose of their prosodic manipulations was to cue the size and location of a discourse boundary, with the boundary's size and location then biasing interpretation.

The effect of prosody in chapters 3 and 4 was a result of a rising vs. falling terminal pitch contour at the end of the first sentence. In both the Coord and Subord interpretations, sentences 1 and 2 are immediately related. Therefore, the rise/fall contrast cannot be indicating a near/far ambiguity. Instead, the rise/fall contrast must be indicating something about how the two sentences are related, e.g. coordination vs. subordination. These results suggest different kinds of prosody could have different kinds of effects on discourse interpretation. Some kinds of prosodic manipulations can cue larger boundaries and disambiguate near/far ambiguities. Other kinds of prosodic manipulations cue the kinds of relationships between sentences, disambiguating Coord/Subord ambiguities. In chapter 5, I proposed that one discourse meaning of terminal rising pitch is specifically discourse coordination, a proposal motivated by the results in chapters 3 and 4.

There are also important methodological differences between Silverman's studies on the one hand and Mayer et al.'s and my own on the other. First, Silverman's speech stimuli are all computer-generated text-to-speech synthetic speech. By contrast, Mayer et al. and I recorded human productions of individual sentences and then synthetically manipulated the prosody of those sentences. Participants in Silverman's study likely knew they were listening to a computer and not a human. It is unclear whether listeners draw on the same resources or use them in the same way in assessing computer-generated speech as compared to human speech.

Silverman himself acknowledges the existence of "problems associated with poor segmental quality in the synthetic speech" (p. 6.20). As discussed above, Silverman sought to address this issue by presenting participants with written transcripts of each discourse, with paragraphing removed. This means his participants were reading along while they listened to the discourses, while in Mayer et al.'s and my own studies listeners were simply listening with no accompanying written transcript. There are conflicting claims about whether listening-while-reading leads to better or worse comprehension. Some research argues that listening-while-reading may hinder comprehension relative to listening alone because it draws more cognitive capacity away from comprehension (Durkin, 1983). Other research argues listening-while-reading can enhance comprehension by providing both visual and auditory cues,

allowing listeners to draw on their own strengths from the variety of sources (Wong, 1986). Hale et al. (2005) test the relative impact of listening vs. listening-while-reading on elementary school students' reading comprehension, as compared to a baseline of reading alone. Their results, while inconsistent and using a small sample size, suggest listening-while-reading had a slightly better impact than listening alone. I am not aware of any research that has looked specifically at whether listening-while-reading has different consequences on the perception of prosody than listening alone. But we should be careful before assuming that results in listening-while-reading experiments would be the same as results in listening-alone experiments.

In addition, the nature of the elicitation of the discourse interpretation judgments was less natural in Silverman's studies. After listening to each discourse, Silverman would explain to each subject the nature of the ambiguity, detailing each alternative interpretation. Then he would ask them to listen to the speech again, focusing specifically on how it was "spoken" (his emphasis) (p. 6.21). By directing listeners' attention so overtly to how the discourse was spoken, listeners may be reacting to the prosody in a different way than in a more natural context, potentially assigning meanings they otherwise do not have. Furthermore, for one discourse's elicitation question, Silverman asks "From the way it is spoken, which of these two does the computer intend step 3 to be?" (his emphasis) (p. App III.9). It is strange to ask listeners to infer the intentions of a computer, which seems to imply that computers have intentions. What a listener might do in response to such a request could be to try to infer what the person who programmed the computer intended. By emphasizing the computer as the creator of the discourse, it highlights that the speech does not originate with a human. These concerns about the naturalness of the task are not intended to say Silverman's findings are not significant as a demonstration that listeners can use discourse prosody in discourse interpretation, but they are important to recognize in assessing the generalizability of the findings. Both Mayer et al. and this dissertation therefore make an important contribution by showing an effect of prosody on discourse interpretation in a more natural context.

In addition, the results of both Silverman and Mayer et al. suggest a cumulative effect of prosodic cues to discourse structure. Mayer et al. found an effect

when both pause duration and pitch were manipulated but no effect when only one prosodic measure was manipulated. Silverman found a stronger effect on interpretation when both pitch and pause duration were manipulated than pitch alone. Unlike both of these studies, I found limited evidence for a cumulative effect of multiple prosodic manipulations. The only prosodic manipulation that mattered was the terminal pitch contour on sentence 1. There may have been some slight improvement when more manipulations were present, as seen in the raw scores, but the differences were not significant. This may be reducible to the specific prosodic contrasts tested; other manipulations could potentially have helped influence interpretation.

Finally, Mayer et al. conducted their experiments in German with native German speakers while Silverman's participants were all native speakers of British English. Given that we know so little about discourse prosody perception at all, much less across languages or dialects, it seems important to keep in mind that speakers and listeners of German, British English and American English may behave differently.

Chapter 5

Rising Intonation as a Marker of Discourse Coordination

The main finding from the experiments in chapters 3 and 4 is that listeners interpret a sentence-final pitch rise to have a special discourse meaning. In this chapter, I will analyze the meaning of that rise, claiming that a rise indicates discourse coordination. This claim not only fits my results, but it also fits with existing claims about sentence-internal listing intonation. Then, I will analyze the potentially contradictory claim in Pierrehumbert & Hirschberg (1990) that sentence-final rises indicate elaboration, a subordinating relation. I will reanalyze the data provided in Pierrehumbert & Hirschberg (1990), bringing their data in line with my findings. Finally, I will discuss implications of the claim that pitch rises indicate discourse coordination, saving for the next chapter a discussion of how this dissertation relates to other work on prosodic disambiguation.

Before embarking on this discussion, however, I will motivate the fact that the pitch rise is the locus of the observed discourse meaning. First, most work on intonational meaning treats the rise as what needs to be explained, with the fall as the more neutral, unmarked or default case (Gunlogson, 2003; Jayez & Dargnat, 2008; Marandin, 2007; Pierrehumbert & Hirschberg, 1990; Reese, 2007). For example, Pierrehumbert & Hirschberg (1990) provide an explanation for the meaning of high boundary tones, while arguing the meaning of low boundary tones is "less clearly marked" (1990, p. 287).

My results in chapters 3 and 4 also provide empirical evidence that the rise is what is having the interpretation effect, not the fall. Those studies involved presenting a spoken discourse three sentences long that could be interpreted in one of two ways.

The Coord interpretation involved seeing each sentence as describing a separate, independent event, while the Subord interpretation involved seeing the first sentence as describing an event that the second and third sentences elaborated. The prosody of those spoken discourses was manipulated into two conditions for each experiment, one that was predicted to bias towards more Coord interpretations and the other that was predicted to bias towards more Subord interpretations. The nature of the prosodic contrasts varied in different experiments, helping to isolate the cause of any effect on interpretation. The study that used subjects from the Psychology Subject Pool, described in chapter 3, contrasted on all five prosodic manipulations (PsychPool12345). The rest of the studies used participants from Amazon's online labor marketplace Mechanical Turk, contrasting different sets of prosodic manipulations. Each Mechanical Turk study is named after the set of contrasting prosodic manipulations:

(5.1) The five prosodic manipulations in the perception studies of chapters 3 and 4

1. Terminal pitch on S1
2. Terminal pitch on S2
3. Pause duration between S1 and S2
4. Pause duration between S2 and S3
5. Mean pitch and intensity on S2 and S3

As a shorthand, the numbers 1 through 5 will be used to refer to these manipulations such that, for example, MTurk12 would contrast on S1 and S2 terminal pitch.

In the studies described in chapters 3 and 4, some contained sets of prosodic manipulations that resulted in an effect on interpretation (MTurk12345, MTurk12, MTurk1, PsychPool12345) and others did not (MTurk2, MTurk345). Of those studies where prosody had no effect, participants chose the Subord interpretation about 55% of the time. This suggests a default response rate of about 55% Subord interpretations. For the studies where there was an effect of the prosodic manipulation on interpretation, the Subord prosody manipulation still resulted in about 55% Subord interpretations, but it was the Coord prosody that was different, resulting in about 55% Coord interpretations. The single feature common across all studies that showed

an effect on interpretation was a contrast in S1-final pitch contour, i.e. a rise vs. a fall. Therefore, in the absence of cues, participants default to 55% Subord interpretation; but when a rise is present, participants change more towards the Coord interpretation. The addition of the rise changes interpretation patterns, suggesting the rise is more of a locus of meaning than the fall.

Finally, there are physiological reasons to believe the rise is carrying the meaning and not the fall. Over the course of a unit of speech production (e.g. an intonational phrase), there is usually pitch declination where pitch drops from beginning to end (e.g. Liberman & Pierrehumbert, 1984). This general pattern can be modulated with extra effort to create pitch accents and terminal rises, but these breaks in the overall declination pattern are marked. Because a sentence-final rise is marked, it may trigger a search for an explanation for why the speaker went through the extra effort required to make a rise. Tomlinson Jr (p.c.) suggests listeners may expect a certain kind of prosody for a default interpretation, e.g. the Subord interpretation, and an "unexpected" prosody could lead listeners to make a different interpretation. A similar conclusion is reached by Wales & Toner (1979), who argue that there is no direct mapping between prosody and syntax, but that prosody may signal the listener to look for an unexpected meaning. The ability of a rise to break the overall declination pattern may make it "unexpected," and result in a search for an alternate meaning. This effort-based explanation also suggests the meaning of the interpretation contrast lies in the rise, juxtaposed with the fall, but not in the fall itself.

Listing intonation as motivation for rising pitch indicating discourse coordination

The intonation of lists provides one motivation for the claim that a terminal rise indicates discourse coordination. In essence, the Coord interpretation of my ambiguous discourses is a list of three independent events. So, if you hear the example in (5.2) and you interpret it as meaning three separate, independent events (the Coord interpretation), you are essentially interpreting the speaker to be giving a list of three things they did.

(5.2) I sat in on a history class. I read about housing prices. And I watched a cool documentary

Work on the intonation of lists, however, has not generally looked at a list of whole sentences, focusing instead of lists of either noun phrases (NPs) or verb phrases (VPs). So, for example, Beckman & Pierrehumbert use the sentence fragment in (5.3) and Ladd gives the example in (5.4) where each of the listed elements is an NP. And Cauldwell & Hewings provide an example of listed VPs in (5.5).

(5.3) Blueberries, bayberries, raspberries, mulberries and boysenberries (Beckman & Pierrehumbert, 1986, p. 273)
(5.4) I need milk and eggs and butter and bread (Ladd, 1980, p. 183)
(5.5) John has got to buy some coffee, wash the floor, and wind the clock (Cauldwell & Hewings, 1996, p. 330).

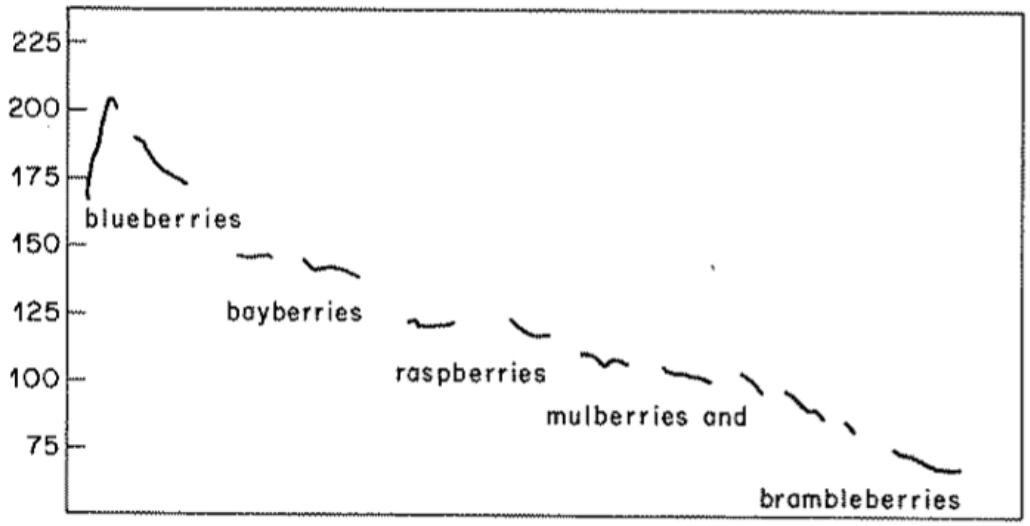
The motivation for looking at lists and listing intonation here, however, is that lists provide a more transparent case of coordination between elements. The elements in a list are all at a similar level of detail and share some similar relationship to the superordinate category defining the list. So, the elements in a list of berries (e.g. blueberries, bayberries, raspberries, mulberries) are all equally members of the set of berries. There is no direct hierarchical relationship between the berries themselves, and in this sense, they are coordinated to each other.

There is a reasonably sized literature on the intonation of lists, and while there is little experimental or corpus work on the actual production of lists, there seems to be general agreement on what listing intonation is, at least in its canonical form. This canonical listing intonation is characterized as a series of rises concluded with a fall (Cauldwell & Hewings, 1996; Hirschberg, 2008; Ladd, 1980; Schubiger, 1958). Ladd gives the example "I need milk and eggs and butter and bread," with a rise on milk, eggs, butter and a fall on bread, as an example of a sentence that would be produced with canonical listing intonation (1980, pp. 183-184).

This canonical form of listing intonation is reiterated by Beckman & Pierrehumbert (1986) as motivation for the feature *downstep* in their intonational

phonology. They provide the example in Figure 5.1, drawn from Liberman & Pierrehumbert (1984), where each item in the list receives a pitch accent.

