The Identification and Function of English Prosodic Features

by

Mara E. Breen

B.A. Liberal Arts
Hampshire College, 2002

SUBMITTED TO THE DEPARTMENT OF BRAIN AND COGNITIVE SCIENCES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN COGNITIVE SCIENCE
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

SEPTEMBER 2007

Signature of Author: _____
Department of Brain and Cognitive Sciences
August 20, 2007

Certified by: _____
Edward A. F. Gibson
Professor of Cognitive Sciences
Thesis Supervisor

Accepted by: _____
Matthew Wilson
Professor of Neurobiology
Chair, Department Graduate Committee

The Identification and Function of English Prosodic Features

by

Mara E. Breen

Submitted to the Department of Brain and Cognitive Sciences
on August 20, 2007 in Partial Fulfillment of the
Requirement for the Degree of Doctor of Philosophy in
Cognitve Science

ABSTRACT

This thesis contains three sets of studies designed to explore the identification and function of prosodic features in English.

The first set of studies explores the identification of prosodic features using prosodic annotation. We compared inter-rater agreement for two current prosodic annotation schemes, ToBI (Silverman, et al., 1992) and RaP (Dilley & Brown, 2005) which provide guidelines for the identification of English prosodic features. The studies described here survey inter-rater agreement for both novice and expert raters in both systems, and for both spontaneous and read speech. The results indicate high agreement for both systems on binary classification, but only moderate agreement for categories with more than two levels.

The second section explores an aspect of the function of prosody in determining the propositional content of a sentence by investigating the relationship between syntactic structure and intonational phrasing. The first study tests and refines a model designed to predict the intonational phrasing of a sentence given the syntactic structure. In further analysis, we demonstrate that specific acoustic cues—word duration and the presence of silence after a word, can give rise to the perception of intonational boundaries.

The final set of experiments explores the relationship between prosody and information structure, and how this relationship is realized acoustically. In a series of four experiments, we manipulated the information status of elements of declarative sentences by varying the questions that preceded those sentences. We found that all of the acoustic features we tested—duration, f0, and intensity—were utilized by speakers to indicate the location of an accented element. However, speakers did not consistently indicate differences in information status type (wide focus, new information, contrastive information) with the acoustic features we investigated.

Thesis Supervisor: Edward A. F. Gibson

Title: Professor of Cognitive Sciences

Chapter 1

What is prosody?

Every native speaker of a language understands that a spoken message consists not only of what is said, but also the way in which it is said. Prosody is the word used to describe the characteristics of the acoustic signal which affect non-lexical meaning. It describes the way in which words are grouped in speech, the relative prominence of words in speech, and the overall tune of speech. It is comprised of psychological features like pitch, quantity, and loudness, the combination of which give rise to the perception of more complex prosodic features like stress (prominence), phrasing (grouping), and tonal movement (intonation).

Focusing on non-lexical meaning, prosody is not concerned with describing acoustic aspects of speech which determine a word's identity. For example, the difference between the noun *PERmit* and the verb *perMIT*, is not considered a prosodic one, even though it is the location of the prominence (stressed syllable, in this case) which determines the difference. Prosody is concerned, rather, with acoustic features that distinguish phrases and utterances from one another. For example, a sign occasionally seen in men's restrooms, reads:

(1) We aim to please. You aim too, please.

The humor in the sign arises from contrast of the prominence and phrasing of the words in the first sentence with the prominence and phrasing of the second. Therefore, the joke arises from prosody.

To study prosody empirically, it is necessary to translate psychological features into acoustic ones, which can be automatically extracted from speech and measured. Therefore, prosodic investigations, including those contained in this thesis, use the following acoustic measures, which have been shown to correspond to listeners' perception. The acoustic correlate of pitch is fundamental frequency (F0), which is measured in Hertz (Hz). The acoustic correlate of quantity is duration, which in the current studies will be measured in milliseconds (ms). Finally, there are several acoustic correlates of perceived loudness. Amplitude is an acoustic correlate of loudness, which is a measure of sound pressure. It is measured in Newtons per square meter (N/m^2). Intensity is another measure of sound pressure, which is computed as the root mean square of the amplitude. Intensity is measured in decibels (dB). Energy is a measure of loudness which accounts for the fact that longer sounds sound louder. As such, it is a measure of the square of the amplitude times the duration of the sound.

The identification of prosodic features

We have already seen in (1) how prosody can determine meaning differences in speech. To characterize the details of this relationship between prosody and meaning, we need to decide what the critical features are, just as to study the relationship between syntax and meaning we need to determine the important syntactic features. One approach to studying this relationship is to impose categories on the acoustic spectrum, which is what systems of prosodic annotation do.

Prosodic annotation

One way in which prosody researchers have addressed the problem of prosodic feature identification is to find out what humans actually hear. This is achieved by training human listeners to utilize coding schemes to tag speech with labels which correspond to perceptual categories. In this way, coders can generate prosodically-annotated corpora which can be used to ask questions about the function and meaning of prosodic features.

The most widely-used system of prosodic annotation is the Tones and Break Indices (ToBI) system of prosodic annotation (Silverman, et al., 1992), which was developed by a group of speech researchers as the standard system of prosodic annotation. However, in the time since the development of the ToBI system, several limitations of ToBI have been suggested (Wightman, 2002). In response to these limitations, prosody researchers have either changed the system to suit their own purposes, or proposed alternative systems. The second section of this thesis will present one such alternative proposal: the Rhythm and Pitch (RaP) system of annotation (Dilley & Brown, 2005), and compare inter-rater reliability for both ToBI and RaP on a large corpus of read and spontaneous speech.

The main challenge that prosodic annotation systems face is how to divide up the continuous acoustic space into relevant categories. Should the categories be simply perceptual in nature? Or should they, conversely, be determined by categories in meaning? Studies of inter-rater agreement, such as the one presented in the second chapter of this thesis, can attest to the effectiveness of a system in determining the perceptual categories of prosody, but cannot speak to the effectiveness of a system's determination of the meaning categories of intonation.

The function of prosody

As already indicated, prosody describes the relative prominence and grouping of words in speech. The question this definition immediately raises, is: why? Why are words grouped in particular ways? Why are some words more prominent than others? In sum, what is the function of prosody for communication?

Multiple functions of prosody have been proposed and explored experimentally which cover a wide spectrum of questions from many levels of language processing, from word recognition, to sentence processing, to discourse processing, to questions about emotion and affect in speech. The current discussion of the function of prosody will focus on the levels of sentence and discourse processing, and will explore the way certain prosodic features contribute to propositional meaning, and how other features contribute to information structure.

The function of phrasing

The first question we will address is: what is the function of prosodic phrasing? Or, why are the words of utterances grouped into phrases? There are several functions for phrasing, including production difficulty, pragmatic/semantic factors, and performance factors. For example, (2) provides examples of how difficulty with lexical access, in this case, of the low-frequency word "sextant," can results in particular word groupings.

(2) I was looking for the // um // sextant.

Another reason for particular patterns of phrasing is exemplified in (3), which provides an example of how pragmatic considerations can result in particular phrasings.

(3) The talk, as you know, starts at noon.

A final proposal about the function of phrasing is that it provides cues to the syntactic and semantic structure of a sentence. This function is demonstrated in (4).

(4) a. The cop saw // the robber with the binoculars.

b. The cop saw the robber // with the binoculars.

Upon hearing the rendition in (4a), a listener is most likely to interpret "with the binoculars" as an adjunct of "the robber," indicating that the robber possesses the binoculars. Conversely, in (4b), the listener is more likely to interpret "with the binoculars" as an argument of "saw," indicating that the binoculars are the instrument that the cop is using to see the robber.

This relationship between phrasing and sentence structure will be the focus of the second section of this thesis. In addition, this section will explore which acoustic cues give rise to the perception of phrase boundaries, by exploring whether acoustic cues to boundaries correlate with results obtained from prosodic annotations of boundaries.

The function of prominence

In addition to exploring the function of phrasing, this thesis will also explore the function of prominence, and, specifically, the proposal patterns of prominence arise from information structure. Information structure describes the role of utterances, or parts of utterances, in relation to the wider discourse, and how these roles change over the course of discourse.

Two wide categories have been proposed for information in discourse. They correspond to information that is either new or old in the discourse. These two categories appear under many different names, and with slightly different meanings, but old, or *given* information, can be thought of that information which is under consideration by all of the participants in the discourse. New, or *focused* information, on the other hand, is, according to Jackendoff, "the information in the sentence that is assumed by the speaker not to be shared by him and the hearer" (1972, p.230).

In addition, there are two categories of focused information: new and contrastive. If an entity is being introduced to the discourse for the first time, it is considered *new.* If an entity is meant to contrast with or correct something already in the discourse, as is *Damon*, in (5), then it is considered *contrastive.*

(5) a. Did Bill fry an omelet?

(6) b. No, *Damon* fried an omelet.

Several theorists have made explicit the relationship between certain types of prominences and categories of information structure (Jackendoff, 1972; Steedman, 2000; Pierrehumbert and Hirschberg,1990). For example, Pierrehumbert and Hirschberg argue that a high pitch on a stressed syllable (a H* accent) indicates that the speaker is adding the prominent entity to the discourse, and it is, therefore, new. Conversely, a steep rise to a high pitch on a stressed syllable (a L+H* accent) indicates that the speaker means to contrast the prominent entity with something already in the discourse.

The experiments in the third section of this thesis explore the relationship between acoustics and meaning in terms of prominence relations. They will explore the acoustic features that underlie the location of focus, and whether different acoustic features underlie different types of focus, specifically, new and contrastive focus.

Chapter 2

The importance of prosodic factors in understanding and producing language is well recognized by researchers studying many aspects of spoken language. However, the relationship between the perception of prosodic events and underlying acoustic factors is complex (e.g., Pierrehumbert, 1980; Choi, et al., 2005). Therefore, a useful and practical means for investigating prosody has been through human annotation of prosodic information. The current chapter will present the theoretical motivation and mechanics of two such annotation systems, ToBI (Tones and Break Indices) (Silverman, K., Beckman, M., Pitrelli, J. Ostendorf, M. Wightman, C. Price, P., Pierrehumbert, J. & Hirschberg, J., 1992)) and RaP (Rhythm and Pitch) (Dilley & Brown, 2005), which are both based on phonological theory. In addition, we will present two large-scale inter-transcriber agreement studies of both systems.

There are two approaches to prosodic annotation: phonetic and phonological. Phonetic intonation systems are analogous to systems of phonetic transcription of phonemes, like the International Phonetic Alphabet (IPA), in that they are concerned with annotating the pitch events of speech without reference to categories or meaning, and they intend the labels in the system to be applicable to any language. Phonological systems, on the other hand, attempt to derive meaningful categories, which differ across languages. The INTSINT (DeChristo & Hirst, 1987) and TILT systems (Taylor, 1992) are phonetic; ToBI and RaP are phonological.

Several attempts at universally applicable phonological prosodic annotation systems have been made (e.g., ToBI, and its related systems for other languages, and RaP). The sections that follow will describe in detail the motivation and mechanics of both the ToBI and RaP systems. In the sections on ToBI, we will first describe the motivation and history of the ToBI system, and some of the theory on which the system is based. Then, we will describe the components of a ToBI annotation and how a coder applies the ToBI system to speech. Finally, we will describe recognized limitations of the ToBI system, including a discussion of empirical research which fails to support the categories embodied by the ToBI labels. Following discussion of the ToBI system, we will present the motivation for the development of the RaP system, and the theory on which it is based. Second, we will provide details about the mechanics of RaP, and how the RaP system is applied to speech. Finally, we will describe the ways in which addresses the limitations of the ToBI system.

ToBI

The ToBI (Tones and Break Indices) system (Silverman, et al., 1992) was developed in the early 1990s by a group of researchers from the fields of psychology, phonology, and computer science. They intended ToBI to be a standard system which could be used not only across different labs but also across disciplines, such that its use would result in a large body of annotated speech. The following will describe the theory on which the ToBI system is based, the inventory and mechanics of the annotation system, and known limitations with the ToBI system.

ToBI History and Theory

A history of the development of the ToBI system for American English, and subsequent variants for other languages can be found in Beckman, Hirschberg, and Shattuck-Hufnagel (2005). The authors note that the ToBI conventions for pitch are

based on the theoretical work of Pierrehumbert and her colleagues (Pierrehumbert, 1980; Beckman & Pierrehumbert, 1986, Pierrehumbert & Hirschberg, 1988). The ToBI conventions for phrasing, or the prosodic grouping of words, are based on the work of Price et al., (1991) and Wightman, Shattuck-Hufnagel, Ostendorf, & Price (1992). An important tenet of the theory on which ToBI is based is the idea that tones in ToBI are determined paradigmatically, meaning that high or low tones are indicated with reference speaker pitch range, and not in relation to the local tones that immediately precede or follow. This approach is hypothesized to result in high and low tones which are comparable irrespective of the context in which they occur (Beckman, et al., 2005). In a following section, this paradigmatic approach to tonal labeling will be contrasted with the syntagmatic approach to tonal labeling embodied in the RaP system, in which tones are labeled as high or low with specific reference to preceding tones (Dilley, 2005).

Label Type	Intended to capture	ToBI	RaP
Metrical RaP: Rhythm tier	strong beat:	N/A	X, X?
	weak beat:		x, x?
	no beat:		no label
Tonal: ToBI: Tones tier RaP: Pitch tier	Prominent sylls:	High: H*, L+H*, H+!H*, !H*, L+!H* Low: L*, L*+H, L*+H	H*, L*, E*
	Non-prom sylls:	None	H, L, E
	Major boundary:	L-L%, H-H%, L-H%, H-L%, !H-L%	H, L, E optionally used to signal pitch change
	Minor boundary:	L-, H-, !H-	
Phrasal ToBI: Break Index tier RaP: Rhythm tier	Major boundary:	4	)),))?
	Minor boundary:	3	),)?
	No boundary:	2, 1, 0	no label

Table 1: Inventory of symbols and associated tiers for ToBI and RaP

ToBI Mechanics

A standard ToBI transcription consists of four tiers of symbolic labels which are time-aligned with the speech signal: an *orthographic tier* for labeling time-aligned text, a *tonal tier* for labeling pitch events, a *break index* tier for labeling perceived disjuncture between words and phrasing, and a *miscellaneous tier* for additional information. Recent proposed modifications to the ToBI system include a fifth tier, termed an *alternative* (or *alt*) tier (Brugos, Shattuck-Hufnagel, & Veilleux, 2006) where alternative choices for tonal and break index labels may optionally be indicated. Determination of prosodic labels is based both on a coder's perceptual impression of prosodic events, as well as on the visual characteristics of the fundamental frequency (F0) contour. The inventory of symbols used by the ToBI system is presented in Table 1. The tonal and break index tiers form the core of a ToBI transcription, and will be described in detail in the following sections.

Tonal Tier

An example of the ToBI tonal tier, and some of its associated labels, is provided in Figure 1[1]. The tonal tier enables the labeling of two kinds of information: pitch accents and phrasal tones. Pitch accents in the ToBI system are indicated on syllables of perceived prominence, i.e. syllables that are more perceptually salient than others in the utterance. Pitch accents in ToBI are binary; a syllable is either accented or unaccented. An important consequence of feature of the system is that the labeling of perceived prominence in ToBI is also binary. This aspect of ToBI will be contrasted below with the labeling or prominence in RaP, which allows for multiple levels of prominence labeling.