Figure 5.1: An f0 contour for the production of a list of berries, showing downstep on each member of the list (Liberman & Pierrehumbert, 1984, p. 171).



The pitch peak reached becomes lower and lower for each subsequent item in the list. While in this case the listing intonation is phrased as a series of tonal targets of descending height, it still sounds like rising pitch on each list item. A similar structure has been replicated by Grabe (1998) with British English.

Much of the early work on listing intonation was based on author intuitions and a few representative, and perhaps constructed, examples. Cauldwell and Hewings (1996) take issue with standard "textbook rules on intonation" (1996, p. 333), arguing instead for a more comprehensive model of intonation. Specifically they are reacting to textbook claims that the intonation of lists "always goes down on the last item (to show that the list is finished), and up on all the items that come before the last (to show that there is more to come)" (Bowler & Parminster, 1992, p. 30). Instead of this simple, deterministic rule for listing intonation, they want to understand the greater complexity of actual listing intonation in production as it varies by context and as different meanings interrelate. In pursuit of this goal, they perform an empirical study of lists in radio programs, finding there is no one-to-one mapping between location in

a list and terminal pitch contour. While they demonstrate that not all lists in production fit the canonical listing intonation, what they call the “textbook rules on intonation,” there is a range of work that seeks to capture the various meanings that different kinds of intonational productions of lists convey.

While Cauldwell and Hewings (1996) demonstrate that not all lists in production fit the canonical listing intonation, scholars have explored the nature and meaning of non-canonical listing intonation. For example, Ladd (1980) contrasts two kinds of listing intonation, one marked and one unmarked. The plain form is claimed to be for contexts where the speaker enumerates all relevant members of the list, while the marked form is when a few associated elements are mentioned and the listener is expected to fill out the rest. The plain form is claimed to be produced with “plain high-rises” on all pre-final elements, while this marked form is produced with "stylized high-rise-- with the rise becoming steady" (1980, pp. 183-184). Ladd provides the following example:

(5.6) An example from Ladd (1980) of stylized high-rise pitch as used in the production of incomplete lists.

A: Hey, these cookies are good. What’s in ‘em?
B: Oh, nothing special, you know--
Flour-- sugar-- butter--
and and and, uh… (1980, p. 183)

The claim is that flour, sugar and butter are only some of the members of the list of what is in the cookies. So while a fully enumerated set would be produced with a plain high-rise at the end of each pre-final element, this partial set instead is produced with a "stylized high-rise" (i.e. a rise that plateaus).

Hirschberg (2008) reiterates this contrast between closed and open-ended lists. Within the autosegmental-metrical framework of the ToBI transcription scheme (Silverman et al., 1992), she claims that H* L-H% (continuative rise) contours are for (presumably non-final) members of a closed set. By contrast, H* H-L% (plateau) contours are for an “open-ended set” (p. 533):

(5.7) The Johnsons are solid citizens.
They **H*** pay their **H*** taxes **H-L%**.
They **H*** attend **H*** PTA meetings **H-L%**.
They're just good people. (p. 533)

Schübiger (1958) discusses another kind of contrast, between lists that are fully planned ahead of time and those that a speaker is figuring out as they go. Schübiger discusses the canonical "rise-rise…-fall" pattern along with four others, and provides interpretations of the kinds of meanings each set of contours conveys. Her contour sets are all for a list containing four elements and are cashed out in terms of element-final rises or falls. The classic rrrf (r=rise, f=fall) contour shape is for lists that are fully planned out ahead of time and are completed sets, i.e. there is no more to come (5.8).

(5.8) (rrrf) ↗↗↗↘
We saw a good deal during those two weeks. We went to Venice↗, Florence↗, Rome↗ and Naples↘.

The alternative ffrf (5.9) is equivalent except the inclusion of an *and* preceding the final element "seems to be the rule" (p. 72).

(5.9) (ffrf) ↘↘↗↘
Which writers do you have to study for your examination? Quite a number of them: Chaucer↘, Shakespeare↘, Milton↘, Pope↗, and Swift↘.

An fffF (F=major fall) set is for lists where the speaker is figuring out as they go what is part of the list, i.e. it is not fully planned ahead of time but still comes to completion (5.10).

(5.10) (fffF) ↘↘↘↘

My husband is very fond of outdoor games. He plays tennis↘, and golf↘, and cricket↘, and polo↘.

By contrast, the contour sets rrrr and ffff are claimed to convey a sense of "more items might follow," i.e. incompleteness (5.11) and (5.12).

(5.11) (rrrr) ↗↗↗↗

You could easily become an interpreter. You know French↗, and German↗, and Spanish↗, and Russian↗.

(5.12) (ffff) ↘↘↘↘

You should see the lovely fruit they grow: apples↘, and pears↘, and peaches↘, and apricots↘, and grapes↘.

This review of work on listing intonation demonstrates a range of kinds of listing intonation. But even with the variability in the actual production of lists, we still conceive of listing intonation as having the canonical form of a series of rises followed by a fall. The variability in actual production may not prevent listeners from drawing on canonical listing intonation in ambiguous contexts to facilitate choosing one interpretation over another.

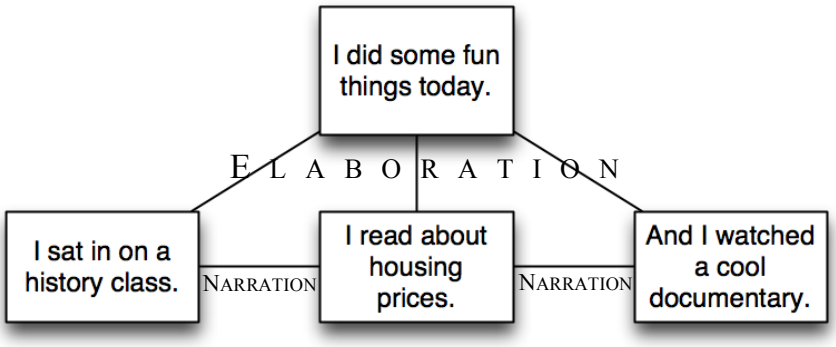
While the previous literature has generally looked at lists of NPs or VPs, lists could also be made up of whole sentences. The canonical form of a list made up of sentences would still entail terminal rises on non-final members of the list with a fall on the final sentence. For example, when the discourse in (5.2) is interpreted as a list of separate activities, it would be conveyed with the canonical set of pre-final rises and concluded with a fall. But non-canonical forms can also scale up to lists of whole sentences. For example, Ladd's stylized rises can still convey a meaning of partial enumeration when applied to whole sentences:

(5.13) A: What did you do today?
B: I went to the stoooore. I stopped by the gyyymmmm. And uh…

In (5.13), speaker B is listing activities they performed that day. But it is only a partial list, and the high-plateau stylized rises convey that partialness. This non-canonical listing intonation contour then functions similarly both within a sentence, as Ladd (1980) describes, as well as between sentences.

In a list where each member of the list is a whole sentence, it is then a small leap to say that the sentences themselves are coordinated to each other. The nature of how we represent that coordination may be different, drawing on discourse structural representations instead of syntactic ones, but in essence the hierarchical nature of the relationship is still one of coordination. Returning to my example in (5.2), the common topic for these three independent activities could be something like "fun things I did today." Perhaps, this would even be made overt by a preceding sentence like "I did some fun things today". The discourse structure would then look something like in Figure 5.2.

Figure 5.2: A graphic representation with an overt topic dominating the discourse in (5.2)



If the elements of the list are full discourse segments (e.g. sentences), as presumed in the above representation, then those discourse segments would be related to each other via discourse relations. More specifically, they would be related by coordinating discourse relations, and all three sentences would together be subordinated to the superordinate topic of "fun things I did today." Within the

terminology of Segmented Discourse Representation Theory (SDRT), they would be coordinated via Narration relations, indicating that each activity happened subsequent to the previous. And all three would together be related to the superordinate topic via an Elaboration relation, indicating they each independently provide more detail about the topic of "fun things I did today." Therefore, it is reasonable to consider the Coord interpretation of my ambiguous discourses as having a list structure, and the sentences of those discourses as being coordinated to one another. It is also reasonable to expect that listeners might draw on listing intonation in their interpretation of what the speaker meant upon uttering the discourse.

Pierrehumbert & Hirschberg (1990): Reconciling a potential contradiction

Pierrehumbert & Hirschberg (1990) provide a potentially contradictory account for the meaning of high terminal pitch. Put simply, they claim high terminal pitch indicates a relation between sentences like elaboration, a subordinating relation in SDRT, while my findings showed high terminal pitch biased away from elaboration, towards coordination. Because of this difference, and because their examples and claims are in similar terms and comparable to the data in my experiments, I will examine their claims with respect to my results in more detail. In my analysis, I will not claim the difference lies in their judgments of felicity or the nature of the prosody, but instead in the proposed structure for their discourse. My reanalysis will bring their data in line with my findings.

I will introduce the approach and claims of Pierrehumbert & Hirschberg (1990) here, but it will quickly become clear that their example is analyzed in terms of two theoretical frameworks that will need introduction. The first is their approach to prosody, the auto-segmental theory of intonational phonology that uses the ToBI annotation system (Silverman, et al., 1992). The second is the Grosz & Sidner model (Grosz & Sidner, 1986), a theory of discourse structure drawn from the artificial intelligence community. To facilitate the discussion, I will therefore also need to introduce these two theories.

The primary goal of Pierrehumbert & Hirschberg (1990) is to develop a compositional approach to intonational meaning. The units of meaning they address

compositionally are from the intonational phonology of Pierrehumbert (1980) represented using the ToBI annotation scheme. This intonational phonology involves proposed abstract phonological categories for two kinds of intonational phenomena: pitch accents and boundary tones (an example of the notation of these categories is described below). These pitch accents and boundary tones are claimed to underlie how speakers and listeners make sense of the intonational structure of speech. What Pierrehumbert & Hirschberg (1990) mean by a compositional approach to meaning is that each distinct pitch accent and boundary tone has a unique, context-independent meaning. These meanings then combine with each other to make more complex meanings. The motivation for such an approach, in contrast to others that assign meanings for whole contours, is that it captures what they see as generalizable, consistent meanings of these discrete features across contexts. Pitch accents, for example, are claimed to "render salient" (p. 288) the information accented, and the status of that information depends on the kind of accent (p. 289). And boundary tones indicate something about how one intonational phrase, the largest unit of intonational structure, relates to another. They write: "boundary tones convey information about relationships among intonational phrases- in particular, about whether the current phrase is to be interpreted with particular respect to a succeeding phrase or not" (1990, p. 287).

In their system, there are two kinds of boundary tones, one high and one low, and each is claimed to convey a different kind of meaning. They say that high boundary tones introduce hierarchically lower discourse segments, e.g. through relations like elaboration. This claim is motivated by the following example:

(5.14)

a. The train leaves at seven
   H* H* H* L H%

b. It'll be on track four
   H* H* L L%

The bolded letters below each sentence are the ToBI transcriptions of the sentences' intonation. In the ToBI annotation, an H* indicates a high pitch accent, an

L indicates a low intermediate phrase accent, an L% and an H% indicate low and high intonational phrase boundary tones respectively. This example shows two sentences produced with two intonational phrases, the first ending with a high boundary tone H% and the second ending with a low boundary tone L%. Pierrehumbert & Hirschberg claim that the H% ending (5.14a) triggers the expectation of an inferable relationship between (5.14a) and (5.14b). By contrast, if (5.14a) ended with an L% the relationship between the two would be "less clearly marked" (1990, p. 287).

In addition, Pierrehumbert & Hirschberg go on to speculate as to the nature of what kind of relationship the H% conveys, proposing that the nature of that relationship could be cashed out in terms of the hierarchical relationships of the Grosz & Sidner model (1986). Grosz & Sidner (1986) propose a theory of discourse organization that comes from the artificial intelligence research community and is deeply pragmatic in nature. Grosz & Sidner argue that conversation is organized as a hierarchy of discourse intentions, where at any one point there may be a high-level goal for the participants that is approached through one or more sub-goals. In this sense, it is the intentions of the speaker that are the organizing units of the discourse and the propositional content of the language uttered is useful only insomuch as it signifies what the speaker's intentions are. Interpreting speaker purposes/intentions/goals is deeply pragmatic because it is not tied specifically to the literal semantic content of the message, but is the result of an inferential process about why the speaker said that message in that context.

In drawing on this model, Pierrehumbert & Hirschberg interpret (5.14) in terms of these discourse purposes, arguing that "the satisfaction of the purpose S [the speaker] has in uttering (5.14b) contributes to the satisfaction of S's purpose in uttering (5.14a) by further elaboration" (1990, p. 287). They posit (5.14a) is in a dominance relationship with (5.14b), because the goal of (5.14b) serves to partially fulfill the goal of (5.14a). They then present the example (5.15) where it is difficult to infer any relationship between the two sentences. The H% still creates a desire to interpret (5.15a) with respect to (5.15b), but because it is difficult to find any reasonable relationship between the two sentences, the discourse ends up sounding odd.

(5.15)

a. The train leaves at seven
H* H* H* L H%

b. There's a full moon tonight
H* H* H* L L%

An L% ending (5.15a), however, would trigger less of an impulse to infer a relationship between (5.15a) and (5.15b) and as a result the discourse would seem less degraded.