Pitch accented syllables are usually accompanied by a pitch excursion, which in ToBI can be a movement to either a high or a low pitch in the speaker's range. Pitch is a psychological concept, and is represented in prosodic annotation as its acoustic counterpart, which is fundamental frequency (F0), measured in Hertz (Hz). There are a total of eight pitch accent types, which are made up of high and low tones, and can be simple, bitonal, or downstepped. The full inventory of ToBI pitch accent labels is presented in Table 1. Simple pitch accents (H*, L*), are assigned to syllables where the perceived prominence is associated with a single tone, which is either a local pitch maximum (H*), or a local pitch minimum (L*). In addition, L* can indicate a perceptually prominent syllable in a stretch of speech that is in the low part of the speaker's range. Bitonal accents, (L+H*, L*+H, and H+!H*) are assigned to prominent syllables where both a high and low tone combine to lend prominence to a syllable. Specifically, the starred tone of a bitonal accent is associated with the stressed syllable, while the unstarred tone leads or trails the starred tone on an unstressed syllable. Finally, there are three "downstepped" variants of the simple and bitonal accents (!H*, L+!H* and L*+!H) which are used when the pitch of a given high tone is lower than that of a preceding high tone in the same phrase.

A total of eight tonal labels are also available for indicating hierarchical phrasal information. Three phrase accents (H-, !H-, and L-), are used to indicate pitch movement at a minor phrasal boundary, while two boundary tones (H% and L%) are used to indicate pitch movement at a major intonational phrase boundary. Because the theory of ToBI maintains that a major phrase always contains one or more minor phrases, a major

[1] Soundfiles corresponding to examples in this paper can be found at: http://tedlab.mit.edu/rap.html etc

phrase is always labeled with one of five phrase accent/boundary tone combinations (H-H%, L-L%, H-L%, !H-L% and L-H%). Each of the preceding labels indicates unidirectional rising or falling pitch movement, except L-H%, which generally indicates bidirectional (falling-rising) movement.

Figure 1 provides an illustration of speech with the associated ToBI tonal labels, and each label on the tonal tier will be described in turn. The H* on "Le-" indicates that a perceptually prominent syllable produced with a high tone. The L- on "-gumes" indicates a low tone associated with a minor (or intermediate) intonational phrase boundary, which will be defined in the next section. The L* on "good" indicates a perceptually prominent syllable in the low part of the speaker's range. The L* on "vi-" also indicates a perceptually prominent syllable associated with a low tone, although here it is associated with a pitch minimum, in that the pitch subsequently rises to the end of the phrase, indicated by the "H-H%" on "vitamins."

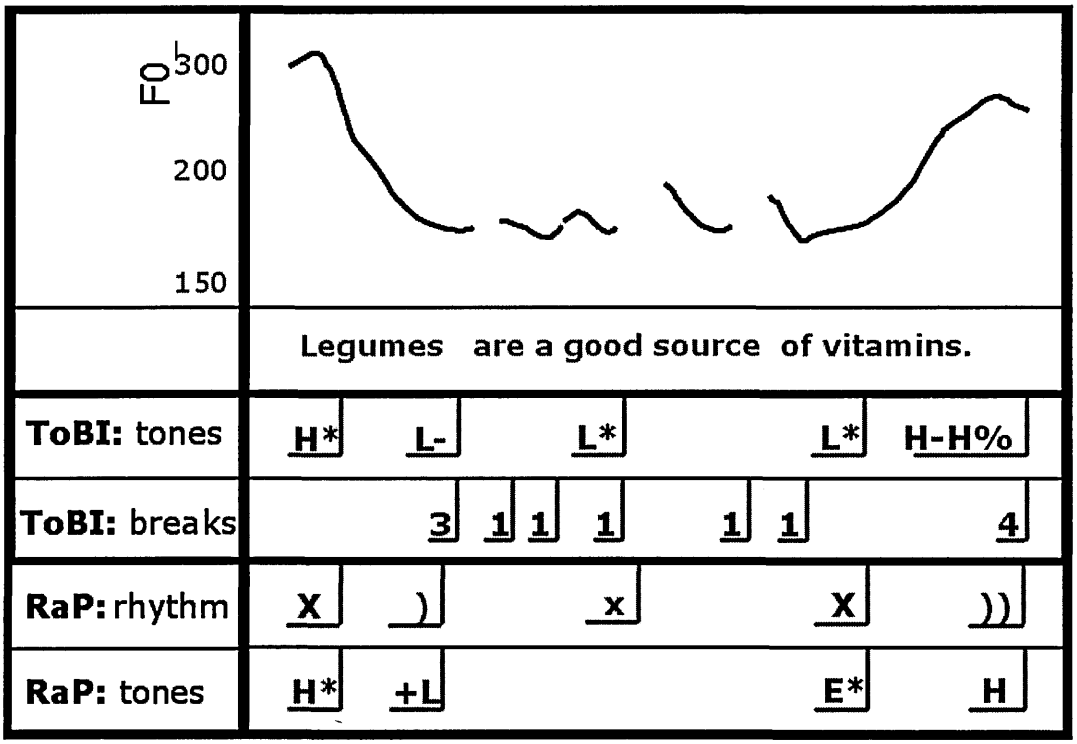


Figure 1. Example transcription of speech using the associated ToBI and RaP labels.

Break index Tier

A break index is a number from 0-4 which is assigned to the end of each word. In general, this number indicates the perceived degree of disjuncture between words, with the exception of the "2" label, which will be explained below. A "1" is used to indicate a small degree of disjuncture, as found at phrase-medial word boundaries. A "0" indicates a tight connection between words during fast speech or cliticization processes, e.g. *didja* for *did you*. A "3" indicates the perception of moderate disjuncture, which in the ToBI system corresponds to a minor phrase boundary. A "4" indicates maximal disjuncture, corresponding to a major phrase boundary.

There are two exceptions to the characterization of break indices as indicating degree of perceived disjuncture. The first stems from the stipulation that particular break indices must be used whenever a particular tonal label is indicated. In particular, a "3" or "4" must be labeled whenever a phrase accent or boundary tone, respectively, is labeled on the tonal tier, regardless of the perceived degree of disjuncture. Second, the break index "2" is used to explicitly indicate a mismatch between tonal movement and disjuncture information. Specifically, a "2" is reserved for word boundaries with "a strong disjuncture marked by a pause or virtual pause, but with no tonal marks; i.e. a well-formed tune continues across the juncture" or "a disjuncture that is weaker than expected at what is tonally a clear intermediate or full intonation phrase boundary" (Beckman & Hirschberg, 1994). Given this dual function, the "2" label can either indicate (1) a large degree of disjuncture comparable to a 4, or (2) a small degree of disjuncture comparable to a 1 (Pitrelli et al., 1994).

Once again, Figure 1 provides examples of ToBI break indices. The "3" associated with "-gumes" indicates a minor boundary after "Legumes," and is obligatorily accompanied by a phrase accent on the tone tier, "L-" in this case, indicating a falling pitch at the boundary. The "4" associated with "vitamins" indicates a major boundary. The label of "4" is obligatorily associated with a phrase accent/boundary tone combination, "H-H%," indicating a tone that rises through the final syllable.

Limitations of ToBI

Although ToBI was designed to be the standard intonation annotation system for the field, issues with the system have arisen throughout the years. The issues that will be addressed in this section include (1) a discussion of experimental evidence that fails to support the categories that are embodied in the ToBI system, (2) the way the ToBI system has moved away from annotation approaches which instruct annotators to "label what they hear" (Wightman, 2002), (3) low inter-transcriber reliability, and (4) the need for labels indicating multiple levels of perceived prominence.

The most serious issue with the ToBI system is that it may not always reliably capture the categories that speakers and listeners impose on intonation. Evidence for this claim comes from a series of production and perception studies which demonstrate that (a) speakers and listeners do not reliably produce or perceive the categories that ToBI defines and (b) speakers and listeners do reliably perceive and produce more categories than those defined by ToBI.

Evidence for the first claim—that speakers do not reliably perceive or produce the categories embodied in the ToBI system—comes from experiments conducted by Bartels & Kingston (1994), Ladd & Schepman (2003), and Watson, Tanenhaus, & Gunlogson (2004), on the distinction that the ToBI system assumes between H* and L+H*. ToBI assumes that these labels represent two categories of accents, both acoustically and semantically. Acoustically, both L+H* and H* are accents aligned with peaks in the high part of the speaker's range. Where they differ, however, is that H* is comprised of a single high tone, whereas L+H* is comprised of two tones, a low target and a high target. As a result, the rise to the peak of H* increases monotonically; conversely, the L+H* is realized as "a high peak target on the accented syllable which is immediately preceded by relatively sharp rise from a valley in the lowest part of the speaker's pitch range" (Beckman & Hirschberg, 1994). Semantically, H* and L+H* are said to differ with respect to the discourse status of the associated accented entity. Specifically, Pierrehumbert and Hirschberg (1990) state that speakers use H* to mark new information

that should be added to the listener's discourse representation, while L+H* is used to mark information that contrasts with something previously mentioned in the discourse.

To investigate the reality of an acoustic and semantic difference between H* and L+H*, Bartels and Kingston (1994) synthesized a continuum of stimuli intended to vary between H* and L+H* by independently manipulating four acoustic characteristics of the target accent. They then presented sentences containing these synthesized accents to naïve listeners and asked them to make a meaning judgment with the assumption that listeners would interpret entities accompanied by an H* as new to the discourse but interpret entities accompanied by L+H* as contrastive with information already in the discourse. The two main findings from their experiment were (a) that peak height, rather than the shape of the rise from a low tone, distinguished H* from L+H* such that the peak was higher for L+H* than for H*, and (b) that there was no clear evidence of a categorical boundary between L+H* and H* in terms of meaning differences.

More evidence for the lack of a categorical acoustic distinction between H* and L+H* comes from Ladd and Schepman (2003), who presented evidence which fails to support the claim that L+H* is comprised of two tonal targets, compared to only one target for H*. If H* is comprised of only one tonal target, then any dip in F0 between two H*s should not associate with a particular syllable. However, Ladd and Schepman demonstrated that speakers will align a low pitch between two H* accents with the /n/ in *Jay Neeson* or the /i/ in *Jane Eason,* depending on the intended production. In addition, listeners were able to use speakers' alignment of the low pitch to disambiguate the syllable membership of ambiguous consonants. These results suggest that the low pitch between two H*s is in fact a tonal target, calling into question the categorical difference between H* and L+H*, because ToBI assumes that it is only the latter accent in which the L has phonological significance.

Finally, Watson, Tanenhaus, & Gunlogson (2004) explicitly tested the idea that H* marks new material while L+H* marks contrastive material in an eye-tracking study. Listeners heard the directions in (3), while interacting with the items in a computerized display. The target item (e.g. camel/candle) in sentence 3c was crossed with the type of accent aligned with it in a 2x2 design.

(3)

a. Click on the camel and the dog.

b. Move the dog to the right of the square.

c. Now, move the *camel/candle* below the triangle.

L+H*/H*

Eye-tracking results demonstrated that listeners were more likely to look to the contrastive referent (the camel) when they heard the L+H* accent than to the new referent (the candle), suggesting that listeners quickly interpreted the L+H* accent as marking contrastive information. In contrast, listeners did not look more quickly at the new referent (the candle) when it was indicated with a H* accent. That is, they looked with equal probability at the camel or the candle when it was produced with a H*, indicating that H* is not preferentially treated as signaling new information. These results suggest, therefore, that H* is not perceived as categorically different from L+H*.

In addition to empirical evidence that the ToBI system does not embody the categories that speakers produce and listeners hear, it has also been criticized for the

grammatical constraints that it embodies, which are a result of the underlying theory which ToBI assumes. Wightman (2002) notes that the development of the system has led to a series of "linkage rules" whereby a label on one tier necessitates a particular label on the other tier. For example, a label of '3' on the break index tier must be accompanied by a phrase accent (L-, H-, or !H-) on the tonal tier. Similarly, a label of '4' on the break index tier must be accompanied by a boundary tone (L-L%, H-H%, L-H%, H-L%, !H-L%) on the tonal tier. Wightman argues that this interdependence between tiers leads labelers away from labeling what they actually perceive.

The third noted limitation of the ToBI system is that, although previous studies have noted high agreement for coarse binary comparisons like accent vs. non-accent, labelers have demonstrated fairly low agreement for fine-grained distinctions such as between accent types, and between phrase types. For example, although labelers in one study exhibited 81% on the binary measure of the presence or absence of a pitch accent, their agreement on the type of pitch accent dropped to 64%, even when the least-agreed-upon categories (i.e. H* vs. L+H*, !H* vs. L+!H*) were collapsed (Pitrelli, et al., 1994).

The final observed limitation of the ToBI system is the fact that it allows for only a binary perceived prominence distinction. That is, syllables in ToBI annotations are either pitch accented or unaccented. Conversely, throughout the history of intonational theory, several researchers have proposed systems in which there are three or more levels of prominence. For example, Halliday (1967) proposes a system in which there are categories of stress and categories of accent. Accent is defined in terms of pitch, but stress is defined with reference to rhythm. Empirically, there is evidence that three categories of accent can be useful. Greenberg, Harvey, and Hitchcock (2002), for example, found systematic differences between the pronunciation of words which were produced with either a "heavy" accent, a "light" accent, or no accent. Finally, Beaver, et al. (2007) have recently argued that second-occurrence focus, though not associated with pitch changes, is still perceived as prominent, and therefore necessitates a way of being indicated as prominent. As will be explained below, RaP instantiates a system with three levels of prominence, in which metrical prominences are defined with reference to rhythm, and pitch accented syllables are defined with reference to intonation.

One of the goals of the current chapter is to investigate whether a system without the above limitations lends itself to higher inter-coder reliability. RaP addresses the described limitations of ToBI in the following ways: First, tonal categories in RaP are based on empirical investigations of the categories that speakers produce and listeners perceive; Second, RaP allows for the independence of labels on different tiers; Finally, RaP allows annotation of three levels of prominence, as opposed to only two in the ToBI system.

RaP

The RaP (Rhythm and Pitch) system (Dilley & Brown, 2005) was developed to meet the needs of the speech research community by building on experimental work and theoretical advances that have taken place since the development of the ToBI system. The following sections will describe the theory on which RaP is based, the mechanics of the labeling system, and the way it addresses the limitations of ToBI.

RaP Theory

RaP is based on the work of Dilley (2005), which in turn draws heavily on the work of Pierrehumbert and colleagues (Pierrehumbert, 1980; Beckman & Pierrehumbert, 1986; Goldsmith, 1976; and Liberman & Prince, 1977). Tonal labels in the RaP system are based on the Tone Interval Theory of intonation and tone proposed by Dilley (2005). Tone Interval Theory differs from the tonal theory upon which the ToBI system is based in two important ways: First, whereas tones in ToBI are usually defined with reference to the global pitch range of the speaker, the tones in RaP are defined with reference to the preceding tone. Second, whereas ToBI postulates only high and low tones, RaP labels reflect three types of tonal relations; a tone may be higher, lower, or at the same level as a directly preceding tone.

Evidence that Tone Interval Theory provides a more accurate picture of the categories underlying speaker's production comes from Dilley's thesis. In a series of imitation experiments, Dilley demonstrated that the relative pitch level of one tone (higher than, lower than, the same as a preceding tone) was perceived and produced categorically. For example, where the ToBI system hypothesizes that H* and L*+H are two categorically different accents, the RaP system predicts no categorical distinction between these two accents.

Overview of the RaP Prosodic Transcription System

A RaP transcription is based on coders' auditory-perceptual impressions of prosodic events. A visual display of the signal is considered an aid, not a requirement, unlike ToBI. A transcription consists of four tiers of symbolic labels which are time-aligned with the speech signal: a *words tier* for labeling time-aligned syllables, a *rhythm tier* for labeling speech rhythm, a *tonal index* tier for labeling tonal information, and a *miscellaneous tier* for additional information. In the following discussion we focus on the rhythm and tonal tiers, which form the core of a RaP transcription.

Rhythm tier

The rhythm tier permits speech rhythm to be labeled by designating each syllable as a metrical beat or not a beat. Several levels are distinguished. The label **X** is used to indicate a syllable which is perceptually a strong metrical beat. The label **x** is used to indicate a syllable which is perceptually a weak metrical beat. In addition, phrasal boundaries are indicated at word boundaries with the following notation: ')' for a minor phrase boundary; '))' for a full phrase boundary; no label for phrase-medial word boundaries. RaP coders may indicate tonal labels in the tonal tier to account for perceived pitch movement at phrasal boundaries, but they are not, as in the ToBI system, obligated to do so.

Tonal tier

By separating tonal information from rhythmic information, the RaP system makes it possible to distinguish syllables which are prominent due to the presence of a

pitch excursion (accented by means of pitch and thus pitch accents), versus syllables which are prominent for rhythmic reasons. The tonal tier consists of labels indicating the tonal targets that correspond to each syllable. All tonal events are indicated with a H (high), L (low), or E (equal) label. In addition, RaP allows coders to indicate two sizes of pitch change: small or large. Small pitch changes are indicated on syllables which incur a pitch change of less than three semitones from the previous syllable, and are indicated with the ! diacritic (e.g. !H, !L). Tonal targets which correspond to metrically prominent (strong beat or weak beat) syllables (which are labeled with 'X' or 'x' in the metrical tier) are called starred tones and are indicated with the * diacritic (e.g. H*). Furthermore, tonal targets that occur on non-metrically prominent syllables (unstarred tones) that precede or follow a starred tone by one syllable are labeled with a '+' (before or after the tonal label depending on the location of the starred tone) indicating their association with the adjacent starred tone (e.g. +H). Finally, as indicated above, tonal labels can be indicated on phrase-final syllables which incur a pitch change.

In summary, there are two ways in which the RaP metrical and tonal tiers contrast with those in ToBI. First, by allowing coders to indicate syllables which are metrically prominent but do not incur a pitch change, RaP allows for three levels of prominence labeling, as opposed to the two that ToBI allows (pitch accented vs. non-pitch accented). Second, by assuming independence between phrasal and tonal labels, RaP allows labelers to indicate perceived disjuncture without the perception of a pitch change.

Motivation for Study One

There are two independent, though equally important, motivations for Study One. One is to conduct an inter-coder reliability study of the ToBI system, with an adequate number of trained coders, a large corpus of speech, and the appropriate statistical measures of agreement between coders. The second is to assess agreement of coders using the RaP system.

Several studies of inter-coder reliability in the ToBI system have been published since the development of the system; however, each has suffered from empirical limitations. In the first study of ToBI inter-coder reliability (Pitrelli, et al., 1994), 26 coders applied the ToBI conventions to 489 words, taken from both read and spontaneous speech corpora. Although this study employed many coders, the agreement results can be questioned because of the small amount of speech that each coder labeled, and because the agreement metric used to compare coders did not take into account the possibility of chance agreement. A more recent study by Syrdal & McGory (2000), employed six coders to label 645 words. Although this study did take into account chance agreement by employing the kappa statistic as described below, the agreement numbers are also questionable because of the composition of the corpus of speech. First, the corpus of speech is very small in comparison to the current study. Second, it is comprised only of two speakers who read the same words, which suggest that the results may not generalize to all speakers or to spontaneous speech. The final study of inter-coder reliability in the ToBI system was conducted by Yoon, et al. (2005). Although this study used appropriate statistical measures and a large corpus of spontaneous speech, including 79 speakers and 1600 words, the speech was labeled by only two coders.

The studies presented in this paper were designed to address the limitations of previous inter-coder reliability studies by employing (a) an adequate number of trained coders, (b) a large corpus of varied speech, and (c) the appropriate statistical measures of

agreement between coders. In addition, they represent the first study of inter-coder agreement for the RaP system.

Study One

The first study was designed to assess how transcribers with no previous prosodic annotation experience would utilize both labeling systems. Undergraduates were trained on both of the systems, and annotated fifty-five minutes of both read and spontaneous speech in both systems. Their annotations were analyzed for agreement on multiple prosodic characteristics, which will be described below.

Method

Participants

Five MIT undergraduates served as coders in the first study. Each coder committed to spending a full summer on the project for which they received either course credit or monetary compensation, at a rate of $8.75/hour for the duration of the project. Although four of the coders had taken an introductory linguistics course, none had any knowledge of prosody research, or any experience with prosodic annotation.

Corpus

There is some evidence that read speech is inherently different from speech which is spontaneously produced. Hirschberg (1995), for example, notes that read and spontaneous speech differ with respect to the intonational contours that speakers choose for questions, and that read speech is faster than spontaneous speech. In order to ensure that the results are applicable to diverse styles of speech, and to investigate differences in coding agreement across different speech styles, materials for the present study were drawn from two speech corpora: the Boston Radio News Corpus of read professional news broadcast speech (Ostendorf, Price, & Shattuck-Hufnagel, 1995), and the CallHome corpus of spontaneous nonprofessional speech from telephone conversations (Linguistic Data Consortium, 1997). The amount of speech from each corpus which was labeled in each system is shown in Table 1.

Table 1. Amount of speech (in minutes and syllables) from each corpus labeled in each system, including number of coders per file. Speakers are the same for both ToBI and RaP-annotated files.

System	Corpus	Minutes	Syllables	Coders/File	Unique Speakers
ToBI	CallHome	15.2	3680	3.5	6
	BRNC	20.9	5939	3.4	6
RaP	CallHome	9.6	2638	4.5	6
	BRNC	9.6	2889	4.7	6
	Total	55.2	15146		12

Coder training

Training and testing on prosodic systems occurred in three successive stages. First, coders trained and were tested on the ToBI system, and then applied this system to the speech corpus. Next, the coders trained and were tested on the RaP system, then applied it to a subset of the corpus which had already been labeled with ToBI. More details about training and labeling of the test materials are given below.

Training and testing of ToBI

Initial training on ToBI involved reading the associated manual and completing the computerized exercises in Beckman and Ayers (1997), as well as receiving one-on-one feedback from an expert coder (the author). In addition, all naïve coders participated in weekly meetings with a group of four expert ToBI labelers throughout the course of the project (the author, and three ToBI experts in the MIT speech community). After two weeks of initial training and feedback, the coders annotated one minute of read speech from the BRNC corpus, which was not included in the ToBI test set. Feedback from two expert coders (the author and Laura Dilley) was provided. Subsequently, the coders annotated one-minute of spontaneous speech from the CallHome corpus, which was again not included in the ToBI agreement analyses. Again, feedback from the two expert coders was provided.

After these two feedback rounds, the coders labeled 90 seconds of speech (approximately 60 seconds read speech, 30 seconds spontaneous), which were once again separate from the ToBI test materials. The annotations were anonymously evaluated by three expert coders (the author, Laura Dilley, and one of the MIT experts) using the following system: One or two points were deducted for each label with which the expert mildly or moderately disagreed, respectively. Three points were deducted when a label was strongly disagreed with and/or presented incorrect ToBI syntax. Experts also employed a subjective grading system ranging from excellent (5) to poor (1), indicating their overall impression of the labels.

Three coders received average grades of 4 or higher from all three expert evaluators on both test files and began annotating the corpus. The other two coders received average grades of 3 from the experts, and were instructed to go back through the ToBI manual, paying attention to the labels they had misused in the test labels. After another week of training, they too began corpus annotation.

Coders spent the next four weeks annotating 26.7 minutes of the corpus with the ToBI system (11 spontaneous, 15.7 read). The order of files in the corpus was the same for every coder, and pseudo-randomly determined so that coders would label approximately equal amounts of radio and spontaneous speech, and not label speech from the same speaker in adjacent files.

Following training and testing on the RaP system (as described below), coders annotated the next 9.4 minutes of the corpus using ToBI (4.2 spontaneous, 5.2 read). Inclusion of a second period of ToBI labeling permitted testing of the hypothesis that higher agreement might result from more labeling experience in general, regardless of the identity of the prosodic labeling system.

Training and testing of RaP

After the initial period of learning and applying ToBI, the coders spent two weeks learning the RaP system. Coders were introduced to RaP using the guidelines laid out in Dilley and Brown (2005)[2]. After an initial week of intensive group training with the manual, coders annotated a one-minute passage of read speech, and received feedback on their annotations from an expert RaP coder (Laura Dilley). Coders then labeled a one-minute passage of spontaneous speech and again received feedback from the expert coder.

---

[2] Available at http://tedlab.mit.edu/rap.html

After these two feedback rounds, the coders all labeled 60 seconds of speech drawn from both the spontaneous and read corpora. The expert RaP coder gave the novice coders quantitative and subjective scores for their annotations, as described above. All coders received scores of "4" or above, and were cleared to begin annotating the corpus according to the RaP conventions.

Coders spent the next four weeks annotating 19.2 minutes of the corpus using the RaP system (9.6 spontaneous, 9.6 read). The files annotated with RaP were a subset of the 26.7 minutes of the corpus labeled in the first four weeks of ToBI annotation.

Data analysis

Agreement metrics

All agreement analyses in the current study were designed to facilitate comparison of current results to those of previous studies of ToBI inter-coder agreement. Two measures of coder agreement were computed for the current study: one based on raw agreement; the other correcting for chance agreement. Whereas in previous studies of ToBI inter-annotator reliability accent labels are aligned with words (consistent with the established Guidelines for ToBI Labeling (Beckman & Ayers, 1997)), accent annotations in this study were aligned with syllables. This alignment scheme allowed us to make direct comparisons between accent placements in both systems. However, it should be noted that this scheme is different from previous studies of agreement, and necessitates two types of raw agreement metrics, explained below.

Following the work of Pitrelli et al. (1994), our raw agreement measures are based on the *transcriber-pair-syllable (TPS)* or *transcriber-pair-word (TPW)* depending on the relevant comparison. For example, agreement on pitch accents was computed syllable-by-syllable, while agreement for phrasing was computed word-by-word. We would not want to compute agreement for phrasing on a syllable-by-syllable basis because labelers would always trivially agree that there are not word-internal phrase boundaries, thereby artificially inflating overall agreement. Agreement using the transcriber-pair-syllable / word is computed as the total number of agreements between pairs of coders on every syllable (or word) over the total number of possible agreements between pairs of coders on every syllable (or word). To get raw agreement numbers for each recording, each corresponding transcriber-syllable-pair or transcriber-word-pair is checked. In this first step, all agreements are counted equally. Then a weighted mean of the aggregate agreements for each recording is taken, weighted by the product of the number of transcriber pairs and the number of units of comparison (i.e. words, syllables) in the recording.

Because labels don't occur with equal frequency, guessing based on the frequency of certain labels would allow labelers to produce relatively high agreement. For example, Taylor & Sanders (1995) note that boundaries occur only 20% of the time. Therefore, chance agreement on the annotation of boundaries is actually 80% (and not 50%), and boundary agreement computed using the TPW method described above would be artificially inflated. To correct for cases where chance agreement differs from 50%, the current study employed the Kappa statistic. The standard Kappa statistic is given by the following:

[1] $$\kappa = (Oa - Ea)/(1 - Ea) \quad (1)$$

where Oa is the observed agreement and Ea is the expected agreement by chance, given the statistical distribution of labels in the population. By convention, a kappa statistic of .6 or higher indicates reliable agreement (Landis & Koch, 1977).

A kappa statistic was computed across for each comparison in the following way: First, chance agreement was calculated for each transcriber pair over each recording, based on the relative frequency of labels in the entire corpus. A weighted mean of chance agreements was then computed by averaging chance agreement from each recording, with each measure of chance agreement of each comparison type weighted by the number of labels of that category in the recording. Finally, the kappa statistic for each relation was derived from the weighted mean raw agreement and the weighted mean chance agreement over the entire relation.

Agreement analyses – Metrical prominence

The first class of agreement concerned the presence and type of metrically prominent syllables, and was computed only for RaP transcriptions, because it is only applied to speech in the RaP system. Agreement on the presence of a metrically prominent syllable consisted of a binary distinction. Coders agreed if they both labeled (a) a strong or weak beat, or (b) no beat. Agreement on the type of metrical prominence was based on a ternary distinction: Coders agreed if they both labeled (a) a strong beat, (b) a weak beat, or (c) no beat.

Agreement analyses - Pitch accents

The second class of agreement concerned the presence and type of pitch accents. Agreement on pitch accent presence was based on a binary distinction in both ToBI and RaP. In ToBI, coders agreed if they both labeled (a) some variety of a pitch accent, or (b) no accent; in RaP, coders agreed if they both labeled (a) some variety of a pitch accent, or (b) no accent. We computed two quantifications of pitch accent type agreement for ToBI, and one version for RaP. In ToBI, one quantification of pitch accent agreement type was based on a ternary distinction: coders agreed if they both labeled (a) some variety of a high accent, (b) some variety of low accent, or (c) no pitch accent. The second quantification of ToBI pitch accent agreement involved a 6-way distinction where all accents were treated as distinct, and compared to the lack of an accent. In this analysis, all downstepped accents, except for H+!H*, were collapsed with their non-downstepped counterparts. In RaP, coders agreed if they both labeled (a) some variety of a high pitch accent (H*, !H*), (b) some variety of low pitch accent (L*, !L*), (c) an equal pitch accent (E*), or (d) no pitch accent.

Agreement analyses – Phrasal boundaries

The third class of agreement concerned the presence and strength of phrasal boundaries. Agreement on phrasal boundaries was only computed for word-final syllables and was based on a binary distinction. In ToBI, coders agreed if they both labeled (a) a minor or major phrase boundary, or (b) no boundary; in RaP, coders agreed if they both labeled (a) a minor or major boundary or (b) no boundary. Agreement on the type of phrasal boundary was limited to syllables on which one or more coders had indicated a boundary and was based on a ternary distinction. In ToBI, coders agreed if they both labeled (a) a minor boundary, (b) a major boundary, or (c) no boundary. In RaP, coders agreed if they both labeled (a) a minor boundary, (b) a major boundary, or (c) no boundary.

Results

	TSP/TWP		kappa	
Comparison	**ToBI**	**RaP**	**ToBI**	**RaP**
Presence of beat (RaP only)		89%		0.78
Strength of beat (RaP only)		77%		0.61
Presence of PA	86%	84%	0.69	0.67
Strength of PA	84%	78%	0.66	0.61
Strength of PA: All accents distinct (ToBI only)	75%		0.52	
Presence of phrasal boundary	82%	92%	0.50	0.78
Strength of phrasal boundary	79%	86%	0.47	0.67

Table 2: Study One agreement results

The agreement results for Study One are presented in Table 2. Each agreement class will be explained in turn.