I agree with the felicity judgments of Pierrehumbert & Hirschberg, i.e. that (5.14) sounds better with an H% ending (5.14a) rather than an L%, and that (5.15) sounds odd with an H% ending (5.15a) while an L% is less odd. The disagreement emerges in terms of how they interpret the structure of the discourse. As stated above, Pierrehumbert & Hirschberg (1990) couch their more specific claim in terms of the Grosz & Sidner (1986) model of discourse structure, meaning the relationship between discourse segments is structured in terms of the relationship between the speaker's intentions underlying those utterances. One problem with this approach involves how one should determine the speaker's intentions that serve as the basis for identifying the discourse's structure. As a result, claims about speaker purposes, especially with isolated two-sentence discourses presented without larger context, leaves the structural representation detached from the actual linguistic production. As a result, the inferred speaker's purposes can be imagined with so much flexibility that it is hard to tie down claims and directly compare the Pierrehumbert & Hirschberg example to my results.

Instead of relying on speaker purposes, there are theories of discourse that model the coherence structure of discourse based more directly on the linguistic material available (Asher & Lascarides, 2003; Kehler, 2002; Mann & Thompson, 1988). These theories have articulated definitions of a range of coherence relations that can hold between sentences, including "elaboration," the relation Pierrehumbert & Hirschberg claim holds between (5.14b) and (5.14a). And theories that draw on coherence relations between sentences have motivated ways of accounting for the

structure of discourse by focusing specifically on the propositional content of the available sentences, along with inferences about how they are related. Because these coherence theories are more tied to the propositional content, and because they have a more explicit account for the meaning of elaboration, I will continue to discuss the example (5.14) in terms of a coherence relation between (5.14a) and (5.14b), one that Pierrehumbert & Hirschberg propose to be one of elaboration. Then their claims can be more directly compared to my results, and the implications of their claims can be more precisely assessed.

Two theories that model the coherence structure of discourse, Rhetorical Structure Theory (RST) (Mann & Thompson, 1988) and Segmented Discourse Representation Theory (SDRT) (Asher & Lascarides, 2003), offer technical definitions for elaboration relations. RST has many different, specialized elaboration relations while SDRT subsumes them all under one single elaboration relation. As our discussion hinges on elaboration more generally, the SDRT definition should more directly suit our needs. One clear definition of an SDRT elaboration relation comes from an annotation manual (Reese, et al., 2007) created for a project called DISCOR (Baldridge, Asher, & Hunter, 2007) that involved developing an SDRT-annotated corpus of newspaper articles. In this manual, the authors write "Elaboration(α, β) holds when β provides further information about the eventuality introduced in α; for example, if the main eventuality of β is a sub-type or part of the eventuality mentioned in α" (Reese, et al., 2007, p. 7). A simple definition of elaboration then is that an elaborating discourse segment is one that provides more information about the eventuality in the discourse segment it elaborates.

If we apply this definition to the Pierrehumbert & Hirschberg example, their claim that (5.14b) elaborates (5.14a) would mean that the train being on track four provides further information about the train leaving at seven. This seems difficult to reconcile with our intuitions about what it means for a train to leave. Sentence (5.14a) ("The train leaves at seven") indicates "time of departure," while sentence (5.14b) ("It'll be on track four") indicates "location of departure". It is unclear how *location of departure* could provide further information about *time of departure*. Rather, both of these provide further information about the departure itself.

In the Pierrehumbert & Hirschberg analysis, discourse (5.14) contains two and only two discourse segments, each of which has a direct correlate in the overt lexical material. I argue instead that the sentences of discourse (5.14) are related by together elaborating some superordinate, but implicit, topic. In this alternative analysis, the *location*(5.14b) of departure is no longer providing more detail about the *time*(5.14a) of departure. Instead, the *location*(5.14b) and *time*(5.14a) together elaborate the implicit topic of "the departure." And because they each provide a similar level of detail about that topic, they would be related to each other via a coordinating relation.

The viability and naturalness of this alternative discourse structure, where one segment describes the time and the other the location of some higher-level topic, may be more visible in a separate example:

(5.16) The concert will start at 9pm at Michigan Theatre.

In this sentence, there are two prepositional phrases that provide further information about the concert, first to specify the time ("at 9pm") and second to specify the location ("at Michigan Theatre"). This sentence could be broken into two sentences and still convey more or less equivalent information:

(5.17) The concert will start at 9pm. It will take place at Michigan Theatre.

The same information could also be conveyed with three sentences:

(5.18) There is a concert. It will start at 9pm. It will take place at Michigan Theatre.

In (5.18), the first sentence most transparently introduces the concert as a topic. Then the second and third sentences provide more information about that concert, namely time and location information. Similarly, example (5.14) can be converted into an informationally equivalent three-sentence discourse:

(5.19) There is a train (meeting your requirements). It leaves at seven. It will be on track four.

This simple re-organization makes the relationship between sentences correspond more closely to the underlying discourse's structure, where leaving at seven and leaving from track four are elaborating the larger topic of the train.

There is another to reason to think an implicit topic analysis is better and this is because of the NP "the train" at the beginning of (5.14a). The speaker in uttering the NP "the train" is presupposing the existence of some unique train. The presence of the existential presupposition in "the train" is visible through a negation test (Levinson, 1983, pp. 177-178).

(5.20)

a. The train leaves at seven.
b. The train does not leave at seven.

In either the affirmative or the negated condition, there is still a presupposition that there is some (salient) train in the world being discussed. It is not required that the train be mentioned previously, as listeners can be expected to reconstruct, or "accommodate" (Lewis, 1979), this presupposition. But it necessitates a context that could facilitate such accommodation. Additionally, the definite description "the train" implies that there is some unique train in the context that is recoverable. Out of the blue, uttering "the train" would be awkward and unclear. But if there is a pre-established topic of somebody taking the train to go somewhere, then it is possible to pick out which train "the train" is referring to, namely the train that said person would be taking. As a result, it is reasonable to assume that (5.14a) does not exist in a vacuum and instead that a superordinate topic is implicitly present.

While most of the discussion thus far has hinged on re-interpreting just the two sentences provided by Pierrehumbert & Hirschberg, we can also add overt contextualizing material and see how that affects the nature of the relationship between (5.14a) and (5.14b). In his discussion of discourse topic, Asher (2004) mentions that topics can be explicit in the discourse as well as implicit. He goes on to suggest, in a discussion of the information sources that can constrain the construction of implicit topics, that "explicit topics may be of use in checking how these information sources really affect discourse topic" (2004, p. 181). In this spirit, one

way to test for the presence of an implicit topic in the discourse is by constructing an explicit topic; if the resultant discourse makes sense, it suggests that the implicit topic analysis is on track. As an example, imagine a person trying to sell a ticket outside of a train station. You are walking by, having no prior experience with this person or even being aware you're in front of the train station. Then you hear:

(5.21) (Walking outside a train station.) Ticket scalper:
a. I have one ticket to London!
b. The train leaves at seven.
c. It'll be on track four.

It seems clear the prior context is not supplying the topic because you were not paying attention to this speaker or your location next to the train station. Instead, the topic of a train voyage to London is introduced by the ticket scalper. As you hear the first sentence, you create a topic of something like "train voyage to London" in your representation of this person's discourse. Then, when you hear "The train leaves at seven. It'll be on track four," you have learned two additional pieces of information about that voyage, namely the time and the location of the train's departure. In this case, sentences b and c of (5.21) are providing further information about, i.e. elaborating, the topic introduced in sentence a of (5.21), and they are doing it at a similar level of detail. Therefore, the first sentence of (5.21), "I have one ticket to London," makes explicit a topic that is likely implicit in the original example (5.14). Because the implicit topic can be made felicitously explicit, the proposal of an implicit topic is a reasonable one.

The implicit topic could also be made more explicit through non-linguistic means, by creating a salient topic in the broader context of the uttering of (5.14). For example, imagine you are in a train station when you see a sign mentioning a train to London. Then someone announces over a loudspeaker:

(5.22) (In train station, having just seen a sign where a train trip to London flashed)
The train leaves at seven. It'll be on track four.

In this context, the flashing of the sign created a topic, something about a train trip to London, and the overt speech of (5.14) simply elaborates that topic.

While adding preceding context can reveal the nature of the relationship between (5.14a) and (5.14b), it can also help to add a sentence afterwards. If there is some topic and both (5.14a) and (5.14b) are elaborating that topic, then we would expect a third sentence to be able to also elaborate that topic at a similar level of detail. To exemplify this strategy, consider the following:

(5.23) The train leaves at seven. It'll be on track four. It will be chaotic at the station.

The implicit topic remains the train's departure. Then, each sentence elaborates that topic, first telling *when* the train leaves, then *where* it leaves from, and finally *how* the departure will play out. The when, where and how are all helping elaborate the topic of the departure. By contributing different kinds of information to the higher-level topic, they are continuing each other, and as such are coordinated to each other.

This same analysis can be applied to the earlier example (5.15) to account for the oddness of the H% in that discourse. Example (5.23) shows how a third sentence can continue the relationship the first two have with respect to some implicit topic. In (5.15), there is no easily inferable relationship, coordinating or subordinating, between (5.15a) and (5.15b). Pierrehumbert & Hirschberg argue the H% triggers a search for how (5.15b) elaborates (5.15a), and because there is no clear way (5.15b) elaborates (5.15a) the discourse sounds degraded. I argue the degraded nature of (5.15) is a result of it being difficult to infer any shared topic for the train leaving at seven and there being a full moon, i.e. it is difficult to imagine a way in which those two sentences could mutually elaborate some implicit topic. One way to test this is to add a third sentence that makes some implicit topic more apparent, and then see if an H% sounds fine. Imagine a scenario like in a heist movie where a group of people are going to rob a train. One member of the group utters the following:

(5.24)

a. The train leaves at seven.
b. There's a full moon tonight.
c. Our men are in position.

In this case the context provides a topic for the discourse, something like "train robbery," and the sentences of (5.24) together elaborate that topic. The relationship between each sentence is no longer difficult to infer, because the larger topic provides a common relationship. And now, when (5.15) is in this larger context, a terminal rise at the end of (5.15a) is fine. So, when (5.15) occurs decontextualized as an isolated, two-sentence discourse, there is no easily inferable relationship between (5.15a) and (5.15b), and as a result an H% ending (5.15a) sounds bad. But in a larger context like (5.24), where the relationship between each sentence is more easily inferable due to the presence of a unifying topic, an H% ending (5.24a) sounds much better. Therefore, the felicity of an H% is not an inherent feature of those two sentences side by side, but a feature of the packaging of those sentences into a larger discourse structure. Furthermore, the sentences of (5.24) are coordinated one to another because they together and equally are elaborating the superordinate topic. So when two unrelated sentences are put into a context where they are coordinated, suddenly a terminal rise sounds good, further evidence that one meaning of a rise is discourse coordination.

There is another way that (5.14) is different from elaborations, namely the reversibility of the two sentences. For many lists, the elements in the list could be listed in any order without changing the propositional content. So, saying "I like plums, apricots and pears" is equivalent to saying "I like pears, plums and apricots". Similarly, (5.14) could be read as either (5.14a)-(5.14b) or (5.14b)-(5.14a), modulo pronouns, and the meaning would remain largely unchanged:

(5.25)

a. The train leaves at seven. It'll be on track four.
b. The train is on track four. It'll leave at seven.

Both orderings of the sentences, seen in (5.25a) and (5.25b), result in more or less equivalent discourses. By contrast, a classic example of elaboration from Asher &

Lascarides (2003) cannot go through the same manipulation without dramatically changing the content of the discourse:

(5.26)

a. Max had a lovely meal last night. He ate lots of salmon.
b. Max ate lots of salmon. He had a lovely meal last night.

In (5.26a), S2 elaborates S1 by introducing the salmon as a way of providing more information about the lovely meal. By contrast, it is difficult to imagine for (5.26b) that the lovely meal provides more information about the salmon. Instead, a more likely interpretation is one where the lovely meal provides some explanation or background for the eating of the salmon. Regardless, the meaning of two sentences related by elaboration cannot undergo reversal without major changes in the meaning of the discourse. The sentences of discourse (5.14), however, can be reversed while leaving the meaning of the discourse largely unchanged. This is further evidence that (5.14a) and (5.14b) are not related by elaboration, but instead are coordinated one to another.

The examples and tests discussed above help motivate the reanalysis of the structure of example (5.14) from one where (5.14b) elaborates (5.14a) to one where (5.14) together elaborate some implicit topic. It may be easier to understand the contrast between the proposed structural representations if they are presented graphically. If we use vertical lines to represent the hierarchical subordination of an elaboration relation, then the Pierrehumbert & Hirschberg (1990) proposed structure would be represented as in (5.27), with a single relation Elaboration((5.14a), (5.14b)).

(5.27)

(5.14a)
|
(5.14b)

By contrast, my implicit topic analysis has an implicit topic that both (5.14a) and (5.14b) elaborate. If (5.14a) and (5.14b) are coordinated, and so at the same level,

then we get the following structure (with coordinated discourse segments linked via horizontal lines):

(5.28)

Implicit topic (ImplTop)
/ \
(5.14a) - (5.14b)

Instead of just the one elaboration relation, this analysis has two relations, ELABORATION(ImplTop,[(5.14a), (5.14b)]) and COORD((5.14a), (5.14b)).

The upshot of this analysis, visible in the graphical representations, is that the elaboration is not between (5.14b) and (5.14a), but between the complex argument of both (5.14b) and (5.14a) and the implicit topic. Furthermore, as (5.14a) and (5.14b) are coordinated to one another, the boundary tone at the end of (5.14a) precedes a coordination. The improved felicity due to a high boundary tone in this position could then mean that high boundary tones indicate an upcoming discourse segment is at the same hierarchical level as the current one.