Metrical prominence

The first class of agreement concerns the location and strength of metrical prominences (beats), and applies only to the RaP labels. Agreement on the binary distinction of beat presence was very high, as indicated by a TPS of 89%, and a kappa of .78. Moreover, agreement on the ternary distinction of beat strength was high, indicated by a TPS of 77% and a kappa of .61.

Pitch accents

The second class of agreement concerns the presence and type of pitch accent and applies to labels in both RaP and ToBI. Agreement on the binary distinction of pitch accent presence (present vs. absent) was equivalent for both labeling schemes, indicated by a TPS of 86% and a kappa of .69 for ToBI, and a TPS of 84% and a kappa of .67 for RaP. An examination of the three comparisons of pitch accent type revealed a correlation between the number of pitch accent types being compared, and labeler agreement. Specifically, the ternary accent distinction in ToBI (high, low, absent) resulted in a TPS of 84%, and a kappa of .66. The four-way accent distinction in RaP (high, low, equal, absent) resulted in a TPS of 78% and a kappa of .61. Finally, the six-way accent distinction in ToBI (H*, L*, L+H*, L*+H, H+!H*, absent) resulted in a TPS of 75% and a kappa of .52.

Phrase boundaries

The third class of agreement concerns the location and strength of boundaries in both ToBI and RaP. Agreement on the presence of a phrasal boundary was a binary distinction in both ToBI and RaP. This agreement was moderate for ToBI (TWP = 82%, kappa = .50) and high for RaP (TWP = 92%, kappa = .78). Agreement on boundary strength was, again, moderate for ToBI (TWP = 79%, kappa = .47) and high for RaP (TWP = 86%, kappa =.67)

Discussion

In Study One, five students with no previous annotation experience labeled 36 minutes of speech with ToBI and 19 minutes of speech with RaP. A comparison of the ToBI agreement numbers from the current study to those of previous examinations of ToBI labeler agreement indicates that these students were proficient labelers. For example, Yoon, et al. (2005) reported a kappa of .75 for a binary distinction of the

presence of a pitch accent, compared to the .70 reported here for both ToBI and RaP. Second, they report a kappa of .51 for a ternary distinction of pitch accent type, compared to the .67 kappa that we observed for the same comparison in ToBI.

The results also demonstrate that it is possible to achieve fairly high agreement across many aspects of these two annotation systems. According to standard interpretations of kappa, values over .60 indicate substantial agreement, and virtually all of the agreement numbers obtained from the first study are above .60. In fact, there are only two cases—the six-way accent distinction in ToBI and phrasal boundary comparisons in ToBI—where agreement falls below .60. Even in these cases, agreement is moderate. These data, therefore, indicate that annotators can achieve high agreement in both annotation systems.

First, in the RaP system, labelers exhibited high agreement for the location and strength of metrical prominences. This result indicates that the RaP system is a useful alternative to ToBI in cases where information about speech rhythm is desired. In addition, this result indicates that annotators can successfully identify multiple levels of perceptual prominence using the RaP system.

Second, labelers exhibited high agreement on the presence and type of pitch accent for both systems. Specifically, the kappa for pitch accent presence was identical for both systems, suggesting that the labelers were using the same criteria across both systems to decide on the location of pitch accents.

One of the stated goals of the current study was to ascertain whether coders exhibited higher agreement for prominence in the RaP system, which allows the differentiation of three levels of prominence, than in the ToBI system, which instantiates only two levels of prominence. This claim cannot be answered by simply comparing agreement on pitch accents across both systems, as it may be the case that the same syllables were not labeled as pitch accents in both systems. A later analysis of the data will explore the extent to which labelers agreed on labels of pitch accents across systems, and what the pattern of disagreements looks like.

Finally, labelers were in higher agreement on the presence and type of phrasal boundary in the RaP system than they were for the ToBI system. This finding may be a result of the interference of ToBI's "linkage rules" with true perception. That is, because ToBI demands that labelers annotate a tonal event wherever they annotate a phrasal boundary, and vice versa, labelers may be more apt to label a phrasal boundary where they don't hear one if they have labeled a tonal event, or more likely to not label a phrasal boundary when they don't hear a tonal event. In the RaP system, conversely, there is no interdependency between labels, so labelers can label using solely their perception of both phrasal disjuncture and tonal events. Once again, an investigation of across-system agreement can help to answer this question.

The finding of lower agreement for ToBI than for RaP on the presence and type of phrasal boundary may indicate that RaP is a better system for phrasal labeling, perhaps because it allows for the independence of labels across tiers. However, this result could also be due to the order of labeling systems employed in the current study. That is, because labelers did the majority of labeling with the ToBI system before they trained and tested on the RaP system, their higher agreement in RaP may have simply been due to an overall practice effect, rather than a difference between phrasal labeling conventions of the two systems. We addressed this latter possibility by comparing labeler agreement on the ToBI labels they completed before training and testing on the

RaP system with ToBI labels completed after training and testing on the RaP system. The results of this analysis are presented in Table 3.

	Before RaP		After RaP	
Comparison	TSP/TWP	Kappa	TSP/TWP	Kappa
Presence of PA	86%	0.69	85%	0.68
Strength of PA: High vs. Low	84%	0.66	83%	0.65
Strength of PA: All accents distinct	75%	0.51	73%	0.52
Presence of phrasal boundary	82%	0.48	86%	0.70
Strength of phrasal boundary	80%	0.44	82%	0.64

Table 3: Agreement on ToBI corpus annotated before and after training on RaP

Comparing agreement on boundary labels in the ToBI system both before and after training in the RaP system indicates that agreement on phrasing in the ToBI system is indeed lower on labels completed before RaP training than after. However, it is not simply the case that labeler were in higher agreement on their labels after they trained on RaP. Rather, the difference in the kappa numbers in the two divisions of the ToBI corpus reflects that fact that coders were using a wider variety of labels in the second division of the corpus. Specifically, if the coders were using a larger distribution of labels, then chance agreement decreased, and, therefore, the kappa scores increased.

Another way to ascertain whether or not RaP allows better agreement on phrasal boundaries is to have coders who are experts in both systems label the same speech with both systems, counterbalancing both the order of the presentation of the speech and the order of the labeling systems. We took this approach in a second study of inter-annotator reliability.

Study Two

Study Two was designed to address the limitations of the first study, which were due to constraints on the availability of RaP training materials, and the availability of expert labelers. For this study, we recruited four expert labelers to label a new corpus of speech using both systems. Because all labelers were experts in both systems, we were able to counterbalance the order of speech each coder labeled, as well as the order of annotation. In this way, we could insure that any differences in agreement between systems was not the result of the labelers being more proficient with either system, nor the result of having labeled the same speech previously using another system.

Method

Participants

Four coders who were experts both in ToBI and RaP served as coders in the second study. Two coders were undergraduates who had served as coders in the first study, and continued to receive either course credit or monetary compensation, at a rate of $8.75/hour, for the duration of the project. The other two coders were the author and Laura Dilley.

Materials

The composition of the corpus for Study Two is presented in Table X. As indicated in the table, speech was selected to ensure a balance between spontaneously produced speech and read speech, and between male and female speakers. None of the material used in the second study had been labeled as part of the first study.

FileName	Duration (sec.)	Speaker
Spont 1	50	male
Spont 2	42	male
Spont 3	46	female
Spont 4	44	female
Total	181	
Radio 1	32	male
Radio 2	58	female
Radio 3	29	female
Radio 4	59	male
Total	178	

Table X. Duration (in seconds) and gender of speaker of speech files used in Study Two.

Procedure

Table X lists the order of speech files and coding systems used by each of the four labelers in Study Two. Each labeler's order of speech files and systems was individually determined to counterbalance the order in which the files were labeled in each system.

In order to ensure the highest possible agreement, coders labeled practice speech files both before they began labeling the entire corpus, and before they switched labeling systems. These practice files averaged 30 seconds. Each labeler then received feedback on his/her labels from the Laura Dilley.

Labelers labeled individually, and never discussed their labels at any point during the study.

Labeler 1		Labeler 2		Labeler 3		Labeler 4	
FileName	System	FileName	System	FileName	System	FileName	System
Practice	ToBI	Practice	ToBI	Practice	RaP	Practice	RaP
Spont 1	ToBI	Spont 1	ToBI	Spont 1	RaP	Radio 3	RaP
Radio 1	ToBI	Radio 1	ToBI	Radio 1	RaP	Spont 3	RaP
Radio 2	ToBI	Radio 2	ToBI	Radio 2	RaP	Radio 4	RaP
Spont 2	ToBI	Spont 2	ToBI	Spont 2	RaP	Spont 4	RaP
Practice	RaP	Practice	RaP	Practice	ToBI	Practice	ToBI
Practice	ToBI	Practice	ToBI	Practice	RaP	Practice	RaP
Spont 1	RaP	Spont 1	RaP	Spont 1	ToBI	Radio 3	ToBI
Radio 1	RaP	Radio 1	RaP	Radio 1	ToBI	Spont 3	ToBI
Radio 2	RaP	Radio 2	RaP	Radio 2	ToBI	Radio 4	ToBI
Spont 2	RaP	Spont 2	RaP	Spont 2	ToBI	Spont 4	ToBI
Radio 3	RaP	Practice	ToBI	Radio 3	ToBI	Practice	RaP
Spont 3	RaP	Radio 3	ToBI	Spont 3	ToBI	Spont 1	RaP
Radio 4	RaP	Spont 3	ToBI	Radio 4	ToBI	Radio 1	RaP
Spont 4	RaP	Radio 4	ToBI	Spont 4	ToBI	Radio 2	RaP
Practice	ToBI	Spont 4	ToBI	Practice	RaP	Spont 2	RaP
Radio 3	ToBI	Practice	RaP	Radio 3	RaP	Practice	ToBI
Spont 3	ToBI	Radio 3	RaP	Spont 3	RaP	Spont 1	ToBI
Radio 4	ToBI	Spont 3	RaP	Radio 4	RaP	Radio 1	ToBI

Spont 4	ToBI	Radio 4	RaP	Spont 4	RaP	Radio 2	ToBI
		Spont 4	RaP			Spont 2	ToBI

Table X. File and system labeling order for all four coders in Study Two.

Agreement Analyses in Study Two were calculated in the same way as they were in Study One.

Results

The agreement results for Study Two are presented in Table X. Each agreement class will be explained in turn.

The first class of agreement concerns the location and strength of metrical prominences (beats), and applies only to the RaP labels. As in the first study, agreement on the binary distinction of beat presence was very high, as indicated by a TPS of 89%, and a kappa of .78. Moreover, agreement on the ternary distinction of beat strength was high, indicated by a TPS of 79% and a kappa of .63.

The second class of agreement concerns the presence and type of pitch accent and applies to labels in both RaP and ToBI. Agreement on the binary distinction of pitch accent presence (present vs. absent) was slightly higher for the ToBI system, indicated by a TPS of 89% and a kappa of .76 for ToBI, and a TPS of 85% and a kappa of .66 for RaP. In contrast with the results of the first study, there was no correlation between the number of pitch accent types being compared, and labeler agreement. Specifically, the ternary accent distinction in ToBI (high, low, absent) resulted in a TPS of 86%, and a kappa of .71. The four-way accent distinction in RaP (high, low, equal, absent) resulted in a TPS of 80% and a kappa of .60. Finally, the six-way accent distinction in ToBI (H*, L*, L+H*, L*+H, H+!H*, absent) resulted in a TPS of 78% and a kappa of .58.

	TSP/TWP		**kappa**	
Comparison	**ToBI**	**RaP**	**ToBI**	**RaP**
Presence of beat (RaP only)		89%		0.78
Strength of beat (RaP only)		79%		0.63
Presence of PA	89%	85%	0.76	0.66
Strength of PA	86%	80%	0.71	0.60
Strength of PA: All accents distinct (ToBI only)	78%		0.58	
Presence of phrasal boundary	91%	90%	0.76	0.76
Strength of phrasal boundary	87%	85%	0.68	0.67

Table X. Agreement results from Study Two.

The third class of agreement concerns the location and strength of boundaries in both ToBI and RaP. Agreement on the presence of a phrasal boundary was a binary distinction in both ToBI and RaP. This agreement was similar for ToBI (TPW = 91%, kappa = .76) and RaP (TPW = 90%, kappa = .76). Agreement on the ternary distinction of boundary strength (high, low, none) was also comparable for both systems, as indicated by a TPW of 87% and a kappa of .68 for ToBI and a TPW of 85% and a kappa of .67 for RaP.

Discussion

In a second study of inter-annotator reliability of both the ToBI and RaP annotation systems, we employed four expert labelers to apply each system to the same six minutes of speech in each system. In contrast to the first study, the order of both speech files and annotation systems was counter-balanced across labelers to control for differences in agreement due to better proficiency in one system over the other, or any possible advantage obtained from labeling the same speech twice. Broadly speaking, agreement numbers for the second study were similar to those from the first study. Once again, virtually all kappa values indicated substantial agreement in that they exceeded .60

As in Study One, agreement was very high for the location and strength of metrical prominences in the RaP system. As is Study One, agreement on the presence and tone of pitch accent was high for ToBI and RaP. The advantage we observed in the first study for boundary agreement and strength in the RaP system was not apparent in the second study.

General Discussion

The current study was conducted both to provide a large-scale investigation of the inter-coder reliability of the ToBI system of prosodic annotation, and to ascertain the inter-coder reliability of a new system of prosodic annotation, RaP, which was designed to address the known limitations of the ToBI system. The two annotation systems are intended to capture several different prosodic categories. Specifically, ToBI endeavors to capture the presence and strength of intonational boundaries (i.e. perceptual breaks) and the presence and tonal characteristics of pitch accents, which serve to give certain syllables perceived prominence. RaP aims to capture the same information about boundaries and accents and, in addition, allows for the coding of speech rhythm.

The results of the first study of rater reliability showed high agreement on every tested comparison for both systems among naïve coders for both ToBI and RaP. First, they demonstrated moderate-to-high agreement on presence and tone of phrasal boundary, though agreement was higher on both of these characteristics for the RaP system. As for pitch accent, agreement correlated with number of accent categories; that is, agreement dropped steadily as agreement went from a 2-way distinction to a five-way distinction. Finally, the results demonstrated high agreement in RaP on the coding of metrical prominence.

A further result from the first study was the observation that coders demonstrated higher agreement on the presence and strength of phrasal boundaries in ToBI in the portion of the corpus that they labeled with ToBI *after* they had learned the RaP system. The kappa results suggest that the labelers were using a wider distribution of boundary labels in the annotations they completed after learning RaP.

The results of the second study replicated the pattern of results of the first study; namely, high agreement for both systems for the categories of boundary presence and strength, high agreement for both systems for pitch accent presence and type, which decreased with the number of pitch accent categories, and high agreement for RaP on pitch rhythm. Moreover, the results of the second study, which was done using the annotations of expert coders, suggest that the coders in the first study may have had higher agreement on RaP categories as a result of learning the RaP system after the ToBI system.

Overall, the results of the two studies indicate that coders can achieve high agreement on both systems, and that prosodic annotation can be considered a valuable tool for use in the study of prosody. Furthermore, they demonstrate that RaP is a viable alternative to ToBI for the annotation of large speech corpora, especially when speech rhythm information, or information about multiple levels of prominence, is desired.

Future analyses of the corpus annotated for the current study will be used to explore questions about whether the differences in the labeling inventories of the ToBI and RaP systems impact agreement across the two systems.

Chapter 3

The focus of the current chapter is on understanding how the syntax of a sentence can be used to predict how a speaker will divide that sentence into intonational phrases. An old joke, illustrated in (1), serves as an example of how different intended syntactic structures can be realized with different intonational phrasing:

(1) Woman without her man is nothing.

Woman // without her man // is nothing.

Woman! Without her // man is nothing.

In (1a), the sentence is disambiguated by the insertion of a break after 'man,' such that 'is nothing' must refer to 'Woman.' Conversely, in (1b), the break is placed *before* 'man,' thereby forcing 'is nothing' to modify 'man.' The written breaks correspond to points of disjuncture in the sentence—places where one intonational phrase ends, and another begins—and so are referred to as intonational boundaries. These two prosodic parses are the direct result of two different syntactic structures, corresponding to two different sentence meanings. The goal of the current research is to determine the nature of the relationship between syntactic structure and intonational boundary placement.

The location of an intonational boundary, such as that which would be produced after "man" in (1a), is indicated by the presence of several acoustic characteristics. A highly salient cue to a boundary location is the lengthening of the final syllable preceding a possible boundary location (Wightman, et al., 1992; Selkirk, 1984; Schafer, et al., 2000; Kraljic & Brennan, 2005; Snedeker & Trueswell, 2003, Choi, et al., 2005). Characteristic pitch changes accompany this lengthening. The most common of these changes are a falling tone (as in that which ends a declarative sentence) or a rising tone (as in the default yes-no question contour) (Pierrehumbert, 1980). Other cues, such as silence between words (Wightman, et al., 1992), and the speaker's voice quality (Choi, et al., 2005), can also cue the presence of a boundary, though the systematic contribution of each cue has not been explored empirically.

There is some debate in the literature about what function the production of intonational phrasing serves. Previous studies have shown that speakers often disambiguate attachment ambiguities with prosody (Snedeker & Trueswell, 2003; Schafer, et al., 2000; Kraljic & Brennan, 2005), and that listeners can use such information to correctly interpret syntactically ambiguous structures (Snedeker & Trueswell, 2003; Kjelgaard & Speer, 1999; Carlson, et al., 2001; Speer, et al., 1996; Price, et al., 1991). However, a question remains as to whether speakers produce boundaries as a communicative cue for the benefit of the listener, or, alternatively, as a by-product of speech production constraints. Data supporting the first possibility comes from Snedeker and Trueswell (2003), who found that speakers only prosodically disambiguated prepositional phrase attachments in globally ambiguous sentences like *Tap the frog with the feather* if they (the speakers) were aware of the ambiguity. This result suggests that the mere act of planning the syntactic structure of the sentence for the purposes of producing it did not induce speakers to reflect the syntactic structure in their intonational phrasing, indicating that some boundaries are not produced as a normal by-product of speech planning. In contrast to Snedeker and Trueswell's results, Schafer et al. (2000) showed that speakers produced disambiguating prosody early in a sentence even when the intended meaning was lexically disambiguated later in the sentence (When

the triangle moves the square // *it*... vs. When the triangle moves // the square *will*...). This result suggests that speakers produce some boundary cues even when such cues are not required for the listener to arrive at the correct syntactic parse of the sentence, and therefore, the production of intonational boundaries is a normal part of speech planning.

Additional support for the position that boundaries are a result of production constraints comes from recent work by Kraljic and Brennan (2005), who showed that speakers consistently produced boundary information whether or not such information would enable more effective communication with a listener. In their first experiment, Kraljic and Brennan employed a referential communication task in which speakers produced syntactically globally ambiguous instructions like (2) for listeners, who had to move objects around a display.

(2) Put the dog in the basket on the star.

The referential situation either supported both meanings of the sentence, or only one. The placement of intonational boundaries can disambiguate the appropriate attachment as follows: If "in the basket" is meant to indicate the location of the dog (the modifier attachment), then speakers can do so with the inclusion of a boundary after "basket," as indicated in (2a). If, conversely, "in the basket" is meant to be the location for the movement (the goal attachment), speakers can indicate this meaning with a boundary after "dog" as in (2b).

(2a) Put the dog in the basket // on the star.

(2b) Put the dog // in the basket on the star.

If prosody is produced as a cue for listeners, then speakers should be more likely to produce disambiguating prosody when the situation has not been disambiguated for the listener. Contrary to this hypothesis, Kraljic and Brennan observed that speakers disambiguated their instructions with prosody even when the referential situation had been disambiguated for them, in cases where the listener did not need an additional prosodic cue to correctly interpret the instruction. Kraljic and Brennan conclude that boundary production is a result of the speaker's processing of the specific syntactic structure s/he is producing. This statement leads directly to the question we will be addressing in this paper: What is the relationship between syntactic structure and speakers' production of intonational boundaries?

The Left-Right Boundary Hypothesis.

Over the past several decades, researchers have attempted to use syntactic structure to predict the placement of intonational boundaries. Gee and Grosjean (1983), Cooper and Paccia-Cooper, (1980) and Ferreira (1988) have all presented models that attempt to account for boundary placement in terms of syntactic structure. Roughly speaking, in each of these models, longer constituents, syntactic boundaries, and major syntactic categories like matrix subjects and verbs correspond with a higher probability of an intonational boundary, although the weighting of each of these factors differ across models.

Watson and Gibson (2004) (W&G) tested the predictions of Cooper and Paccia-Cooper's, Gee and Grosjean's and Ferreira's models by having naïve speakers produce a series of sentence structures varying in length and syntactic complexity.

In order to encourage natural, fluent productions, W&G employed an interactive communication task, where speakers were paired with listeners and encouraged to produce normal fluent speech. First, speakers read a target sentence silently. The speakers then produced the target sentences aloud for listeners, knowing that the listeners would have to answer comprehension questions about what the speakers produced. In this way, speakers were aware that they were conveying information to listeners, and so would be more likely to produce the sentences with the cues they would use in normal, interactive, language. In addition, speakers were familiar with the material they were producing, and would be less likely to produce disfluencies, an outcome that W&G intended in their choice of task.

W&G identified several aspects of Gee and Grosjean's, Cooper and Paccia-Cooper's and Ferreira's models which enabled successful boundary placement predictions with respect to their data set. First, the size of the most recently completed constituent was a good predictor of boundary occurrence, regardless of the constituent's position in a syntactic tree of the sentence. Assuming a model whereby boundary production behavior is driven by the needs of a speaker, as supported by Kraljic and Brennan (2005), W&G suggested that this effect is due to a 'refractory period' in which the speaker must recover from the expenditure of resources involved in producing a sentence. A second generalization that W&G observed is that the three models all predict more boundaries to occur before the production of longer constituents. Again, W&G interpreted this effect as a result of speaker needs; specifically, that speakers take more time to plan longer upcoming constituents. This possibility is supported by evidence from Ferreira (1991) that speakers take longer to initiate speaking when a sentence has a longer utterance-initial NP, and from studies by Wheeldon and Lahiri (1997) showing that speaker's latency to begin producing a sentence increases as a function of the number of phonological words in the sentence.

Based on their observations of the pervasiveness of the two preceding phenomena, W&G proposed a model of boundary production based on speaker resources. They hypothesized that speakers place boundaries where they do to facilitate *recovery* and *planning*. Specifically, W&G suggested that speakers use boundaries to recover from the expenditure of resources used in producing previous parts of an utterance, and to plan upcoming parts of an utterance. As such, they calculated the probability of a boundary at a certain point (which they subsequently call the "boundary weight") as a function of the size of the preceding material (Left-hand Size—LHS) and the size of the upcoming material (Right-hand Size—RHS). It should be noted that, although in their 2004 paper W&G refer to the material that precedes a candidate boundary location and the upcoming sentence material as the LHS and RHS, respectively, we will, in this paper, emphasize the relationship of these two components to the speaker's production process, and so will refer to them throughout as Recovery and Planning, respectively.

Size in this model is quantified as the number of phonological phrases in a region. A phonological phrase is defined as a lexical head and all the maximal projections on the head's left-hand side (Nespor and Vogel, 1986). Specifically, a phonological phrase contains a head (noun or verb) and all the material that comes between that head and the preceding one, including function words, pre-nominal adjectives, and pre-verbal adverbs. The following examples, from Bachenko and Fitzpatrick (1990) demonstrates how a sentence is divided into phonological phrases:

a. A British expedition | launched | the first serious attempt.

b. We saw | a sudden light | spring up | among the trees.

W&G chose phonological phrase boundaries as the most likely candidates for intonational boundaries as several researchers had noted that, in fluent utterances, speakers rarely place boundaries within phonological phrases (Nespor & Vogel, 1986; Gee and Grosjean, 1983). This phonological phrase constraint greatly limits possible boundary points within a sentence.

In addition to the size constraint on boundary production defined in terms of phonolgical phrases, W&G hypothesized that boundaries would not occur between heads and their arguments. This constraint was motivated by the work of Selkirk (1984), who proposed the Sense-Unit Condition constraint on boundary production. The Sense-Unit Condition stipulates that constituents within an intonational boundary must together form a sense unit, meaning that each constituent within the intonational phrase must participate in either a modifier-head or argument-head relationship with another constituent in the phrase. An example is given in (4).

(4)

a. The mayor of Boston | was drunk again.

b. The mayor | of Boston was drunk again.

(4a) provides an allowable phrasing of the sentence "The mayor of Boston was drunk again," but (4b) is disallowed by the Sense-Unit Condition because the two phonological phrases "of Boston" and "was drunk again" do not form a sense unit. One does not depend on or modify the other; rather, they both depend on "The mayor."

As a first attempt at quantifying semantic-relatedness according to the Sense-Unit Condition, W& G constrained the Recovery and Planning weights as follows: First, the Recovery weight of a constituent immediately preceding a possible boundary location is zero if the following constituent is a dependent of the preceding constituent; Second, the Planning weight of a constituent following a possible boundary location is zero if that constituent is an argument of the immediately preceding constituent.

W&G operationalized their predictions with the Left/Right Constituent Boundary Hypothesis (LRB), which is stated in (3):

(3) Left Constituent / Right Constituent Boundary (LRB) Hypothesis: Two independent factors are summed to predict the likelihood of an intonational boundary at a phonological phrase boundary in production.

Recovery (LHS): The number of phonological phrases in the largest syntactic constituent that has just been completed. A constituent is completed if it has no rightward dependents.

Planning (RHS): The number of phonological phrases in the largest upcoming syntactic constituent if it is not an argument of the most recently processed constituent.

In more recent work, Watson, Breen, and Gibson (2006) investigated the nature of the semantic-relatedness constraint on boundary production. They found that rather than there being a restriction on the placement of boundaries between heads and their immediately adjacent arguments, a better model fit is obtained if boundaries are disallowed between heads and their immediately adjacent *obligatory* arguments. Specifically, they found that although speakers can be induced to place boundaries between nouns and their non-obligatory arguments if the argument is long (e.g. The

reporter's arrival // *at the crash of the subway* vs. The reporter's arrival *at the crash*), speakers virtually never place boundaries between verbs and their obligatory arguments, even when the argument is long (e.g. The reporter arrived *at the crash of the subway*). The current study does not provide evidence in support of a particular formulation of a semantic-relatedness constraint, as the obligatoriness of arguments is not systematically manipulated in the current study. However, we will assume the most recent published formulation of the LRB (indicated in 4), in which boundaries are disallowed between heads and their adjacent obligatory arguments.

Left Constituent / Right Constituent Boundary (LRB) Hypothesis: Two independent factors are summed to predict the likelihood of an intonational boundary at a phonological phrase boundary in production.

1) Recovery (LHS): The number of phonological phrases in the largest syntactic constituent that has just been completed. A constituent is completed if it has no obligatory rightward dependents.

2) Planning (RHS): The number of phonological phrases in the largest upcoming syntactic constituent if it is not an obligatory argument of the most recently processed constituent.

According to the original definition in (3), the LRB accounted for a significant amount of variance in boundary placement ($r^2 = .74, N = 85, p<.001$). When compared to the predictions made by the three other algorithms, the LRB was able to make similarly accurate predictions, using fewer parameters. It should be noted that W&G's regression analysis is somewhat inflated in two ways: First, it was computed on *all* word boundaries, including places where boundaries virtually *never* occur in fluent speech (e.g. between articles and nouns) (Watson & Gibson, 2004). Indeed, when the same regression was performed only on word boundaries that coincided with phonological phrase boundaries (the locations where the LRB predicts boundaries will occur), the LRB accounted for a diminished, though still significant, amount of the variance in boundary production ($r^2 = .55, N = 41, p<.001$). Second, W&G computed their regression analyses using the average boundary proportion of each of the eight constructions. Here again, the r-squared value is high because averaging across conditions decreases the amount of variance. The regression analyses presented in the current paper will be computed only at phonological phrase boundaries, both on a *trial-by-trial* basis, and on the *average* boundary production across trials of the same condition, to allow comparison with W&G's original results.

Although the LRB performed well, there are two significant problems with W&G's first study, which the current study addresses. First, the LRB is a post-hoc model, tested on the data it was designed to explain. Although W&G did conduct one subsequent test of the LRB (Expt 2 in W&G 2004), this experiment involved only one syntactic construction. Because of this, the current study was designed to evaluate the LRB's predictions on a new set of syntactically varied materials. In addition, non-syntactic factors, such as the interpretation of relative clauses, could be contributing to boundary placement in the original set of items on which the formulation LRB was based. The relative clauses contained in many of the items could be interpreted restrictively or non-restrictively. W&G suggest that, in the absence of context, speakers followed the simpler non-restrictive reading, since this reading does not require the instantiation of a contrast set (as a restrictive relative clause does) (see Grodner, Gibson, & Watson, 2005). They hypothesized that speakers tend to place boundaries before or after non-restrictive

relative clauses because of factors that go beyond Planning and Recovery. Specifically, because non-restrictive relative clauses constitute new or aside information, speakers may be more likely to set them off from the rest of the sentence in their own intonational phrase. Indeed, in the third experiment of their 2004 paper, W&G found that when they established contexts which disambiguated the restrictiveness of a relative clause with prior context, speakers were more likely to produce boundaries before relative clauses with a non-restrictive reading. Given this finding that the introduction of a new clause can induce speakers to produce intonational boundaries for non-syntactic reasons, the stimuli used for the current experiment did not contain relative clauses.

In addition to addressing problems with the original study, the current study was designed to specify the predictions of the Recovery (LHS) component of the LRB. In earlier work, Watson and Gibson considered other versions of the Recovery component, which took into account more than the size of the most recently completed constituent, but settled on the definition in (3) as a reasonable starting point. The current study was designed to systematically compare three formulations of Recovery:

Incremental: A logical first proposal for how to measure left-hand distance is as a measure of the distance back to the last boundary the speaker produced. We will refer to this formulation as the *Incremental* version of recovery. Under this view, intonational boundaries are a result of the *physical* needs of the speaker. The need to breathe or to reset pitch induces the speaker to place boundaries at regular intervals such that the likelihood of a boundary increases with the amount of material a speaker has produced since the last boundary was produced. To test this alternative, we designed items that included length manipulations at two different points in the sentence in order to see whether the placement of a boundary earlier in the sentence influenced the probability of boundary placement later in the sentence.