Under this analysis of Pierrehumbert & Hirschberg's example, their data and my data are no longer in conflict. Their example no longer shows that rising terminal pitch indicates upcoming subordination. Instead, the generalization is that rises, or high boundary tones, indicate a hierarchically equal relationship between the current and upcoming sentence, i.e. that the prior and the subsequent will share the same kind of relationship to some hierarchically superordinate segment.

The discourse meaning of rising pitch

Given this analysis, what kind of generalization can we make about the meaning of a terminal pitch rise? Pierrehumbert & Hirschberg (1990) suggest that boundary tones can indicate relationships between sentences. I think this intuition is sound; it was the nature of what relationship a rise indicates that was problematic.

When a new discourse segment is attached to a larger discourse structure, there are generally three hierarchically distinctive ways it could be attached. As contrasted in this dissertation's studies, a current discourse segment could be attached

to its immediately preceding one at the same hierarchical level (by a coordinating relation) or at one level lower in the hierarchy (by a subordinating relation). In my examples, a rising terminal precedes more material at the same hierarchical level while a falling terminal indicates material at a lower hierarchical level.

The third way a new discourse segment could attach to the discourse is at a higher hierarchical level. Sometimes called a discourse pop, these jumps up in the discourse structure often correspond to the beginning of a new topic or paragraph. We have reason to believe that a sentence preceding such a jump up in the discourse hierarchy would tend to end with falling pitch. Production studies that have looked specifically at the prosody at topic or paragraph boundaries have found that speakers tend to have final lowering of pitch preceding the boundaries (Lehiste, 1982; Silverman, 1987; Yule, 1980). This suggests that when the next sentence is hierarchically lower *or* higher in the discourse (i.e. subordination or discourse pops), the previous sentence is terminated with falling intonation; on the other hand, when the next sentence is hierarchically at the same level (i.e. coordination), the previous sentence ends with rising intonation. The benefit of this account is it is simple and empirically testable. It may even be potentially diagnostic in ambiguous contexts of whether a new sentence is hierarchically at the same or different level as the previous one. If all else is equal, a hierarchically ambiguous attachment preceded by rising terminal pitch may be more likely to be coordinated than subordinated.

There are also distributional reasons to like this account of rising pitch indicating discourse coordination, because a new discourse segment being subordinated seems to be more common than a new one being coordinated. In the newspaper article used in the production study discussed in chapter 2, sixty segments were attached by subordination only and thirty by coordination. Moreover, this article was selected from the 23 articles in the DISCOR corpus because it actually had the highest proportion of coordination relations. The table below shows the distribution of relations in the corpus, showing 487 out of 648 total relations, or 75%, were subordinating.

Table 5.1: Distribution of discourse relations in new_data portion of DISCOR database

Subord		Coord	
Precondition	3	Alternation	3
Commentary	5	Narration	3
Attribution	28	Parallel	4
Explanation	32	Consequence	7
Source	69	Result	18
Elaboration	161	Contrast	40
Background	189	Continuation	86
Total	487	Total	161

It seems clear, at least for the newspaper articles used in the DISCOR corpus, that coordinating relations are less common and thus more marked. As terminal pitch rises also seem to be marked (see discussion above), then one would expect rises to be less common than falls. The fact that rises and coordinating relations are likely both marked suggests this analysis may be on the right track.

While I have proposed a correspondence between terminal pitch rises and discourse coordination, I would refrain from expecting a categorical relationship between the two. Even my own data, while demonstrating a significant effect of the rise on interpretation, was far from categorical. Moreover, there are a number of other meanings that rises can convey, e.g. a yes/no question or a speaker's commitments (Gunlogson, 2003). What a specific terminal pitch contour means in a specific context may depend on many factors, but one of those factors is the relation to a following sentence in the discourse.

Being empirically testable, this hypothesis that a rise indicates discourse coordination sets the stage for a range of follow-up projects. For example, how dependent are the results in these studies on the specific structure of the discourses used? Would the interpretation of terminal pitch rises continue to be linked to discourse coordination in other kinds of ambiguous discourses? Would the same result show up with two-sentence instead of three-sentence discourses? What if there is more preceding or subsequent material?

Prosody may also be able to help with the larger, longer-term goals of figuring out the nature of discourse structure more generally. Theories of discourse have some fairly uncontroversial examples, what Asher & Vieu (2005, p. 600) call the prototype relations, but others are less clear. And the distinction between coordinating and subordinating relations itself may be problematic in some ways; for a discussion of problems with and an alternative to the Coord/Subord contrast, see (Stede, 2007/08). Prosody might be able to help reveal the nature of the Coord/Subord contrast, or perhaps whether the strict binary of Coord and Subord is actually the best way to categorize discourse relations.

Chapter 6

Conclusions and Future Studies

This dissertation has presented studies examining both the production and the perception sides of discourse prosody. One purpose of pairing production and perception was to compare how production patterns relate to perception effects. The synthesized manipulations of prosody in the perception studies of chapters 3 and 4, listed in (6.1), were motivated from two sources.

(6.1)

1: Terminal pitch contour on S1 (sentence 1)
2: Terminal pitch contour on S2
3: Pause duration between S1 and S2
4: Pause duration between S2 and S3
5: Mean pitch and intensity on S2 and S3

First, these manipulations were grounded in the performance of one particularly successful speaker who was attempting to convey one meaning of an ambiguous discourse. In a study at the University of North Carolina, Chapel Hill, results showed only this one speaker out of twelve was consistently able to communicate to listeners her intended meaning of ambiguous discourses (Tyler, et al., 2011). The second motivation for the manipulations comes from larger-scale trends identified in discourse prosody production. These patterns, discussed in chapter 2, tend to show larger discourse boundaries and hierarchically higher discourse segments correlating with longer pauses and higher post-boundary maximum pitch and intensity. The results of the study in chapter 2 also showed correlations between larger discourse boundaries and higher max pitch, higher max intensity and longer pauses, as well as

coordinated discourse segments showing relatively longer preceding pauses, higher max pitch and higher max intensity than subordinated discourse segments.

The one manipulation that was found to drive the perception effect was the terminal pitch contour on sentence 1. This prosodic feature was not measured in chapter 2 and terminal pitch contours generally have not been examined in discourse prosody production studies. The decision to include this manipulation was motivated by the performance of the one successful speaker in the UNC study, not overall results of production studies. This suggests that production studies may be measuring features of the prosody that are more easily measured, but not necessarily the ones that are most important for listeners in the interpretation of discourse structure.

One follow-up question is whether the rising terminal pitch contrast that was relevant in perception actually shows up in the production data of chapter 2. This would test whether the rise/fall contrast was present in the data and simply not measured, or whether the speakers did not exploit a terminal rise/fall contrast in ways that correlated with their production of the discourse. To answer this question, I returned to the recordings in Chapter 2 and coded the terminal pitch contours of all discourse segments produced by all 10 speakers. Each terminal pitch contour was coded as either rising, flat or falling. So, given that rising pitch biased towards coordination interpretations in the perception studies, were coordinated segments more likely to be produced with rising terminal pitch than subordinated segments? This question was tested statistically with a Generalized Linear Mixed model with CoordSubord as a fixed effect and rising terminal pitch as a binary outcome (rise or no rise). The statistical model contained control variables for quotation, duration and number (see Chapter 2 for more about these variables). I also wanted to separate the effect on terminal pitch of a discourse segment being sentence-final or not from the effect of CoordSubord. To control for this sentence-finality, I included a control variable in the model that captured whether the discourse segment ended a sentence of not. There was also a random effect for subject to control for inter-subject variation. CoordSubord was found not to be a significant predictor of the presence of a terminal rise ($t=1.039$, $p=.299$). This shows that a segment's being coordinated vs. subordinated did not affect the likelihood of a speaker producing rising terminal pitch.

It is possible that the relevant contrast for speakers is not between rising and non-rising terminal pitch, but between non-falling and falling terminal pitch. In the perception studies, falling terminal pitch biased listeners towards more subordinated interpretations. Might coordinated segments be more likely to be produced with non-falling terminal pitch contours than subordinated segments? This was tested using a Generalized Linear Mixed model with the same set of predictor and control variables described above, but with falling pitch as a binary outcome (fall vs. non-fall). CoordSubord was found not to be a significant predictor of falling vs. non-falling terminal pitch ($t=.022$, $p=.917$). This indicates that the likelihood of a speaker producing non-falling terminal pitch was also not affected by the segment being coordinated vs. subordinated.

These results tell us that in the production study in Chapter 2, there is no independent correlation between a segment's status as coordinated or subordinated and whether it is completed with rising vs. non-rising pitch, or falling vs. non-falling pitch. And yet, the perception studies show that rising pitch can bias listeners towards a coordinated interpretation, at least for the discourses used in those studies. Taken together, these findings suggest a complex relationship between the production and perception of discourse prosody. The prosodic variation that is easy to measure or commonly found in discourse production may not be relevant in perception, or prosodic variation that is important in discourse perception may not be measured in production. It will be important to study both production and perception together, because perception results can inform production studies in addition to production studies informing perception studies.

In chapter 1, I reviewed ways prosodic disambiguation was similar and different for analogous structures in both sentences and discourses. I will now extend this discussion to include the results of this dissertation. In their work on prosodic disambiguation of a range of syntactic ambiguities, Price et al. (1991) examine the ability of prosody to distinguish appositional constructions from attached NPs or PPs (see Table 1.1). The appositional construction is similar to the Subord interpretation of my discourses, where some constituents provide more information about another. The alternative interpretation is different, however, as my Coord interpretations are

not similar in structure to the attached NP/PP structure. Nevertheless, we may gain insight into the relationship between sentence and discourse by comparing the productions biasing towards appositions on the one hand and my Subord interpretations on the other. To exemplify, the appositional construction "the Daleys" in (6.2a) tended to be produced with major prosodic breaks before and after the apposition. By contrast, (6.2b) tended to be produced with only a small break before and/or after the phrase "the dailies."

(6.2)

a. The neighbors who usually read, the Daleys, were amused.
b. The neighbors who usually read the dailies were amused.

It seems speakers set off "the Daleys" by creating large boundaries on either side, which were often cued by breath intake or a long pause (p. 2962). The Subord interpretations in my data, analogous to the appositions, were communicated through falling terminal pitch at the end of the first sentence. So while the sentence-level version is marked with pauses, the discourse-level version was conveyed with pitch. This is more evidence suggesting that discourse disambiguation relies more on pitch variation while sentential disambiguation relies more on duration.

What has not been addressed in the literature is a full, sentence-level analog to my discourse ambiguities. This would involve the same words being potentially interpreted as conveying three separate independent constituents or one constituent that is then further specified by the following two. For example, the spoken sentence in (6.3a) could be interpreted as either (6.3b) or (6.3c):

(6.3)

a. I gave gifts to my friends Joe and Melba
b. I gave gifts to my friends Joe and Melba. (J and M are speaker's friends)
c. I gave gifts to my friends, Joe, and Melba. (J and M are not speaker's friends)

In (6.3b), Joe and Melba are the speaker's friends, while in (6.3c) they are not. (6.3c) presents a list of targets of the speaker's gift-giving: the set of all of the

speaker's friends as well as the individuals Joe and Melba. The contrast then is between a list construction and one where a category is first specified (*my friends*) and members of that category are delimited. The bracketing of such a construction would be like the following:

(6.4)

a. [my friends] [Joe] [Melba]
b. [my friends [Joe and Melba]]

The two structures then have a different kind of boundary after *my friends* in the two interpretations. Similarly, the discourses in this dissertation are ambiguous between the three sentences describing independent events and S2 and S3 elaborating S1. The bracketing would be like the Joe and Melba example above:

(6.5)

a. [S1] [S2] [S3]
b. [S1 [S2 S3]]

For both the discourse-level and sentence-level examples of this kind of ambiguous construction, listing intonation (see chapter 5) could bias towards the three separate, independent constituents interpretation. That is, if you were to pronounce the discourse in (6.6) you would be biasing interpretation toward a list of three independent recipients. Similarly, chapters 3 and 4 demonstrate that saying (6.7) with the indicated terminal pitch contours biases interpretation towards each sentence describing separate, independent events. It seems then that listing intonation functions similarly at both sentence-internal and discourse levels of linguistic structure (see chapter 5 for more on listing intonation).

(6.6) I gave gifts to my friends↗, Joe↗, and Melba↘ (↗=rise, ↘=fall),

(6.7) I sat in on a history class↗. I read about housing prices↗. And I watched a cool documentary↘.

By contrast, the non-list interpretation would have different prosody at the sentence and discourse levels. The appositional construction of the Joe and Melba example in (6.4b) would have a prosody without any breaks, flowing fluently between *my friends* and *Joe*. This would be the most natural production of this kind of meaning, and would naturally map onto that meaning in perception. By contrast, the Subord interpretation of my discourses would generally be produced with a break and a terminal pitch contour between each sentence. Furthermore, the experiments in chapter 3 and 4 show that a production like S1↘ S2↗ S3↘ biases towards the Subord interpretation. While it might be unusual, it would not be infelicitous to produce (6.4b) with the discourse prosody of the subordinated interpretation:

(6.8) I gave gifts to my friends↘, Joe↗, and Melba↘

This production seems to indicate somehow that Joe and Melba are parenthetical. But producing the sentence in (6.8) does not convey a different meaning from the sentence in (6.3b), it could just sound awkward. Perhaps, then, it is not for semantic reasons that we don't say sentences like (6.8). It may seem strange to have terminal pitch contours like in (6.8) because the constituents being marked are simply so small.

In Chapter 1, I examined the claim that ambiguities can be prosodically disambiguated when their two meanings correspond to two different bracketings. Table 1.1 showed that there are many different kinds of bracketing contrasts that prosody can distinguish. This dissertation has begun to illuminate some of the ways prosody relates to discourses, and how those relationships may differ from or be similar to prosody's relationship to sentences. For example, the appositional construction and the Subord interpretation contain similar meaning relations but get produced with different prosody. By contrast, listing intonation can operate on lists composed of whole sentences or sub-sentential constituents like NPs and VPs (see chapter 5). Our understanding of discourse benefits from this fuller understanding of the ways discourse and sentence structures are similar and different in a range of contexts.

Relevance for language processing

Another approach to prosodic disambiguation has been to focus less on which structures speakers and listeners *can* disambiguate to see in what contexts and for what reasons speakers and listeners *do* disambiguate. While this literature has focused on a subset of syntactic ambiguities, we should not assume *a priori* that speakers relate to different kinds of ambiguities in the same way in all contexts. Our understanding of the context-dependence of prosodic disambiguation would benefit from a fuller treatment of the range of structures speakers and listeners *can* disambiguate, including discourse ambiguities. I will now briefly review some of the questions addressed in this literature, questions that could get different answers with different ambiguities.

One major focus has been to explore prosodic disambiguation in more natural contexts, to determine how generalizable the earlier findings are to a wider range of contexts (Allbritton, McKoon, & Ratcliff, 1996; Schafer, Speer, Warren, & White, 2000; Snedeker & Trueswell, 2003). Allbritton et al. (1996) wanted to see whether speakers normally produced disambiguating prosodic cues for ambiguous sentences in disambiguated contexts. They found that when these speakers were not told of the ambiguity, they did not produce disambiguating cues. When speakers were aware of the ambiguity and told to produce the sentences to convey one of the meanings, speakers with professional training were able to produce sufficiently disambiguating prosody for listeners to retrieve their intended meaning. But naïve, untrained speakers were still unable to provide sufficient disambiguating cues even when trying. Their study demonstrates that a speaker's experience, awareness of the presence of an ambiguity, and desire to convey a particular meaning can affect their production of disambiguating cues.

Schafer et al. (2000) seek to elicit more natural, conversational productions of ambiguous sentences, moving beyond using read speech like in Allbritton et al. (1996). They elicit these more natural productions by having speakers participate in a game task that elicits specific syntactically ambiguous utterances. They find that speakers produce disambiguating cues in fully disambiguated contexts, even without

being informed of the ambiguity, and listeners were able to retrieve the intended meanings. There are thus conflicting results, where Allbritton et al. found naïve speakers did not provide disambiguating cues and Schafer et al. found they did.

Snedeker and Trueswell (2003) propose that prosody interacts with the context in complex ways that could account for these conflicting results. First, they wanted to create an experiment where both speaker and listener are present. This is a setting that more closely approximates natural conversation, unlike Allbritton et al. (1996) and Schafer et al. (2000) who ran their production study independent of their perception study. For those two studies, speakers produced ambiguous sentences without a listener present, and those productions were recorded and later played for listeners to interpret. In the first study in Snedeker & Trueswell (2003), when speakers and listeners had ambiguous sentences in ambiguous contexts where both meanings were plausible, speakers produced disambiguating prosody that listeners were able to use to retrieve the intended meaning. In a second study, the speaker's context and the listener's context were different. For the speaker, the context heavily biased toward one meaning, while for the listener the context was just as ambiguous as before. Therefore, if the speaker continues to produce disambiguating cues, they are doing it despite the context making clear what the sentence would mean. In this second study, speakers did not produce disambiguating cues, and as a result listeners were unable to recover the intended meaning. Snedeker & Trueswell (2003) conclude that speakers only produce disambiguating prosody in contexts where other cues would not already disambiguate.

Kraljic & Brennan (2005) extend research on prosodic disambiguation to test for effects of audience design. They manipulate speaker and listener knowledge to see for whom disambiguating cues are created. They find speakers produce those cues as a result of their own needs, not listener needs. They also found, unlike Snedeker & Trueswell (2003), that speakers produced disambiguating cues regardless of speaker awareness of the ambiguity or a context that disambiguated or not.

One problem in this literature and a potential explanation for sometimes conflicting results, as pointed out by Hirsch & Wagner (2011), is that the ambiguities being used are sometimes structurally different. Snedeker & Trueswell (2003) use

high/low attachment ambiguities like (6.9) while Kraljic & Brennan (2005) use left/right ambiguities like (6.10).

(6.9) Tap the frog with the flower.
(6.10) Put the dog in the basket on the star.

Differences in results between the two studies could be due to contextual factors or the difference in structure. Hirsch and Wagner's results suggest the kind of ambiguous structure tested does make a difference for prosodic disambiguation. A more controlled account of which structures are actually being examined will improve our understanding not just what structures *can* be disambiguated but which ones speakers and listeners *do* disambiguate.

Future research

The results of the studies presented in chapters 3 and 4 make clear that listeners can use prosody to interpret the meaning of an ambiguous discourse. Furthermore, chapter 2 demonstrates along with the wider literature that the structure of a discourse can affect how speakers produce discourse. This indicates a relationship between the structure of discourse and prosody. The structural ambiguities used in chapters 3 and 4 have truth conditional effects, and so are at least partially semantic. This shows a relationship between the semantic and phonological components of an English speaker/listener's knowledge of English. Part of the input to the phonological system includes knowledge of the structure of discourse, and part of the interpretation of discourse meaning involves the interpretation of the discourse's prosody. That is, part of a speaker/listener's grammatical knowledge is knowledge of the structure of discourse. And this knowledge has behavioral effects, e.g. on the production and perception of prosody. An adequate account of a native speaker/listener's knowledge of American English must incorporate some representation of the structure of discourse in order to account for these findings.

More specifically, this dissertation has illuminated a systematic relationship between one particular kind of ambiguous discourse structure and the prosodic

contrast of rising vs. fall pitch. While the relationship is now clearer in this specific context, it is not at all clear how far-reaching this connection between discourse and prosody can be generalized. There are many follow-up studies that could help illuminate the scope of the discourse-prosody interface.

There are many possible ambiguous structures, with different amounts of preceding material, following material, or ambiguous material, that could be tested for prosody's ability to disambiguate them. An initial follow-up study would be to test ambiguous discourses that are composed of only two sentences, e.g. (6.12) instead of (6.11).

(6.11) I sat in on a history class [S1]. I read about housing prices [S2]. And I watched a cool documentary [S3].
(6.12) I sat in on a history class [S1]. And I read about housing prices [S2].

It is unclear whether the biasing effect of a rise towards the Coord interpretation would show up if there were only a single sentence S2 that was either subordinated or coordinated to S1. Might the fact that the discourse only contains two sentences, and that as a result a listing interpretation would comprise the less customary list-length of only two items, mean that the rise of listing intonation could not have the same effect on interpretation? Another variant would be if there was some preceding sentence S1 that was coordinated to S2 and an S3 was then ambiguously attached to S2, either by coordination or subordination. Could a rise/fall contrast bias interpretation of the attachment of this S3?

There is also a way to run a follow-up study extending my studies that would allow them to speak more directly to the existing research on prosodic disambiguation. Unlike most of the literature on prosodic disambiguation, I did not construct my stimuli specifically to test whether the size or location of a prosodic boundary could disambiguate my discourses. For example, in high vs. low attachment ambiguities (e.g. *tap the frog with the flower*), a larger boundary before *with the flower* biases interpretation towards a higher attachment of that PP. Instead, the motivation for my prosodic manipulations came in part from the production data discussed in chapter 2, but mostly from the productions of the one speaker in the UNC study who was

successful at communicating her intended meaning. A relevant follow-up study would construct stimuli with my discourses that use the size and location of prosodic boundaries to bias interpretation. If prosodic boundaries can disambiguate my discourses, then the rising pitch effect I found is simply an alternative way to do so. But if they cannot, then there is some distinct relationship between the kind of meaning contrast in my ambiguous discourses and rising pitch on the one hand, and prosodic boundaries and the ambiguous structures of other studies on the other.

More concretely, the Coord interpretation of my discourses has equal boundaries between each sentence. By contrast, the Subord interpretation has a different kind of boundary after S1 than after S2 (Coord:[S1] [S2] [S3] VS Subord: [S1 [S2 S3]]). A different kind of prosodic boundary, e.g. in terms of pause durations, could bias listeners' interpretations of the discourse. How to produce such a contrast is not immediately apparent, however. One theory would be that the boundary after S1 is larger than after S2 in the Subord interpretation, and as a result a longer pause between S1 and S2 should bias towards Subord. On the other hand, the production data of chapter 2 shows that subordinated discourse segments tend to be produced with shorter preceding pauses than coordinated segments. Would longer pauses after S1 bias towards subord? Or coord? Or does pause duration simply not affect the interpretation of this kind of ambiguous discourse?

There is also a question of how the relative ambiguity of the discourses relates to their perceived naturalness. The discourses that were used in chapters 3 and 4 were normed to be as ambiguous as possible. If speakers try to avoid unnecessary ambiguity in their speech, might these ambiguous discourses then be perceived as less natural than the less ambiguous discourses? If the ambiguous discourses are perceived to be less natural, then it would be difficult to generalize the results from those discourses to discourse more generally. One way to test this is by collecting naturalness judgments on the discourses and testing for a correlation between ambiguity and naturalness. Another way would be to run the same studies as those in chapters 3 and 4 using the less ambiguous stimuli. What effect does prosody have, if any, on the interpretation of discourses that are logically ambiguous but in practice strongly biased towards one or the other interpretation? These studies could identify

how generalizable the results for prosodic effects on discourse interpretation are for different degrees of naturalness and ambiguity.

These discourses have also been constructed so as to make the relations between sentences consistently either Narration (the coordinating relation) or Elaboration (the subordinating relation)[5]. And yet, I have discussed my results in terms of rises indicating coordination, not just Narration. This raises the question of whether pitch rises could similarly bias interpretation towards other kinds of coordinating relations, e.g. Result, Contrast, Parallel, Continuation. Does the generalization that rises indicate discourse coordination extend to other coordinating relations?

These proposed follow-up studies are incremental extensions of the studies carried out in this dissertation, and they are intended to extend our understanding of the relationship between prosody and discourse. But these incremental steps are part of a larger research goal, namely to better understand both the nature and communication of discourse structure. Research into the global structure of discourse generally uses semantic phenomena like anaphora and temporal relations between sentences to motivate proposed structures, so adding prosody into the discussion provides a new approach to these issues. There are multiple kinds of cues to the structure of discourse, and a more complete theory should take advantage of and incorporate them all. For this reason, it will be important to explore the relationships between various cues to discourse structure and how they interact. For example, does prosody contribute anything to discourse interpretation over and above discourse markers, which make relationships between sentences explicit? In what contexts are different cues to discourse structure used? Are they interchangeable to some degree? What reasons are there that in some contexts the structure of discourse is made more explicit while in others it is left more implicit? Also, how can we account for inter-

---

[5] These relation names are drawn from Segmented Discourse Representation Theory (SDRT). For a fuller discussion and definitions of these and other discourse relations, see (Asher & Lascarides, 2003).

subject variability in the use of cues to discourse structure? And what can answers to these questions tell us about how the structure of discourse is similar to/different from the structure of sentences?

In addition to these questions about how we communicate the structure of discourse, there is the more basic question of what it is that we know when we know the structure of discourse. It is a kind of grammatical knowledge, because it affects truth conditions and has effects in terms of behaviors like the production and perception of prosody. But while coherence relations between sentences seem to be pragmatic, as they tend to be cancelable, it is less clear how they relate to other kinds of pragmatic meaning, e.g. implicatures, presuppositions, conventional implicatures in the sense of Grice (1989) or Potts (2005).

Conclusion

In this dissertation, I have examined how prosody relates to discourse structure in production and perception. The structure of discourse is sometimes made explicit with forms like discourse markers, but often is implicit and requires complex inferencing on the part of interlocutors. Most work that has studied explicit markers of discourse structure has focused on lexical cues. For instance, the examples provided by Jasinskaya (2007) in her review of explicit markers of discourse are lexical (p. 11). This dissertation has explored prosodic cues to discourse structure in production and the use of prosodic cues to discourse structure in perception. This helps extend prosody's utility as a non-lexical marker of discourse structure. And given how complex, often implicit and underspecified discourse can seem, our understanding of discourse will benefit from examining all the cues at our disposal. That is how we can best account for what speakers know when they know the structure of discourse, and how speakers and listeners are able to communicate the structure of discourse to each other.