Integration Distance: Watson and Gibson (2001) first proposed the Recovery component as a measure of integration distance of the upcoming constituent, or, the distance back to the head with which the upcoming constituent was integrating. We will refer to this formulation as the *Integration Distance* version of Recovery. Under this view, speakers place boundaries before words that must be non-locally integrated because such integration is more complex than local integration, as evidenced by work in self-paced reading showing that readers slow down when reading words that must be integrated with a non-local head compared to words that integrate with a local head (Gibson, 1998; Grodner & Gibson, 2005; Gordon, Hendrick, & Johnson, 2001). The complexity of non-local integration as observed in reading could also be more complex in production, and would induce speakers to produce boundaries at such points to recover from production difficulty. We tested this alternative in two ways by designing items in which (1) integration distance was varied while the size of the most recent completed constituent was held constant, or (2) integration distance was held constant while the size of the most recent constituent varied.

Semantic Grouping: The method of computing the Recovery weight that W&G used in their original formulation of the LRB, in which Recovery is computed as the size of the largest most recently completed constituent, will be referred to as the *Semantic Grouping* version of Recovery. Under this view of boundary production, speakers produce words and constituents that rely on one another for meaning in the same intonational group and separate into different intonational words and constituents which don't depend on one another. In order to test this view, we manipulated the length of the most recently

completed constituent to see if the presence of more semantically related material before a possible boundary location led to a greater incidence of boundaries at that location.

Length manipulations designed to test the three versions of the Recovery component and one version of the Planning component of the LRB were obtained by varying the length of three post-verbal constituents in sentences that took the form: Subject Verb Direct Object Indirect Object modifier. A post-verbal direct object was either short (*the chapter)* or long (*the chapter on local history*); an indirect object was either short (*to the students)* or long (*to the student of social science).* Finally, a modifier was either short (*yesterday*) or long (*after the first midterm exam*). These three independent manipulations resulted in eight conditions. An example item is presented in (5).

(5) a. Short direct object, Short indirect object, Short modifier

The professor assigned the chapter to the students yesterday.

b. Long direct object, Short indirect object, Short modifier

The professor assigned the chapter on local history to the students yesterday.

c. Short direct object, Long indirect object, Short modifier

The professor assigned the chapter to the students of social science yesterday.

d. Long direct object, Long indirect object, Short modifier

The professor assigned the chapter on local history to the students of social science yesterday.

e. Short direct object, Short indirect object, Long modifier

The professor assigned the chapter to the students after the first midterm exam.

f. Long direct object, Short indirect object, Long modifier

The professor assigned the chapter on local history to the students after the first midterm exam.

g. Short direct object, Long indirect object, Long modifier

The professor assigned the chapter to the students of social science after the first midterm exam.

h. Long direct object, Long indirect object, Long modifier

The professor assigned the chapter on local history to the students of social science after the first midterm exam.

Predictions

To investigate the three possible quantifications of Recovery, and to evaluate the overall accuracy of the LRB, we compared the LRB's predictions to actual speakers' boundary placement at four phonological phrase boundaries in each sentence. These points, indicated in (6), include: (1) between the subject and the verb, (2) between the verb and the direct object, (3) between the direct object and the indirect object, and (4) between the indirect object and the modifier.

(6) The teacher $|_1$ assigned $|_2$ the chapter (on local history) $|_3$ to the students (of social science) $|_4$ yesterday / after the first midterm exam.

The LRB does not predict any differences in boundary occurrence at the first two sentence positions. Between the subject and the verb, the total LRB weights are comparable because the size of all three versions of the Recovery component (i.e. the length of the subject) is the same in every condition, and the Planning component (i.e. the entire verb phrase) in all conditions is always relatively large (≥ 4 phonological phrases). Therefore, the LRB predicts minimal differences across conditions. Between the verb and the direct object, the LRB weights are the same for the following reasons: First, the size of all three versions of the Recovery component is the same in all conditions (i.e. two phonological phrases corresponding to the subject and the verb); second, in every condition of every item, the direct object is an obligatory argument of the preceding verb, so the Planning component will have a weight of zero because it disallows boundaries between heads and their obligatory arguments.

The three different versions of Recovery make different predictions at both the third and fourth points in the test sentences, and we will elaborate each in turn.

Incremental

In order to test the predictions of an incremental version of Recovery, we included two length manipulations within each sentence. In this way, we could see whether the presence (vs. absence) of a boundary at a location early in the sentence would lead to fewer boundaries at the later location, irrespective of the size of the adjacent constituents at the second location. For example, we hypothesized that if speakers did not place boundaries after a short direct object, as in (5a), they might be more likely to place a boundary after a short indirect object. Although the Semantic Grouping version of Recovery would predict few boundaries between the indirect object and modifier in this example, an Incremental version would predict more simply because the speaker has produced at least two phonological phrases since his/her last boundary.

Integration Distance

The stimuli were created to allow for two critical comparisons to investigate the Integration Distance version of Recovery. The first is between sentences (5b) and (5c) or (5f) and (5g). In (5b), at "yesterday," the size of the most recently completed constituent is one phonological phrase ("to the students") whereas the integration distance back to "assigned"—which "yesterday" must be integrated with—is three phonological phrases ("the chapter / on local history / to the students"). Conversely, in sentence (5c), while the size of the most recently completed constituent has increased to two phonological phrases ("to the students / of social science"), the integration distance back to the verb is still three phonological phrases ("the chapter / to the students / of social science"). Whereas Semantic Grouping predicts more boundaries before the modifying adverb "yesterday" in condition (5c) than in (5b), and in (5g) than in (5f), Integration Distance predicts an equal boundary probability between the two conditions.

The predictions of Integration Distance also differ from those of Semantic Grouping at the phonological phrase boundary between the indirect object and the modifier when the direct object has been either long or short but the indirect object remains constant. Specifically, when the indirect object length is matched, Integration Distance predicts more boundaries when the direct object is long than when the direct object is short, because when the direct object is long, the modifier has a longer distance to integrate back to the verb than when the direct object is short. In contrast, Semantic Grouping

predicts an equal probability of boundary production in either case, as the most recently completed constituent (the indirect object) is the same length.

Semantic Grouping

The Semantic Grouping version of Recovery predicts main effects of length at the two testing points. Between the direct object and the indirect object—"the chapter (on local history)" / "to the students (of social science)"—Semantic Grouping predicts a main effect of direct object length such that speakers will place more boundaries after long direct objects than short direct objects. *Semantic Grouping* also predicts a main effect of indirect object length such that speakers will place more boundaries before long indirect objects than short indirect objects. Between the indirect object and modifier, *Semantic Grouping* predicts main effects of indirect object length, such that speakers will place more boundaries after long indirect objects than short indirect objects. It also predicts a main effect of modifier length such that speakers will place more boundaries before long modifiers than short modifiers.

Method

Participants

Forty-eight native English speakers from the MIT community participated in the study for $10.00 each. Participants were run in pairs, and each member of the pair was randomly assigned to either the role of Speaker or the role of Listener. Data from three of the twenty-four pairs of subjects could not be used due to poor recording quality. Productions from eighteen of the remaining twenty-one pairs were coded for intonational boundaries by two blind coders. Productions from all twenty-one successfully recorded subjects were analyzed for word duration and silence data.

Materials and Design

Length of direct object, length of indirect object, and length of modifier were manipulated in a 2x2x2 design to create thirty-two stimulus sets like those in (5).

The long direct object and long indirect object conditions were created by adding a modifier phrase (or non-obligatory argument phrase) to the object noun phrase (e.g. the bouquet *of thirty roses*, the turkeys *with homemade stuffing*, the chapter *on local history*). All direct objects and indirect objects in the short condition had three syllables, while the long conditions had seven or eight syllables. The short modifiers were temporal modifiers (in 23 items) or adverbs (nine items) comprised of one or two words (two to four syllables), but always only one phonological phrase (e.g. *secretly, last night, on Sunday*). The long modifiers were always temporal modifiers containing five words, which were comprised of 3-4 phonological phrases.

Twelve of the 32 experimental items contained an information structure manipulation in the long modifier condition. Specifically, the long modifier in these items contained a new clause (e.g. *after filming had already begun, before the crime was committed*). In order to discover whether speakers place more boundaries before the introduction of a new clause, we computed analyses of variance on two subsets of items: those with new clauses in the modifier and those without. The analyses were virtually identical to the analyses presented below, which were computed across all thirty-two items. Thus, the presence/absence of a clause boundary does not seem to have affected

the probability of the productions of an intonational boundary. Any divergence across the three analyses will be noted.

The materials were presented in a Latin Square design, resulting in eight lists. Each participant saw only one of the lists, which was presented in a random order. Experimental items were randomly interspersed with 44 fillers, which were comprised of items from two other unrelated experiments, with different syntactic structures. A full set of experimental items can be found in the Appendix.

Procedure

The experiment was conducted using Linger, a software platform for language processing experiments.[3] Two participants—a speaker and a listener—were included in each trial, and sat at computers in the same room such that neither could see the other's screen. The speakers were instructed that they would be producing sentences for their partners (the listeners), and that the listeners would be required to answer a comprehension question about each sentence immediately after it was produced. Each trial began with the speaker being presented with a sentence on the computer screen to read silently until s/he understood it. The speaker then answered a multiple-choice content question about the sentence, to ensure understanding. If the speaker answered correctly, s/he proceeded to produce the sentence out loud once. If the speaker answered incorrectly, s/he was given another chance to read the sentence, and to answer a different question about it. The speaker always produced the sentence after the second question whether or not s/he got the second question right.

The listener sat at another computer, and saw a blank screen while the speaker went through the procedure described above for each sentence. After the speaker produced a sentence out loud for the listener, the listener would press the space bar on his/her computer, whereupon s/he was presented with a multiple-choice question about the content of the sentence that was just produced. Listeners were provided feedback when they answered a question incorrectly.

Trials where one or both of two blind coders identified a disfluency in the production were excluded from analysis, following the method of W&G, accounting for 4.6% of the data. Trials where either a) the speaker answered both comprehension questions incorrectly or b) the listener answered his/her comprehension question incorrectly accounted for 3.2% of the data. We conducted all analyses reported below on a) all trials and b) only trials without any incorrect responses, and the results of both analyses were not different. Therefore, we report the results from all fluent trials below.

Boundary identification

Each sentence was recorded digitally, and analyzed using the PRAAT program (Boersma & Weenink, 2006). Each production was coded by two expert coders (neither of whom was an author) for intonational boundaries using a subset of the ToBI intonational labeling conventions (Silverman et al., 1992). Both coders were blind to the predictions of the experiment. The strength of a boundary was marked by each of the coders using the following standard break indices and disfluency notation: 4 – full intonational phrase boundary (IP); 3 – intermediate phrase boundary (ip); 0, 1, 2 – no phrase boundary; P – hesitation pause; D – disfluency. Most of the non-phrasal boundaries were coded as "1". We therefore collapsed all of 0, 1, and 2 as the category 1. The raw numerical labels (i.e.

[3] Linger was written by Doug Rohde, and can be downloaded at: http://tedlab.mit.edu/~dr/Linger/

1, 3, 4) were grouped in two different ways for two separate analyses reported below. Trials which contained hesitations or disfluencies were excluded from analysis.

Boundary identification is not a straightforward, objective task. Although several acoustic measures, such as silence and lengthening, have been found to correlate with raters' identification of intonational boundaries (see Wightman, et al., 1992), even expert coders are not in perfect agreement about the presence of boundaries (Pitrelli, et al., 1994; Dilley, et al, 2006; Syrdal & McGory, 2000 Yoon, et al., 2004). We therefore computed the inter-rater reliability of the two coders. Furthermore, we computed the correlation of the coders' identification of boundaries with both the duration of the pre-boundary word and any silence that accompanied the boundary location.

Each trial was annotated for word boundaries and silence by one of three coders, none of whom were authors on the current paper, or ToBI coders for the present study. These coders were blind to the hypotheses of the study. Using output from the PRAAT program, the coders annotated all word boundaries, as well as any perceptual silence that was distinct from silence due to stop closure.

Inter-rater reliability

Reliability between ToBI coders was measured by calculating the proportion of the instances in which the two transcribers agreed on the label of a word boundary over the number of possible agreements, as described in Pitrelli et al. (1994). To avoid artificial inflation of the agreement numbers, we excluded the final word in each sentence from analysis, as these words, being utterance final, always received an obligatory break index label of '4' from both coders, and therefore contributed a trivial point of agreement to the analyses. As such, we computed, for example, nine points of agreement for a sentence composed of ten words.

We computed agreement between the coders in two different ways. First, we compared the coders' raw break indices (BIs) (1, 3, 4), resulting in a total agreement of 78.8%. Second, we computed agreement when we compared ips (BI of 3) or IPs (BI of 4) boundaries to the absence of a boundary (BI of 1), which resulted in an overall agreement measure of 82.0% between the two coders. Finally, we computed agreement when both raters indicated an IP (BI of '4'). This calculation resulted in overall agreement of 94.9%.

All of the above agreement numbers are consistent with previous measures of boundary agreement in ToBI (e.g. Pitrelli, et al., 1994; Dilley, et al., 2006). However, in order to effectively use the data we collected from two coders, the two measures of boundary proportion that we will present in the results section come from the average of the two coders labels. We will present one set of analyses which we conducted on the average of the coder's boundary decisions where only IPs (i.e. Break indices of '4') are considered to be boundaries. The second set of analyses is based on the coders' data where IPs and ips (i.e. Break indices of '4' or '3') are considered to be boundaries. In both cases, the binary distinction of each individual coder (of boundary vs. not boundary) was expanded into a ternary distinction where a phonological phrase could be (a) 0: a non-boundary as indicated by both coders, (b) .5: a boundary for one of the coders and not the other, or (c) 1: a boundary indicated by both coders.

Acoustic correlates of boundaries

To further test the accuracy of the coders' labels of intonational boundaries, we correlated several acoustic measures with the coders' ToBI labels. We gathered measures of (a) the duration of each word that preceded a phonological phrase boundary, (b) the duration of any silence that followed a phonological phrase boundary, and (c) the sum of each of these two measures. We then correlated each of these three acoustic variables with the two formulations of boundary labels described above; first, where only IPs are considered to be boundaries; second, where IPs and ips are considered to be boundaries.

When we considered only IPs as boundaries, we observed a correlation between pre-boundary word duration and average coder boundary label ($r^2 = .091$, $N = 2636$, $p < .001$), indicating that boundaries were more likely to be identified after lengthened words. We also observed a correlation between post-boundary silence and boundary label ($r^2 = .318$, $N = 2636$, $p < .001$), such that measurable silence was more likely to occur when coders indicated a boundary. Finally, we observed a correlation between the coders' average boundary label and the sum of the word duration and post-word silence ($r^2 = .234$, $N = 2636$, $p < .001$). Because the correlation between the presence of silence and the labeling of an IP was the highest of these three analyses, we used the presence of silence as the dependent measure in a series of ANOVAs conducted at each critical constituent.