Appendix A: Full text of newspaper article used in Chapter 2 production study with paragraphing removed, as presented to participants

Politics & policy: blacks' increasing vocal opposition to violence is matched by strong opposition to crime bill ---- by Joe Davidson staff reporter of The Wall Street Journal. The Rev. Jesse Jackson, the often fiery Rainbow Coalition president, was subdued, reflective, nearly rhymeless. At a recent hearing of the Congressional Black Caucus brain trust on crime, he spoke solemnly, his voice breaking, of how some young black men feel "more secure in jails than on our streets." With tears in his eyes, he spoke of death in his own neighborhood here and the precarious position of black youth. "Nearly half of all murder victims are black," he said. "More blacks kill each other each year than were killed in the entire history of lynching." Yet, the Rev. Jackson assailed one of the prime legislative vehicles for dealing with that explosion of violence – the Senate-passed crime legislation that President Clinton backed in his State of the Union address. The measure, he declared, is an "ill-conceived bill" and a "Draconian . . . expensive non-remedy." The bill has widespread bipartisan support in the Senate. Lawmakers contend it represents the toughest and most comprehensive government attack yet on violent crime, an issue at the top of the public's list of concerns in opinion polls. But at a time when African-Americans increasingly are speaking out against black criminals and the "gangsta rap" that seems to glorify violence, the Black Caucus and others say the Senate bill is too concerned with punishment, and not enough concerned with the alleviation of the conditions that cause crime. The strong opposition to the measure presents a problem for President Clinton, whose support for the legislation places him at odds with a core group of Democrats who elected him. Citing Mr. Clinton 's embrace of one provision of the Senate bill – mandatory life sentences for criminals convicted of three violent felonies – the dean of the Black Caucus, Democratic Rep. John Conyers of Michigan, decries the "lock-'em up and throw away the key" approach that "only fools the public into believing that we're doing something about crime." The White House will try to assuage at least some opponents' concerns as Congress undertakes to reconcile the Senate bill with a much different House measure. Justice Department officials, who were criticized for not visibly exerting influence over the Senate bill last year, will play a more overt role in removing or modifying the more extreme provisions this year. Deputy Attorney General Philip Heymann plans to testify at House crime legislation hearings, and Mr. Clinton himself held out the carrot of help to endangered youth in his speech to Congress. "We have got to stop pointing our fingers at these kids who have no future," he said, "and reach our hands out to them." The question, though, is whether enough changes can be made to the bill to soften opposition to it. In addition to the Black Caucus, a range of others -- including the American Bar Association, the American Civil Liberties Union, the National Conference of State Legislators and many federal judges and prosecutors -- oppose stringent sentencing provisions in the bill. Other less controversial provisions in the 22.3 billion dollar legislation include authorization for 100,000 additional police officers, drug treatment and other crime-prevention programs. Some black leaders, such as the leadership of the Nation of Islam, have long spoken out

against crime and for the kind of values that make it unacceptable. But the mainstream civil-rights leadership generally avoided the rhetoric of "law and order," regarding it as a code for keeping blacks back. Law and order didn't mean justice, Mr. Jackson used to say, but "just us." In the past, many were hesitant to speak about crime in public because "the larger community would talk about 'lock them up and throw the key away' and hide behind black leaders in doing it," explains Rep. Craig Washington, the Houston Democrat who led the caucus hearing. Now there is escalating discourse within the black community about what it can and must do to stop crime. Just after the new year, Mr. Jackson held the first of several conferences focusing on just that. "The premier civil-rights issue of this day is youth violence in general and black-on-black violence in particular," he has said. His conference also noted the structural conditions that encourage crime – the sorry state of the black economy, high unemployment, poor education and a legacy of racism. "The black leaders recognize that if they don't step out front and engage in the discussion, that basically our young people are turning themselves into slaves," says Rep. Washington. Within the black community, there is "more public concern and debate about the appropriate level of response to increasing crime and violence." Many of the black leaders involved in the growing debate retain strong objections to the Senate bill, with its large number of mandatory minimum sentences, death penalties and federalization of local crimes. One of the Senate measures strongly opposed by most members of the Black Caucus has as its author one of its own, Illinois Democratic Sen. Carol Moseley-Braun. Her amendment would restrict prosecutorial discretion -- a point opposed by Attorney General Janet Reno -- by directing U.S. attorneys to prosecute as adults 13-year-olds charged with committing violent crimes with firearms. The provision would federalize many crimes currently prosecuted by the states. Yet, notes federal Judge Maryanne Trump Barry of Newark , N.J., who is chairwoman of the criminal law committee of the Judicial Conference of the U.S., there is no federal juvenile justice system to handle such cases -- no federal juvenile prisons, for instance, and no federal youth probation officers. The National Conference of State Legislatures is so opposed to the federalization of state crimes -- another provision in the bill, pushed by GOP Sen. Alfonse D'Amato of New York, would federalize all violent handgun crimes – that it recently wrote President Clinton to say "the Senate bill is inimical to principles of federalism, and we must oppose it." And a measure that would require states to adopt certain federal sentencing guidelines, such as mandatory minimum sentences, to get federal prison building funds is "coercive policy," complains Jon Felde, NCSL's general counsel. There are numerous mandatory minimum provisions in the legislation that Mr. Washington fears could be used in an unfair fashion against blacks who may be charged more harshly than whites for similar acts. And federal judges have "consistently, vehemently, and virtually unanimously opposed" mandatory minimum sentences, Judge Barry wrote to Senate Judiciary Committee Chairman Joseph Biden, Democrat of Delaware, in November. Other measures that caucus members say could be used in a discriminatory way are those that would make it a federal crime to conspire to participate in a criminal street gang and that provides the death penalty for drug kingpins even if no death can be shown to have resulted directly from their illegal activity. The Justice

Department has warned Congress that it thinks the drug kingpin provision is unconstitutional; the anti-gang measure will also be hit on constitutional grounds in the House. But Sen. Biden insists that the final legislation will include enough significant prevention and punishment provisions that liberals and conservatives alike will be able to endorse it. After all, he says, "everybody is kind of singing from the same hymnal on the broad strokes."

Appendix B: Full text of newspaper article used in Chapter 2 production study as segmented according to SDRT in the DISCOR corpus

0. Politics & policy :
1. blacks ' increasing vocal opposition to violence is matched by strong opposition to crime bill ----
2. by Joe Davidson
3. staff reporter of The Wall Street Journal
4. The Rev. Jesse Jackson , the often fiery Rainbow Coalition president , was subdued , reflective , nearly rhymeless .
5. At a recent hearing of the Congressional Black Caucus brain trust on crime ,
6. he spoke solemnly ,
7. his voice breaking ,
8. of how some young black men feel "more secure in jails than on our streets .
9. " With tears in his eyes , he spoke of death in his own neighborhood here and the precarious position of black youth .
10. " Nearly half of all murder victims are black , "
11. he said.
12. " More blacks kill each other each year than were killed in the entire history of lynching. "
13. Yet , the Rev. Jackson assailed one of the prime legislative vehicles for dealing with that explosion of violence –
14. the Senate-passed crime legislation that President Clinton backed in his State of the Union address .
15. The measure , he declared , is an "ill-conceived bill " and a " Draconian . . . expensive non-remedy. "
16. The bill has widespread bipartisan support in the Senate.
17. Lawmakers contend it represents the toughest and most comprehensive government attack yet on violent crime ,
18. an issue at the top of the public 's list of concerns in opinion polls .
19. But at a time when African-Americans increasingly are speaking out against black criminals and the "gangsta rap " that seems to glorify violence ,
20. the Black Caucus and others say
21. the Senate bill is too concerned with punishment , and not enough concerned with the alleviation of the conditions that cause crime .
22. The strong opposition to the measure presents a problem for President Clinton,
23. whose support for the legislation places him at odds with a core group of Democrats who elected him .
24. Citing Mr. Clinton 's embrace of one provision of the Senate bill –
25. mandatory life sentences for criminals convicted of three violent felonies –
26. the dean of the Black Caucus , Democratic Rep. John Conyers of Michigan , decries the "lock-'em up and throw away the key " approach that "only fools the public into believing that we 're doing something about crime . "
27. The White House will try to assuage at least some opponents ' concerns
28. as Congress undertakes to reconcile the Senate bill with a much different House measure .

29. Justice Department officials , who were criticized for not visibly exerting influence over the Senate bill last year , will play a more overt role in removing or modifying the more extreme provisions this year .
30. Deputy Attorney General Philip Heymann plans to testify at House crime legislation hearings ,
31. and Mr. Clinton himself held out the carrot of help to endangered youth in his speech to Congress .
32. "We have got to stop pointing our fingers at these kids who have no future ,"
33. he said,
34. " and reach our hands out to them . "
35. The question , though , is whether enough changes can be made to the bill
36. to soften opposition to it .
37. In addition to the Black Caucus , a range of others -- including the American Bar Association , the American Civil Liberties Union , the National Conference of State Legislators and many federal judges and prosecutors -- oppose stringent sentencing provisions in the bill.
38. Other less controversial provisions in the $ 22.3 billion legislation include authorization for 100,000 additional police officers , drug treatment and other crime-prevention programs .
39. Some black leaders , such as the leadership of the Nation of Islam , have long spoken out against crime and for the kind of values that make it unacceptable .
40. But the mainstream civil-rights leadership generally avoided the rhetoric of "law and order , "
41. regarding it as a code for keeping blacks back .
42. Law and order didn't mean justice ,
43. Mr. Jackson used to say ,
44. but " just us ."
45. In the past , many were hesitant to speak about crime in public
46. because " the larger community would talk about 'lock them up and throw the key away ' and hide behind black leaders in doing it, "
47. explains Rep. Craig Washington ,
48. the Houston Democrat who led the caucus hearing .
49. Now there is escalating discourse within the black community about what it can and must do to stop crime.
50. Just after the new year , Mr. Jackson held the first of several conferences focusing on just that .
51. " The premier civil-rights issue of this day is youth violence in general and black-on-black violence in particular , "
52. he has said .
53. His conference also noted
54. the structural conditions that encourage crime –
55. the sorry state of the black economy , high unemployment , poor education and a legacy of racism .
56. " The black leaders recognize
57. that if they do n't step out front and engage in the discussion ,
58. that basically our young people are turning themselves into slaves , "
59. says Rep. Washington .
60. Within the black community , there is "more public concern and debate about the appropriate level of response to increasing crime and violence . "

61. Many of the black leaders involved in the growing debate retain strong objections to the Senate bill , with its large number of mandatory minimum sentences , death penalties and federalization of local crimes .
62. One of the Senate measures strongly opposed by most members of the Black Caucus has as its author one of its own ,
63. Illinois Democratic Sen. Carol Moseley-Braun.
64. Her amendment would restrict prosecutorial discretion
65. -- a point opposed by Attorney General Janet Reno
66. -- by directing U.S. attorneys to prosecute as adults 13-year-olds charged with committing violent crimes with firearms .
67. The provision would federalize many crimes currently prosecuted by the states .
68. Yet, notes federal Judge Maryanne Trump Barry of Newark , N.J. ,
69. who is chairwoman of the criminal law committee of the Judicial Conference of the U.S. ,
70. there is no federal juvenile justice system to handle such cases -- no federal juvenile prisons , for instance , and no federal youth probation officers .
71. The National Conference of State Legislatures is so opposed to the federalization of state crimes
72. -- another provision in the bill , pushed by GOP Sen. Alfonse D'Amato of New York , would federalize all violent handgun crimes –
73. that it recently wrote President Clinton to say
74. " the Senate bill is inimical to principles of federalism , and we must oppose it ."
75. And a measure that would require states to adopt certain federal sentencing guidelines , such as mandatory minimum sentences , to get federal prison building funds is " coercive policy , "
76. complains Jon Felde , NCSL 's general counsel .
77. There are numerous mandatory minimum provisions in the legislation that Mr. Washington fears could be used in an unfair fashion against blacks who may be charged more harshly than whites for similar acts.
78. And federal judges have " consistently , vehemently , and virtually unanimously opposed " mandatory minimum sentences ,
79. Judge Barry wrote to Senate Judiciary Committee Chairman Joseph Biden, Democrat of Delaware , in November .
80. Other measures that caucus members say could be used in a discriminatory way are those that would make it a federal crime to conspire to participate in a criminal street gang
81. and that provides the death penalty for drug kingpins
82. even if no death can be shown to have resulted directly from their illegal activity .
83. The Justice Department has warned Congress
84. that it thinks the drug kingpin provision is unconstitutional ;
85. the anti-gang measure will also be hit on constitutional grounds in the House .
86. But Sen. Biden insists
87. that the final legislation will include enough significant prevention and punishment provisions
88. that liberals and conservatives alike will be able to endorse it .
89. After all , he says ,

90. "everybody is kind of singing from the same hymnal on the broad strokes . "

Appendix C: Praat pitch settings used in automatic measurements in Chapter 2

Appendix C Table 1: Praat pitch settings used in automatic measurements in Chapter 2

	Praat Default Setting	F0max setting	F0min setting
Voicing threshold	0.45	0.6	0.75
Octave cost	0.01	0.01	0.07
Voicing/voiceless cost	0.14	0.14	0.21

Appendix D: Paraphrase analysis in Chapter 2

Appendix D Table 1: Results of paraphrase analysis in Chapter 2

Topic #	Boundary Size (level 3)	Discourse Segment	Topic Content	Speakers who mention topic (n=10)
1			A Senate crime bill, including Jesse Jackson's concerns about it and Senate support for it	9
2	X	19	Bill is too focused on punishment and not enough on prevention; it is a "lock 'em up and throw away the key" approach	9
3	X	35	Can changes be made to bill to soften opposition to it	3
4	X	49	Black community discussing what it can do to stop crime	2
5	X	56	Quote: Importance for black leaders to address issue of crime	
6	X	80	Other potentially discriminatory measures in bill	1

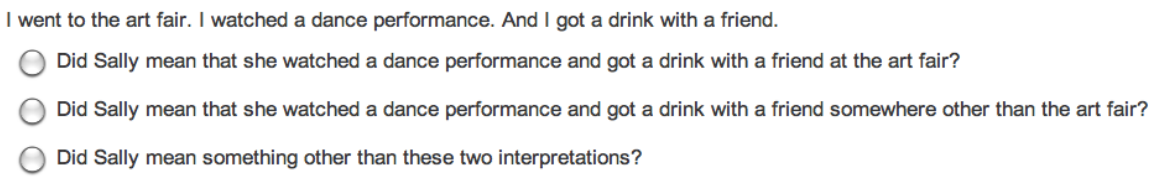
Appendix E: Norming study for stimuli used in studies in Chapters 3 and 4, with a table of full set of discourses selected for the studies

The development of a set of ambiguous discourse stimuli was necessary for the study of prosodic effects on discourse interpretation. Not only was it important to have a sufficiently large set of logically ambiguous discourses, it was important to test how practically ambiguous they were, i.e. whether listener-readers actually get both of the potential meanings originally intended. If it's possible to get two meanings but in practice people only notice one, then the discourse is not practically ambiguous. Having both meanings available was considered important to test for prosodic effects on interpretation. As a result, this norming study was run on the text of each discourse to determine their underlying bias and select the most ambiguous ones.

Method

The author and an undergraduate research assistant together created a list of 102 discourses, each designed to have two primary possible interpretations. The discourses are all three sentences long, with one interpretation having the second and third sentences as elaborating an event described in the first sentence and the other interpretation where all three sentences describe independent events.

Appendix E Figure 1: The target questions in the Qualtrics survey for the norming study



Procedure

The norming took place in an online survey through the Qualtrics survey research tool (Qualtrics Labs Inc., 2009). All participants participated through

Amazon's Mechanical Turk (Amazon Mechanical Turk, 2011) and were directed to the Qualtrics survey, where after completion they returned to Mechanical Turk and submitted their task for payment. Only participants in the US were allowed to participate. The first question asked if they were native speakers of American English, a listed pre-requisite for participation. Then, they answered the target questions. And finally they answered some demographic questions about gender, age, education level and knowledge of foreign languages.

Participants

Forty-seven total participants took part in this norming study via Amazon's Mechanical Turk service in exchange for payment. An initial group of ten subjects participated, after which it became clear five of the discourses had questions that did not match the discourse. These five discourses were fixed for the remaining 37 subjects.

For the first 10 subjects who saw five discourses that had questions that had nothing to do with the discourse presented, eight subjects gave the "other" interpretation for nearly all of these mismatches. I interpret this as indicating that they read the discourse and questions carefully enough to recognize the questions had nothing to do with the discourse. The other two subjects responded with one of the two main interpretations, suggesting they had not read carefully. They also finished the survey in around five minutes, while the others took around 23 minutes. Because they did not recognize the mismatch and finished so much faster, their data were excluded. I also used the experience with these subjects to establish a criterion where subjects who completed the survey in under 10 minutes would be excluded. Only one other subject was thusly excluded. In conclusion, of the 47 subjects who participated in this study, the responses from 44 serve as the basis of the analysis.

The 44 participants included in the analysis averaged 37 years of age and varied from a high school education to having a graduate degree:

Appendix E Table 1: Education levels for participants in norming study

Education Level	Total subjects
Did not complete high school	0
High school	2
Some undergraduate education	15
Undergraduate degree	16
Some graduate education	2
Graduate degree	9

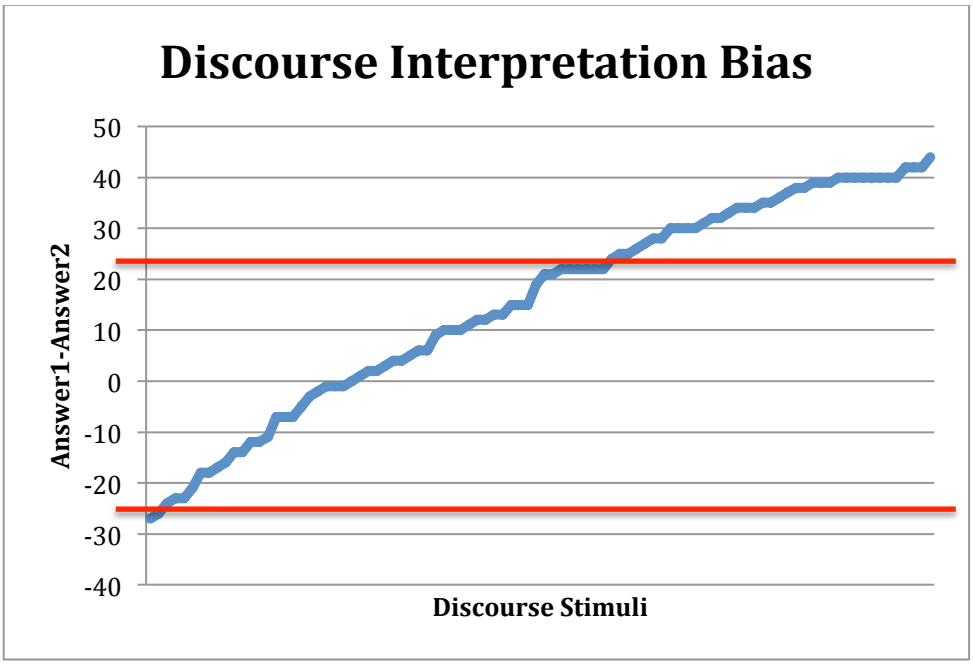
This indicates that 95% of this study's participants have at least some college education, suggesting that the participants in a Mechanical Turk study are actually quite well educated. 70% of the participants were female.

Results

One goal for the discourse stimuli was for the two proposed interpretations to be preferred over the "other" interpretation, meaning that participants are actually choosing between the two proffered interpretations and not getting distracted by other possible interpretations. Of all the discourses presented, six had the "other" interpretation chosen more than 20% of the time, and an additional two had the "other" interpretation chosen more than 10% of the time. These eight discourses are excluded from subsequent analysis and from the final list.

Another goal of the norming was to identify discourses where the two proposed interpretations were as close to equally preferred as possible. The following graph plots the discourses from most biased towards the coordinate interpretation (lower on y-axis) to most biased towards the subordinate interpretation (higher on y-axis):

Appendix E Figure 2: Graph of 102 normed discourses along x-axis, arranged from most biased towards coordination interpretations to most biased towards subordination interpretations. The y-axis shows the difference between the number of subordination interpretations and coordination interpretations, with higher positive numbers indicating a subordination bias and negative numbers indicating a coordination bias.



There is an overall preference for the subordinated interpretation as indicated by more discourses being above the zero line, indicating equibias, than below. The most ambiguous discourses were those that were closest to the zero line, indicating they received a more equal number of Coord and Subord interpretations. The 48 most ambiguous discourses are those that fall between the horizontal lines on the graph above.

Discussion

The results of the norming study separated discourses that were affected by "other" interpretations from those that were not. They also plotted the discourses for their underlying bias toward the coordinated or subordinated interpretation. A final set of 52 discourses were chosen that had under 10% of their interpretations chosen as "other" and a second best interpretation chosen at least 25% as often as the preferred interpretation. This final set of discourses is listed below. The Bias column is defined

as the absolute value of Answer1-Answer2. Therefore, the lower the number, the more equibiased the discourse.

Appendix E Table 2: The full set of discourses used in the perception studies in Chapters 3 and 4. Bias indicates the difference between the number of participants who chose the Coord interpretation and the number who chose the Subord interpretation. The discourses are ordered from most ambiguous (least biased) to least ambiguous.

Bias	Discourse Text
0	I visited my uncle in Detroit. I saw a movie. And I went for a run.
1	I spent the day at work. I played some ping pong. And I experimented with paper airplane designs.
1	I went to the gas station. I bought an apple. And I picked up some wine.
1	I sat in on a history class. I read about housing prices. And I watched a cool documentary.
1	I finished my senior project. i taught some kids how to tango. And I put on a show at school.
2	I did some work for class. I read about dogs. And I took some pictures.
2	I partied at my friend's house. I changed my status on facebook. And I spilled juice on my shirt.
2	I went to the art fair. I bought some dinner. And I saw a performance by the Pink Flamingoes.
2	I hung out with my boyfriend. I did some homework. And I played guitar.
3	I competed in a race. I built a raft. And I gave my friend Jason a pep talk.
3	I took a trip in my convertible. I played some disc golf. And I ate a lot of beef jerky.
4	I cleaned the kitchen. I vacuumed my new rug. And I took out the trash.
4	I got my living room ready for a party. I fixed the fire alarm. And I put away my clothes.
5	I went to the market. I met up with my advisor. And I ate some good food.
5	I laid in bed for a while. I ate a bowl of chicken soup. And I played with my cat.
6	I worked on a project with my neighbor. I baked a cake. And I put up decorations.
6	I relaxed on the sand. I played some chess. And I read a novel.
7	I am getting trained for my job at the mall. I am learning to be a better public speaker. And I am figuring out how to use my new smartphone.
7	I squeezed in a workout. I walked to my parents' house. And I helped my dad move some furniture.
9	I went for a hike. I hung out with my buddies. And I scavenged for seashells.
10	I go on dates whenever I can. I go to museums. And I occasionally go out for a drink.
10	I ran some errands. I picked my dad up from the airport. And I got take-out Chinese food for dinner.
10	I ate some breakfast. I enjoyed the sunshine. And I read a few chapters of my book.
11	I worked an eight hour shift. I did some crossword puzzles. And I got yelled at by a sketchy homeless guy.

12	I walked around the art fair. I gave a friend a pep talk. And I thought about the war in Afghanistan.
12	I played fetch with my dog. I practiced my frisbee technique. And I watched a soccer game.
12	I planned a practical joke. I bought a bucket of paint. And I wrote a letter.
12	I went to the grocery store. I got Starbucks coffee. And I picked up my photo prints.
13	I overreacted to a friend's comment. I went for a long walk. And I wrote in my diary for an hour.
13	I visited the state fair. I learned how to knit. And I saw my favorite band.
14	I stopped by my hometown. I wrote a bunch of thank-you notes. And I bought a new outfit at the mall.
14	I babysat for my neighbors. I baked a pie. And I gave my brother a call.
15	I went on a road trip. I played some disc golf. And I ate a lot of beef jerky.
15	I went to the library. I listened to a presentation about music. And I got a cup of coffee.
15	I went on a date. I saw a bunch of movies. And I almost fainted.
18	I did some work. I read a book. And I talked to my boss.
19	I went to English class. I drew some cartoons. And I gave a presentation.
21	I played on the computer. I read the newspaper. And I chatted with a friend.
21	I went to my brother's birthday party. I got a drink with Sharon. And I played some darts.
21	I worked on my computer. I listened to some music. And I looked at some photos.
22	I spent some time in Chicago. I went to the beach. And I saw an old friend from college.
22	I took care of some business. I bought some painting supplies. And I cashed a check.
22	I went home for Easter. I watched the NBA playoffs. And I ran in a race for the first time.
22	I took the dogs for a walk. I picked some wild berries. And I dropped off a letter at the mailbox.
22	I stopped by the market. I did some people watching. And I saw an accordion performance.
22	I picked up some stuff for my mom. I got some bird seed for the bird feeder. And I bought a couple rose bushes.
23	I went to my neighborhood block party. I cleaned my picnic table. And I went for a bike ride.
23	I played a game. I turned on my computer. And I relaxed on the couch.
24	I waited tables for a couple hours. I drank a glass of wine. And I watched the end of the football game.
24	I went to the hospital. I bought some flowers. And I ran into my neighbor.
26	I prepared for the tennis tournament. I did some meditation. And I got a

	massage.
4[6]	I played some ultimate frisbee. I caught up with my friend David. And I got bitten by a dog.

[6] This discourse was one of the five that originally had a mismatch between the discourse and the question, and so had a different total. While the other mismatches were excluded, this one was fairly equibiased based on the remaining 37 participants' data and so was included to get the set of discourses up to 52. The final four discourses, including this one, were part of the initial four practice discourses and so are not part of the target discourses.

References

Akkaya, C., Conrad, A., Wiebe, J., & Mihalcea, R. (2010). *Amazon Mechanical Turk for subjectivity word sense disambiguation*. Paper presented at the Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Los Angeles, California.

Allbritton, D. W., McKoon, G., & Ratcliff, R. (1996). Reliability of Prosodic Cues for Resolving Syntactic Ambiguity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(3), 714-735.

Allison, P. (2004). Convergence Problems in Logistic Regression. In M. Altman, J. Gill, M. McDonald & l. Wiley online (Eds.), *Numerical issues in statistical computing for the social scientist* (pp. xv, 323 p.). Hoboken, N.J.: Wiley-Interscience.

Amazon Mechanical Turk. (2011), from http://www.mturk.com

Arnold, J. E. (2008). Reference production: Production-internal and addressee-oriented processes. *Language and Cognitive Processes, 23*(4), 495-527.

Asher, N. (2004). Discourse Topic. *Theoretical Linguistics, 30*(2-3), 163-201.

Asher, N., & Lascarides, A. (2003). *Logics of Conversation*. xxii+526pp, Cambridge, UK: Cambridge U Press.

Asher, N., & Vieu, L. (2005). Subordinating and Coordinating Discourse Relations. *Lingua, 115*(4), 591-610.

Auran, C. (2007). *Discourse cohesion and its prosodic marking in French: interactions between intonation unit onsets and anaphoric pronouns in speech perception*. Paper presented at the ICPhS XVI, Saarbrücken.

Auran, C., & Hirst, D. (2004). *Anaphora, Connectives and Resetting: Prosodic and Pragmatic Parameters Interactions in the Marking of Discourse Structure.* Paper presented at the Speech Prosody, Nara, Japan.

Baldridge, J., Asher, N., & Hunter, J. (2007). Annotation for and Robust Parsing of Discourse Structure on Unrestricted Texts. *Zeitschrift fur Sprachwissenschaft, 26*, 213-239.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (under review). Random effects structure in mixed-effects models: Keep it maximal.

Beckman, M., & Pierrehumbert, J. (1986). Intonational Structure in Japanese and English. *Phonology Yearbook, 3*, 15-70.