When we considered IPs or ips to be boundaries, we observed a correlation between pre-boundary word duration and average coder boundary label ($r^2 = .153$, $N = 2625$, $p < .001$), indicating that boundaries were more likely to be identified after lengthened words. This correlation suggests that word duration may be a better cue for intermediate boundaries than for full intonational boundaries as this correlation is better than the one performed on IPs above. We also observed a correlation between post-boundary silence and boundary label ($r^2 = .097$, $N = 2625$, $p < .001$), such that measurable silence was more likely to occur when coders indicated a boundary. This correlation suggests that silence is a stronger cue to full intonational boundaries (IPs) than to intermediate boundaries (ips) as this correlation is lower than the one performed on IPs above. Finally, we observed a correlation between the coders' average boundary label and the sum of the word duration and post-word silence ($r^2 = .205$, $N = 2625$, $p < .001$). Because the combination of word duration and post-word silence proved to correlate better with coders' labels of '3' or '4' than either cue alone, we used this measure in a different series of ANOVAs conducted at each critical constituent boundary.

The above correlations are noteworthy for the following reason: Because the experiment was conducted using a fully between-subject Latin-square design, each subject produced only one token of each word used in this analysis. Moreover, though the actual words being compared were all two-syllable words with initial stress, they differed greatly in terms of segmental characteristics. Individual speaker and word variation add a large amount of noise to this analysis. However, despite the variability, we still observed a strong relationship between boundary perception and acoustic measures. The fact that the acoustic measures correlated with the perceived boundaries suggests that the coders may have been using these and other acoustic cues to code boundary locations.

Data analysis

We will present the results of the experiment in several different ways to investigate the effectiveness of the three versions of the Recovery and the one version of Planning which we have proposed to account for boundary placement. First, we present the results of a series of ANOVAs, where the dependent measure corresponds to four different quantifications of boundaries, to compare the specific predictions of the models at the four critical points indicated in (6). Second, we present results of a series of regression models which test the success of predictions across all the phonological phrase boundaries in each sentence, as indicated in (7).

(7) The teacher $|_1$ assigned $|_2$ the chapter $|_3$ on local history $|_4$ to the students $|_5$ of social science $|_6$ yesterday.

We will first present the results of analyses where 3's or 4's are considered to be boundaries, following W&G. Using these criteria, boundaries were indicated by both coders 33.1% of the time, and by one of the two coders 48.1% of the time. Along with these analyses, we will present the results of ANOVAs conducted with the sum of word duration and following silence as the dependent measure, as this acoustic measure correlated most highly with combined '3' and '4' labels.

We will next present the results of the analyses where only 4's are considered to be boundaries. Using these criteria, boundaries were indicated by both coders 6.1% of the time, and by one of the two coders 14.2% of the time. Along with these analyses, we will present the results of analyses conducted on post-boundary silence, as this measure correlated most highly with coders' labels of '4.'

In addition to the analyses of variance performed at each critical phonological phrase boundary in the sentences, we also performed a series of regression analyses, to see which version of the LRB model accounted for the most variance in the speakers' productions. We will present regression analyses performed on the *means* of boundary productions at each phonological phrase boundary across all of the sentences, and analyses performed on a *trial-by-trial basis*.

It should be noted that the regressions testing the Incremental version of recovery differ from those testing Semantic Grouping and Integration in three important ways.

First, regressions using the latter two versions of Recovery can easily be performed on the sentence means, as the predictions of these models do not change across trials; In contrast, the predictions of the Incremental version *do* change across trials as they are, by definition, contingent upon what has previously happened in the sentence. In order to perform regression analyses on means in this case, we computed the average distance (in phonological phrases) back to the location of the last boundary using the following method. Consider (8), where $P_1 \ldots P_n$ indicate phonological phrases and b_i indicates the proportion of boundaries between P_i and P_{i+1}:

(8) $P_1 \mid P_2 \mid P_3 \mid P_4 \ldots$
$\quad\quad b_1 \quad b_2 \quad b_3 \ldots$

At the first possible boundary location (between P_1 and P_2), the average distance back to the last boundary is always one phonological phrase (i.e. the distance back to the beginning of the sentence). At the second possible boundary location (between P_2 and P_3), the average distance back is $[1 + (1\text{-}b_1)]$. At the third candidate boundary location (between P_3 and P_4), the average distance back is $1 + (1\text{-}b_2) + (1\text{-} (1\text{-} b_2) * (1\text{-} b_1))$. The proportion of boundaries at later locations is computed similarly.

Second, analyses on the Semantic Grouping and Integration Distance versions of Recovery can easily be performed on the average of the two coders' data. However, because the incremental predictions are contingent on the labels of a particular coder, they cannot be averaged. Therefore, we performed regressions on the two coders' data individually in order to test the Incremental version of Recovery.

Finally, it is not possible to use the Incremental version of Recovery to predict the acoustic measures of word duration and silence. The Incremental version of Recovery's predictions about where boundaries will occur is based on the location of previous boundaries. Without a dichotomous way of using duration and silence to define the location of boundaries, there is no way of assessing where the previous boundary has been placed. Moreover, even if there were a way of defining the location of the last boundary, it is not clear how increased distance from that boundary would translate into the continuous variables of word duration and silence.

Results

Boundaries as Intermediate or Full Intonational Phrases: ANOVAs

We performed a series of 3 x 2 analyses of variance with Participants or Items as the random factor at the sentence positions indicated in (6), which were a) after the Subject NP, b) after the Verb, c) after the direct object, and d) after the indirect object. The dependent variables were the averaged ratings of the two coders when IPs or ips were considered boundaries and the corresponding acoustic measure of the duration of the pre-boundary word and following silence. The independent variables in each analysis were a) the length of the direct object (short, long), b) the length of the indirect object (short, long), and c) the length of the modifier (short, long). In all cases, analyses were conducted only on fluent trials.

Subject NP:

A 3 x 2 ANOVA with boundary percentage after the Subject NP as the dependent measure revealed no effect of direct object length on boundary placement, $F1(1,17) < 1$, $F2(1,31)=1.752$, $p=.195$. There was, however, a marginal main effect of length, $F1(1,17)=4.99$, $p<.05$, $F2(1,31)=3.39$, $p=.075$, such that more boundaries occurred after the subject NP when the upcoming indirect object was long than when it was short. There was no effect of modifier length, $F1<1$; $F2(1,31) = 2.17$, $p=.15$. There were no effects of direct object length, indirect object length, or modifier length on the duration of the phrase-final word and following silence (4 of 6 F's < 1; all p's > .14).

Verb:

A 3 x 2 ANOVA revealed no effects of direct object length, indirect object length, or modifier length on the on the probability of boundary placement after the verb, as quantified by ToBI label (4 of 6 F's <1; all p's > .16), or by the sum of phrase-final word duration and silence (4 of 6 F's <1; all p's >.24)

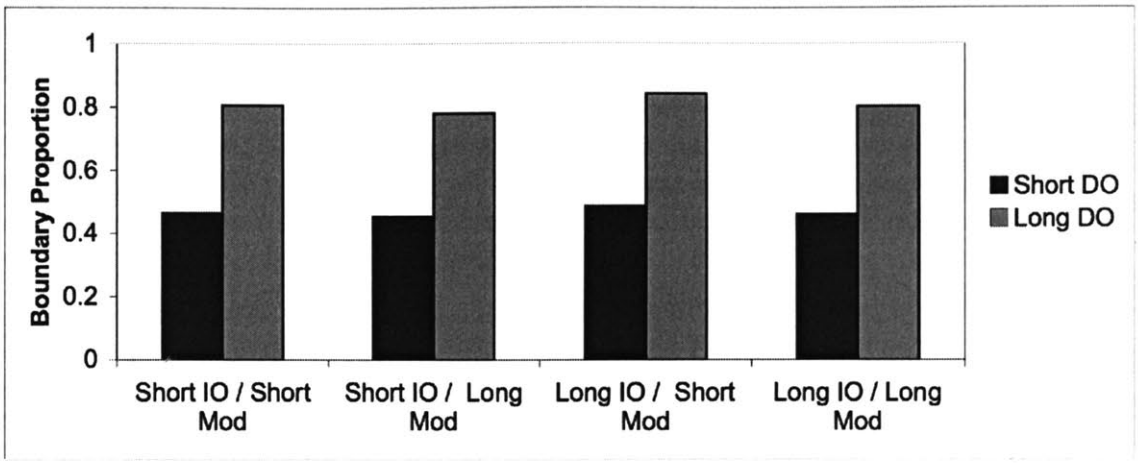


Figure 1: Proportion of boundaries placed after the Direct Object NP in all conditions.

Direct object:

The percentages of boundary placement between the direct object and indirect object, as determined by coders' ToBI labels, are presented in Figure 1. There was a main effect of direct object length such that boundaries occurred more often after long direct objects than short direct objects $F1(1,17)=76.17$, $p<.001$, $F2(1,31)=132.95$, $p<.001$. An effect of indirect object length such that speakers placed more boundaries before long indirect objects than short indirect objects approached significance in the participants' analysis ($F1(1,17)=2.76$, $p=.11$), but not in the items' analysis, $F2(1,31) < 1$. There was also no effect of modifier length in this position, $F's<1$.

In the acoustic analyses, we also observed a main effect of direct object length such that the final word of a long direct object was longer and followed by longer silence (61ms) than a short direct object (55ms), $F1(1,20)=14.35$, $p<.001$, $F2(1,31)=4.33$, $p<.05$. There were no effects of indirect object length or modifier length at this position (All $F's < 2$; all $p's > .20$).

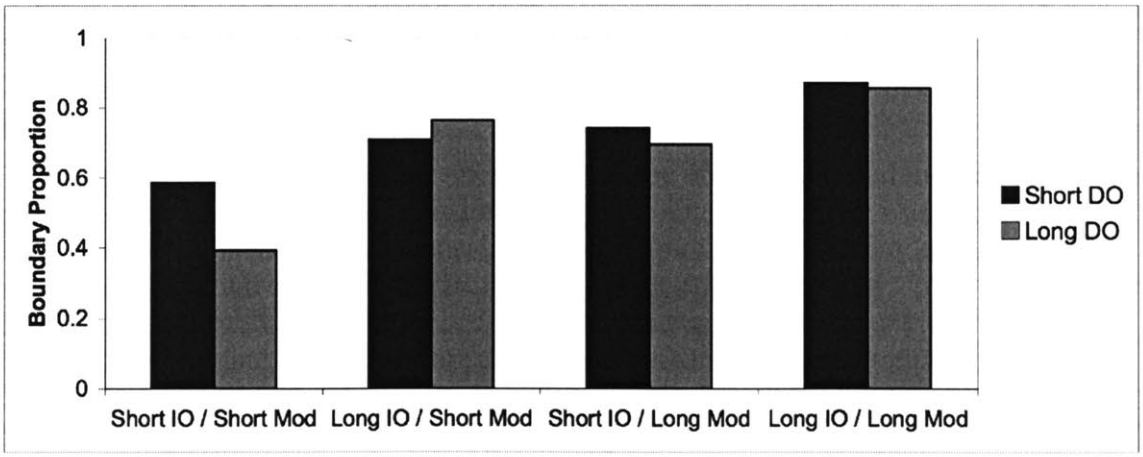


Figure 2: Proportion of boundaries placed after the indirect object NP in all conditions.

Indirect object:

The percentages of boundary placement between the indirect object and the modifier are presented in Figure 2. There was a main effect of indirect object length such that more boundaries occurred following long indirect objects than short indirect objects, $F1(1,17) = 63.09$, $p<.001$; $F2(1,31) = 38.49$, $p<.001$. There was also a main effect of modifier length such that more boundaries preceded long modifiers than short modifiers, $F1(1,17) = 48.23$, $p<.001$; $F2(1,31) = 34.74$, $p<.001$.

In the acoustic analyses, we observed a main effect of indirect object length such that the combined duration of the final word of a long indirect object and any following silence (56ms) was longer than that of a short indirect object (51ms), $F1(1,20)=14.35$, $p<.001$, $F2(1,31)=4.33$, $p<.05$. We also observed a main effect of modifier length such that the combined duration of the final word of the indirect object and any following silence was longer when the modifier was long (56ms) than when the modifier was short (51ms), $F1(1,20) = 2.82$, $p<.05$, $F2(1,31) = 7.29$, $p<.05$.

We performed two critical comparisons at the indirect object/modifier boundary to test the predictions of the Integration version of Recovery. First, we compared the proportion of intonational boundaries between the indirect object and the modifier when the distance back to the verb (i.e. integration distance) was matched. This condition was satisfied in cases where the direct object was long and the indirect object was short, or when the direct object was short and the indirect object was long. An independent samples t-test comparing these two conditions was conducted on the coders' data revealed a significant effect of indirect object length on the boundary labels $t(259) = -5.94$, $p<.001$, and on the acoustic correlates (word duration and silence), $t(259) = -3.06$, $p<.005$), such that speakers placed more boundaries between indirect object and modifier when the indirect object was long, regardless of the integration distance back to the verb. This result is in contrast to the predictions of the Integration version of Recovery.

The second test of Integration Distance concerned cases where the size of the most recently completed constituent stayed the same, while the integration distance back to the verb varied. This condition is satisfied in when the indirect object is the same length, but the direct object is varied. Independent samples t-tests conducted on the coders' ratings and the acoustic measures revealed no difference between boundary placement after a long indirect object regardless of the length of the direct object (all p's>.4), indicating that the Integration Distance version of Recovery did not account for boundary placement after the indirect object phrase.

In accordance with the predictions of the Incremental version of Recovery, we observed a significant interaction between direct object length and indirect object length such that more boundaries preceded the modifier in the short direct object / short indirect object condition than in the long direct object/short indirect object condition $F1(1,17)=6.40$, $p<.05$; $F2(1,31)=9.63$, $p<.005$. In addition, we observed a significant 3-way interaction between direct object length, indirect object length and modifier length ($F1=5.46$, $p<.05$; $F2=4.31$, $p<.05$), such that in cases where the modifier was long, speakers overwhelmingly placed a boundary before it, regardless of what had preceded that point. This interaction only reached significance in the participants' analysis of items that did not introduce new clauses in the modifier phrase, $F1\ (1,17) =5.57$, $p<.05$; $F2\ (1,19) =2.41$, $p=.14$).

Boundaries as Full Intonational Phrases: ANOVA

We performed a series of 3 x 2 analyses of variance with Participants or Items as the random factor at the sentence positions indicated in (6), which were a) after the Subject NP, b) after the Verb, c) after the direct object, and d) after the indirect object. The dependent variables were the averaged ratings of the two coders when only IPs were considered boundaries, and the corresponding acoustic measure of the duration of post-word silence. The independent variables in each analysis were a) the length of the direct object (short, long), b) the length of the indirect object (short, long), and c) the length of the modifier (short, long). In all cases, analyses were conducted only on fluent trials.

Subject NP

A 3 x 2 ANOVA with boundary as the dependent measure revealed no effect of direct object length (F's < 1), no effect of indirect object length (F's < 1), and no effect of MOD length (F1 (1, 31) =2.09, p = .17; F2 (1,17) = 1.88, p = .18). A 3 x 2 ANOVA with duration of silence as the dependent measure revealed no effect of direct object length (F's <1), and no effect of indirect object (F1<1, F2=1.24, p=.27), but a marginal effect of modifier length, such that longer silence followed the Subject NP when the sentence contained a long modifier (20ms) than when it contained a short modifier (10ms), F1(1,20) =4.66, p<.05 F2(1, 31) = 3.324, p = .08.

Verb

A 3 x 2 ANOVA revealed no effects of direct object length, indirect object length, or modifier length on the on the probability of boundary placement after the verb, as quantified by ToBI label (All F's <2; all p's > .25), or by post-word silence (4 of 6 F's < 2; all p's > .12).