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2010). *Using Mechanical Turk as a Subject Recruitment Tool for Experimental Research*. Retrieved from http://huber.research.yale.edu/materials/26_paper.pdf

Boersma, P., & Weenink, D. (2009). Praat: doing phonetics by computer. Retrieved from http://www.praat.org

Bowler, B., & Parminster, S. (1992). *Headway Pre-Intermediate Pronunciation*. Oxford: Oxford University Press.

Cauldwell, R., & Hewings, M. (1996). Intonation rules in ELT textbooks. *ELT Journal, 50*(4), 327-334. doi: 10.1093/elt/50.4.327

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*, 335-359.

Couper-Kuhlen, E. (2001). Interactional prosody: High onsets in reason-for-the-call turns. *Language and Society, 30*(1), 29-53.

Danlos, L. (2010). Strong generative capacity of RST, SDRT and discourse dependency DAGSs. In A. Benz & P. Kühnlein (Eds.), *Constraints in discourse*. Amsterdam: John Benjamins Publishing Co.

Davidson, J. (1994, January 27). Blacks' increasing vocal opposition to violence is matched by strong opposition to crime bill, *The Wall Street Journal*.

Degen, J. (2012). Random effects structure in mixed-effects models. *die welt ist alles, was der blog ist* Retrieved April 23, 2012, from http://jdegen.wordpress.com/2012/01/03/random-effects-structure-in-mixed-effects-models/

den Ouden, H. (2004). *Prosodic realizations of text structure.* University of Tilburg, Tilburg.

den Ouden, H., Noordman, L., & Terken, J. (2009). Prosodic realizations of global and local structure and rhetorical relations in read aloud news reports. *Speech Communication, 51*(2), 116-129.

Downs, J. S., Holbrook, M. B., Sheng, S., & Cranor, L. F. (2010). *Are your participants gaming the system?: screening mechanical turk workers*. Paper presented at the Proceedings of the 28th international conference on Human factors in computing systems, Atlanta, Georgia, USA.

Durkin, D. (1983). *Teaching them to read* (4th ed.). Boston: Allyn & Bacon.

Esser, J. (1988). *Comparing reading and speaking intonation*. Amsterdam: Rodopi.

Fintel, K. v. (1994). *Restrictions on Quantifier Domains.* University of Massachusetts.

Gibson, E., Piantadosi, S., & Fedorenko, K. (2011). Using Mechanical Turk to Obtain and Analyze English Acceptability Judgments. *Language and Linguistics Compass, 5*(8), 509-524. doi: 10.1111/j.1749-818X.2011.00295.x

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should We Trust Web-Based Studies? A Comparative Analysis of Six Preconceptions About Internet Questionnaires. *American Psychologist, 59*(2), 93-104. doi: 10.1037/0003-066x.59.2.93

Grice, H. P. (1989). *Studies in the way of words*. Cambridge, Mass.: Harvard University Press.

Grosz, B., & Hirschberg, J. (1992). *Some Intonational Characteristics of Discourse Structure.* Paper presented at the Proceedings of the 2nd International Conference on Spoken Language Processing, Banff, October.

Grosz, B., & Sidner, C. (1986). Attention, Intentions, and the Structure of Discourse. *Computational Linguistics, 12*(3), 175-204.

Gunlogson, C. (2003). *True to Form: Rising and Falling Declaratives as Questions in English*. New York: Routledge.

Hale, A. D., Skinner, C. H., Winn, B. D., Oliver, R., Allin, J. D., & Molloy, C. C. M. (2005). An investigation of listening and listening-while-reading accommodations on reading comprehension levels and rates in students with

emotional disorders. *Psychology in the Schools, 42*(1), 39-51. doi: 10.1002/pits.20027
Heeger, D. (2003). Signal Detection Theory, from http://www.cns.nyu.edu/~david/handouts/sdt/sdt.html
Herman, R. (2000). Phonetic markers of global discourse structures in English. *Journal of Phonetics, 28*(4), 466-493.
Hirsch, A., & Wagner, M. (2011). *Syntactic differences in the reliability of prosodic disambiguation*. Paper presented at the ETAP 2 (Experimental and Theoretical Approaches to Prosody), McGill University.
Hirschberg, J. (2008). Pragmatics and Intonation *The Handbook of Pragmatics* (pp. 515-537): Blackwell Publishing Ltd.
Hirschberg, J., & Grosz, B. (1992). *Intonational Features of Local and Global Discourse Structure*. Paper presented at the Proceedings of the Speech and Natural Language Workshop.
Hobbs, J. R. (1985). On the coherence and structure of discourse: Center for the study of language and information.
Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*(4), 434-446. doi: 10.1016/j.jml.2007.11.007
Jasinskaya, E. (2007). *Pragmatics and Prosody of Implicit Discourse Relations: The Case of Restatement.* Unviersität Tübingen.
Jayez, J., & Dargnat, M. (2008). *The Semantics of French Continuative Rises in SDRT*. Paper presented at the Constraints in Discourse.
Kahn, J. (2012). Modeling, from http://jmkahn.web.unc.edu/modeling/
Keating, P. (2004). D-prime (signal detection) analysis Retrieved November 30, 2011, from http://www.linguistics.ucla.edu/faciliti/facilities/statistics/dprime.htm
Kehler, A. (2002). *Coherence, Reference, and the Theory of Grammar*. Stanford, California: Center for the Study of Language and Information.
Kehler, A., Kertz, L., Rohde, H., & Elman, a. J. (2008). Coherence and Coreference Revisited. *Journal of Semantics (Special Issue on Processing Meaning), 25*(1), 1-44.
Kittur, A., Chi, E. H., & Suh, B. (2008). *Crowdsourcing user studies with Mechanical Turk*. Paper presented at the Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems, Florence, Italy.
Laan, G. P. M. (1997). The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style. *Speech Communication, 22*(1), 43-65.
Ladd, D. R. (1980). *The structure of intonational meaning: evidence from English*. Bloomington: Indiana University Press.
Lehiste, I. (1973). Phonetic Disambiguation of Syntactic Ambiguity. *Glossa, 7*(2), 197-222.
Lehiste, I. (1975). The phonetic structure of paragraphs In A. Cohen & S. G. Nooteboom (Eds.), *Structure and Process in Speech Perception: Proceedings of the Symposium on Dynamic Aspects of Speech Perception* (pp. 195-203). Berlin – Heidelberg – New York Springer.

Lehiste, I. (1982). Some phonetic characteristics of discourse. *Studia Linguistica, 36*(2), 117-130.
Lehiste, I., Olive, J., & Streeter, L. (1976). Role of duration in disambiguating syntactically ambiguous sentences. [10.1121/1.381180]. *J. Acoust. Soc. Am., 60*(5), 1199.
Levinson, S. C. (1983). *Pragmatics*. Cambridge [Cambridgeshire] ; New York: Cambridge University Press.
Lewis, D. (1979). Scorekeeping in a language game. *Journal of Philosophical Logic, 8*(1), 339-359. doi: 10.1007/bf00258436
Liberman, M., & Pierrehumbert, J. (1984). Intonational Invariance under Changes in Pitch Range and Length. In M. Aronoff & R. Oehrle (Eds.), *Language Sound Structure* (pp. 157-233). Cambridge MA: MIT Press.
Lieberman, P. (1967). *Intonation, perception, and language*. Cambridge: M.I.T. Press.
Ljolje, A. (2002). *Speech recognition using fundamental frequency and voicing in acoustic modeling.* Paper presented at the Proceedings of ICSLP, Denver, USA.
Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: a user's guide*. Mahwah, N.J.: Lawrence Erlbaum Associates.
Mann, W. C., & Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text, 8*(3), 243-281.
Marandin, J.-M. (2007). *Contours as constructions*.
Mayer, J., Jasinskaja, E., & Kölsch, U. (2006). *Pitch Range and Pause Duration as Markers of Discourse Hierarchy: Perception Experiments* Paper presented at the Ninth International Conference on Spoken Language Processing, Potsdam, Germany.
Müller, F. E. (1996). *Affiliating and Disaffiliating with Continuers: Prosodic Aspects of Recipiency*.
Ostendorf, M., Price, P., Bear, J., & Wightman, C. W. S. (1990). *The Use of Relative Duration in Syntactic Disambiguation.* Paper presented at the 3rd DARPA Workshop on Speech and Natural Language, San Mateo, CA.
Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making\, 5*(5\), 411\-419\.
Pierrehumbert, J., & Hirschberg, J. (1990). The Meaning of Intonation in the Interpretation of Discourse. In P. Cohen, J. Morgan & M. Pollack (Eds.), *Intentions in Communication* (pp. 271-311). Cambridge MA: MIT Press.
Polanyi, L. (1988). A Formal Model of the Structure of Discourse. *Journal of Pragmatics, 12*(5-6), 601-638.
Potts, C. (2005). *The logic of conventional implicatures*. Oxford: Oxford University Press.
Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S., & Fong, C. (1991). The use of prosody in syntactic disambiguation. *Journal of the Acoustical Society of America, 90*(6), 2956-2970.
. Qualtrics Labs Inc. (2009)
Quené, H., & van den Bergh, H. (2004). On multi-level modeling of data from repeated measures designs: a tutorial. *Speech Communication, 43*(1-2), 103-121. doi: 10.1016/j.specom.2004.02.004

Quené, H., & van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language, 59*(4), 413-425. doi: 10.1016/j.jml.2008.02.002

Reese. (2007). *Bias in Questions.* PhD, The University of Texas at Austin, Austin, Texas.

Reese, B., Denis, P., Asher, N., Baldridge, J., & Hunter, J. (2007). Reference Manual for the Analysis and Annotation of Rhetorical Structure.

Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., & Tomlinson, B. (2010). *Who are the crowdworkers?: shifting demographics in mechanical turk*. Paper presented at the Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems, Atlanta, Georgia, USA.

Schafer, A. J., Speer, S. R., Warren, P., & White, S. D. (2000). Intonational disambiguation in sentence production and comprehension. *J Psycholinguist Res, 29*(2), 169-182.

Schubiger, M. (1958). *English intonation; its form and function*. T√¯ºbingen: M. Niemeyer.

Silverman, K. E. A. (1987). *The structure and processing of fundamental frequency contours.* Unpublished Doctoral Dissertation, University of Cambridge, Cambridge, U.K.

Silverman, K. E. A., Beckman, M., Pierrehumbert, J., Ostendorf, M., Wightman, C. W. S., & Price, P. (1992). *ToBI: A standard scheme for labeling prosody.* Paper presented at the Proceedings of the 2nd International Conference on Spoken Language Processing, Banff, Canada.

Smith, C. L. (2004). Topic transitions and durational prosody in reading aloud: production and modeling. *Speech Communication, 42*(3-4), 247-270. doi: 10.1016/j.specom.2003.09.004

Snedeker, J., & Trueswell, J. (2003). Using Prosody to Avoid Ambiguity: Effects of Speaker Awareness and Referential Context. *Journal of Memory and Language, 48*(1), 103-130.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). *Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks*. Paper presented at the Proceedings of the Conference on Empirical Methods in Natural Language Processing, Honolulu, Hawaii.

Sprouse, J. (2011a). A test of the cognitive assumptions of magnitude estimation: Commutativity does not hold for acceptability judgments. *Language, 87*(2), 274-288.

Sprouse, J. (2011b). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. [Research Support, U.S. Gov't, Non-P.H.S.]. *Behavior research methods, 43*(1), 155-167. doi: 10.3758/s13428-010-0039-7

Sprouse, J., & Almeida, D. (to appear). Assessing the reliability of textbook data in syntax: Adger's Core Syntax. *Journal of Linguistics*.

Stede, M. (2007/08). RST Revisited: Disentangling Nuclearity. In C. Fabricius-Hansen & W. Ramm (Eds.), *'Subordination' versus 'coordination' in*

sentence and text – from a cross-linguistic perspective. Amsterdam: Benjamins.
Swerts, M. (1997). Prosodic Features at Discourse Boundaries of Different Strength. *The Journal of the Acoustical Society of America, 101*(1), 514-521.
Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychology and diagnostics: collected papers*. Mahwah, N.J.: Lawrence Erlbaum Associates.
Tanner, W. P., Jr., & Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review, 61*(6), 401-409.
Tyler, J., Kahn, J., & Arnold, J. (2011). *Speakers use prosody to communicate discourse structure, and listeners use that prosody in comprehending discourse structure*. Paper presented at the Experimental and Theoretical Advances in Prosody (ETAP) 2, McGill University, Montreal.
Van Kuppevelt, J. (1995). Main Structure and Side Structure in Discourse. *Linguistics, 33*(4(338)), 809-833.
Wagner, M., & Watson, D. G. (2010). Experimental and theoretical advances in prosody: A review. *Language and Cognitive Processes, 25*(7), 905 - 945.
Wales, R., & Toner, H. (1979). Intonation and ambiguity. In W. E. Cooper & E. C. Walker (Eds.), *Sentence Processing: Psycholinguistic Studies Presented to Merrill Garrett*. New York: Halsted Press.
Wichmann, A. (2000). *Intonation in Text and Discourse*. Essex: Pearson Education Limited.
Wickens, T. D. (2002). *Elementary signal detection theory*. Oxford ; New York: Oxford University Press.
WISE. (2006). Signal Detection Theory Tutorial Retrieved November 30, 2011, from http://wise.cgu.edu/sdtmod/index.asp
Wong, B. Y. L. (1986). Problems and issues in the definition of learning disabilities. In J. K. Torgesen & B. Y. L. Wong (Eds.), *Psychological and educational perspectives on learning disabilities* (pp. 1–26). New York: Academic Press.
Yule, G. (1980). Speakers' Topics and Major Paratones. *Lingua, 52*(1-2), 33-47.