Direct Object

The proportion of boundaries that occurred between the direct object and the indirect object are plotted in Figure 3. A 3 x 2 ANOVA with boundary as the dependent measure revealed a main effect of direct object length (F1(1,17) = 34.00, p < .001; F2(1,31) = 56.60, p < .001), such that more boundaries occurred after long direct objects (24.6%) than short direct objects (7.2%). The analysis also revealed a main effect of indirect object length (F1(1,17) = 11.42, p <.005; F2(1,31) = 4.92, p < .05), such that speakers placed more boundaries before long indirect objects (18.6%) that before short indirect objects (13.2%). Finally, there was a suggestion of an effect of modifier in the participants analysis, in the opposite of the predicted direction (F1(1,17) = 5.09, p < .05), but this effect was not reliable in the items analysis (F2(1,31) = 1.69, p = .21).

A 3 x 2 ANOVA with duration of silence as the dependent measure revealed a main effect of direct object length (F1=6.75, p<.05; F2=4.30, p<.05) such that longer silences followed long direct objects (45ms) than short direct objects (21ms). There was no effect of indirect object length (F1=2.05, p=.17; F2=1.88, p=.18), and no effect of modifier length, F's<1.

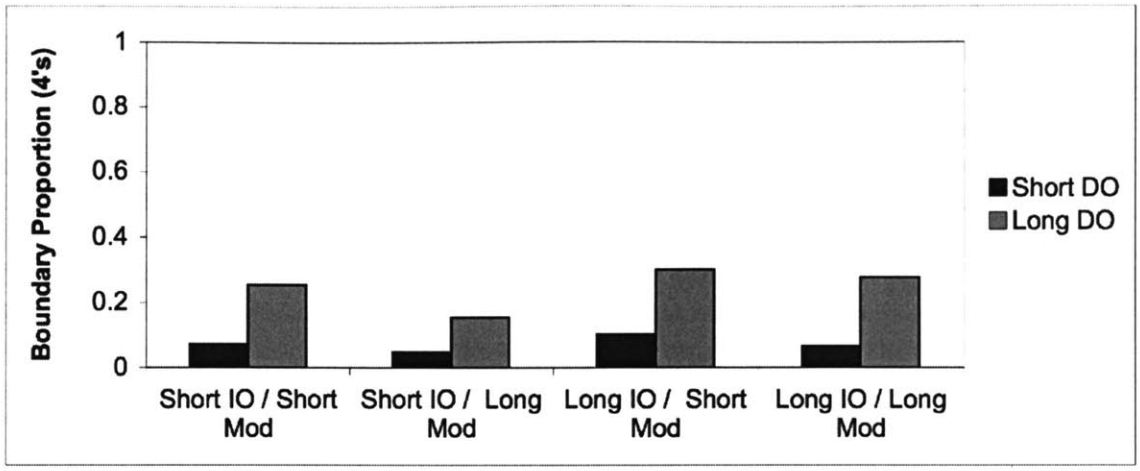


Figure 3: Proportion of boundaries (as determined by average of coder's '4' labels) between the direct object and the indirect object in all conditions.

Indirect Object

The proportion of boundaries that occurred between the indirect object and modifier are plotted in Figure 4. A 3 x 2 ANOVA with boundary as the dependent measure revealed no main effect of direct object length (F's < 1). The analysis did reveal a main effect of indirect object length ($F1(1,17) = 5.75$, $p<.05$; $F2(1,31) = 12.00$, $p < .005$), such that speakers placed more boundaries after long indirect objects (22.6%) than before short indirect objects (14.0%). Finally, this analysis revealed a main effect of modifier length ($F1(1,17) = 39.42$, $p < .001$; $F2(1,31) = 39.23$, $p < .001$), such that speakers placed more boundaries before long modifiers (28.9%) than before short modifiers (8.3%).

A 3 x 2 ANOVA with duration of silence as the dependent measure revealed no effect of direct object (F's<1), and no effect of indirect object ($F1=2.81$, $p=.11$; $F2=1.63$, $p=.21$); however, this analysis did reveal a main effect of modifier length ($F1=15.95$, $p<.005$; $F2=12.83$, $p<.005$), such that longer silences occurred before long modifiers (42ms) than before short modifiers (12ms).

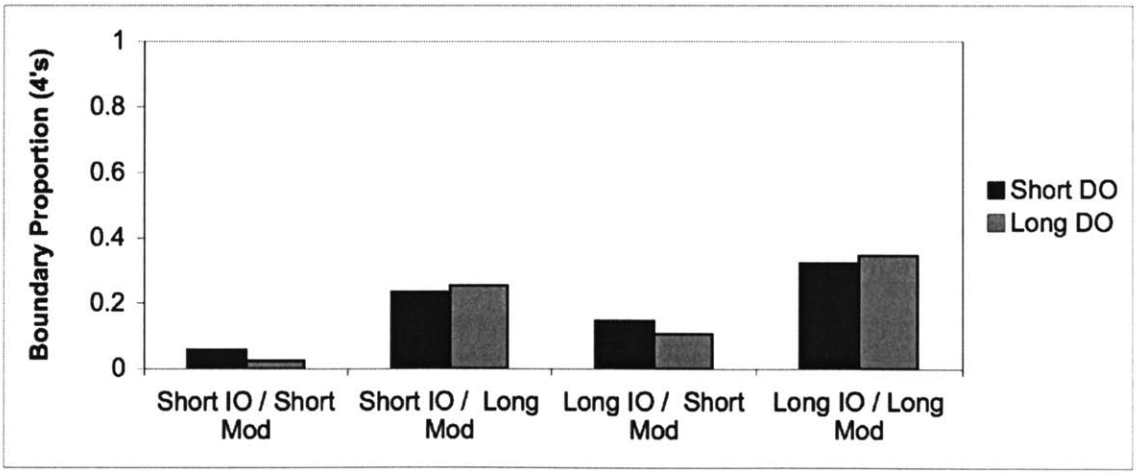


Figure 4: Proportion of boundaries (as determined by average of coder's '4' labels) between the indirect object and the modifier in all conditions.

We performed two critical comparisons at the indirect object/modifier boundary to test the predictions of the Integration version of Recovery. First, we compared the proportion of intonational boundaries between the indirect object and the modifier when the distance back to the verb (i.e. integration distance) was matched. This condition was satisfied in cases where the direct object was long and the indirect object was short, or when the direct object was short and the indirect object was long. An independent samples t-test comparing these two conditions was conducted on both the coders' boundary data. This test revealed a significant effect of indirect object length ($t(260) = -2.32$, $p<.05$) such that speakers placed more boundaries between the indirect object and the modifier when the indirect object was long (22.9%), than when the indirect object was short (13.7%), regardless of the integration distance back to the verb. This result is in contrast to the predictions of the Integration version of Recovery. In the analysis of post-word silence, this effect was not significant, $t(308) = -1.20$, $p=.23$.

The second test of Integration Distance concerned cases where the size of the most recently completed constituent stayed the same, while the integration distance back to the verb varied. This condition is satisfied in cases where the indirect object is the same length, but the direct object is varied, a condition which was satisfied in two places in this current experiment. When the indirect object is long, speakers place boundaries after the indirect object 22.9% of the time when the direct object is short and 22.4% of the time when the direct object is long. When the indirect object is short, speakers place boundaries after the indirect object 14.2% of the time when the direct object is short and 13.7% of the time when the direct object is long. Two independent samples t-tests conducted on the coders' boundary ratings revealed no difference between boundary placement after a long indirect object regardless of the length of the direct object ($t_1(261) = .128$, $p = .898$; $t_2(263) = .125$, $p = .900$), indicating that speakers produced comparable numbers of boundaries when the size of the largest recently-completed constituent was the same, regardless of the Integration Distance back to the verb. The pattern of results was the same for the analysis of post-word silence, ($t_1(306) = .024$, $p = .980$; $t_2(312) = .125$, $p = .717$). Once again, these results support the Semantic Grouping version of Recovery over the Integration version of Recovery.

Finally, the Incremental version of Recovery predicted a two-way interaction between the length of the direct object and the length of the indirect object, as pictured in Figure 1, such that speakers would place more boundaries after a short direct object and a short indirect object than after a long direct object a short indirect object. However, this interaction did not approach significance in the analysis of boundaries ($F's <1$) or in the analysis of post-word silence ($F's<1$), failing to show support for the Incremental version of Recovery.

Boundaries as Intermediate or Full Intonational Phrases: Regressions

A series of multiple regression analyses was conducted to predict the average coders' label of a '3' or '4' from the three versions of Recovery—Incremental (2 versions), Integration Distance, Semantic Grouping—and the single version of Planning. The regressions were performed on a trial-by-trial basis, and on the condition means. A summary of the regression models tested is presented in Table 1.

As is evident from the table, every model tested accounted for a significant amount of the variance in the speakers' production of boundaries. However, the best model in both groupings of the data is the one using the Semantic Grouping version of Recovery. For example, when tested on all of the production data, trial-by-trial, this

model accounted for more variance than any of the other three models, $R^2 = .44$, $F(2,2325) = 898.86$, $p < .001$. In addition, in both models which include the Semantic Grouping version of Recovery, both the left and the right factors accounted for a significant amount of the variance, indicating that an increase in a) the size of the largest recently-completed left-hand constituent or b) the right-hand constituent increase the probability that a speaker will produce a boundary at a candidate phonological phrase boundary. The model using the Integration version of Recovery, tested on all of the data, trial-by-trial, also accounts for a significant amount of the variance in speakers' boundary production, $R^2 = .37$, $F(2,2325) = 685.04$, $p < .001$). It predicts that an increase in a) the distance back to the head with which an upcoming constituent must be integrated or b) the size of the upcoming constituent will increase the probability of a boundary at a candidate phonological phrase boundary. Although significant, this model does not account for as much variance as the Semantic Grouping version. Finally, the models using the Incremental version of Recovery, based on each of the speakers' data, both account for a significant amount of the variance in the individual coders' production data. However, these models are less successful than the model using the Semantic Grouping version of Recovery. In several cases, the Incremental factor is negatively correlated with boundary production, indicating that an increase in the distance back to the last boundary leads to a *decrease* in the probability that a speaker will produce a boundary at a candidate phonological phrase boundary.

		Beta	SE	p-value	R^2	F and p-value
Trial-by-trial	Coder 2 Incremental Left	0.08	0.01	0.001	0.07	F(2,2490) = 90.88, p<.001
	Planning Right	0.06	0.01	0.001		
	Coder 1 Incremental Left	0.03	0.01	0.017	0.14	F(2,2330) = 196.74, p<.001
	Planning Right	0.13	0.01	0.001		
	Integration Left	0.20	0.01	0.001	0.37	F(2,2325) = 685.04, p<.001
	Planning Right	0.04	0.05	0.001		
	Semantic Grouping Left	0.38	0.01	0.001	0.44	F(2,2325) = 898.86, p<.001
	Planning Right	0.02	0.00	0.001		
Condition Means	Coder 2 Incremental Left	0.22	0.03	0.001	0.66	F(2,37) = 35.54, p<.001
	Planning Right	0.09	0.02	0.001		
	Coder 1 Incremental Left	0.32	0.09	0.001	0.51	F(2,37) = 18.91, p<.001
	Planning Right	0.16	0.03	0.001		
	Integration Left	0.19	0.02	0.001	0.77	F(2,37) = 62.21, p<.001
	Planning Right	0.04	0.02	0.008		
	Semantic Grouping Left	0.38	0.02	0.001	0.91	F(2,37) = 193.43,p<.001
	Planning Right	0.02	0.01	0.041		

Table 1: Summary of regression models tested on data where 3's or 4's are boundaries. Models were tested on all fluent trials on a trial-by-trial basis, and on the means of each condition.

Boundaries as Full Intonational Phrases: Regression

A series of multiple regression analyses was also conducted to predict the average coders' label of a '4' from the three versions of Recovery—Incremental (2 versions), Integration Distance, Semantic Grouping—and the single version of Planning. The

regressions were performed on all fluent trials, on a trial-by-trial basis, or on the condition means. A summary of the regression models tested is presented in Table 2.

As is evident from the Table, the results of these regressions pattern with those conducted on the data when boundaries were considered to be labels of '3' or '4.' The models that use the Semantic Grouping version of Recovery account for as much or more variance as the other versions of Recovery.

		Beta	SE	p-value	R^2	F and p-value
Trial-by-trial	Coder 2 Incremental Left	0.02	0.00	0.001	0.02	F(2,2526) = 24.44, p<.001
	Planning Right	0.02	0.00	0.001		
	Coder 1 Incremental Left	0.00	0.01	0.333	0.05	F(2,2526) = 59.43, p<.001
	Planning Right	0.05	0.01	0.001		
	Integration Left	0.06	0.01	0.001	0.09	F(2,2330) = 111.42, p<.001
	Planning Right	0.02	0.00	0.001		
	Semantic Grouping Left	0.11	0.01	0.001	0.10	F(2,2330) = 123.25, p<.001
	Planning Right	0.01	0.00	0.001		
Condition Means	Coder 2 Incremental Left	0.04	0.01	0.001	0.62	F(2,37) = 30.39, p<.001
	Planning Right	0.02	0.01	0.001		
	Coder 1 Incremental Left	0.07	0.09	0.001	0.68	F(2,37) = 39.37, p<.001
	Planning Right	0.07	0.01	0.001		
	Integration Left	0.06	0.01	0.001	0.61	F(2,37) = 28.30, p<.001
	Planning Right	0.02	0.01	0.005		
	Semantic Grouping Left	0.10	0.01	0.001	0.65	F(2,37) = 34.86, p<.001
	Planning Right	0.02	0.01	0.03		

Table 2: A summary of regression models tested on data where 4's are boundaries. Models were tested on all fluent trials on a trial-by-trial basis, and on the means of each condition.

Boundaries as Word Duration and Silence: Regression

A series of multiple regression analyses was also conducted to predict the length of the final word of a phonological phrase, and following silence, from either the Integration Distance or Semantic Grouping version of Recovery, and the single version of Planning. The regressions were performed on a trial-by-trial basis, or on the condition means, and were performed on all of the data, as described in the Method section. A summary of the regression models tested is presented in Table 3.

Similar to the regression analyses reported above, the regressions that use the Semantic Grouping version of Recovery perform better than those which use the Integration version of Recovery.

Table 3: A summary of regression models tested on the data comparing the predictions of each model to the duration of the final word of each phonological phrase and any following silence. Models were tested on all data on a trial-by-trial basis, and on the means of each condition.

		Beta	SE	p-value	R^2	F and p-value
By-trial	Integration Left	0.03	0.00	0.001	0.09	F(2,2491) = 123.86, p<.001
	Planning Right	0.03	0.02	0.001		
	Semantic Grouping Left	0.08	0.01	0.001	0.13	F(2,2491) = 182.74, p<.001
	Planning Right	0.02	0.00	0.001		
Means	Integration Left	0.02	0.01	0.001	0.58	F(2,37) = 25.64, p<.001
	Planning Right	0.02	0.00	0.001		
	Semantic Grouping Left	0.06	0.01	0.001	0.85	F(2,37) = 105.03,p<.001
	Planning Right	0.01	0.00	0.001		

Discussion

The aim of the current study was to investigate the relationship between syntactic structure and intonational phrasing, and, specifically, to test the predictions of the Left-Right Boundary Hypothesis (LRB), proposed by Watson & Gibson (2004). The LRB predicts the probability of a boundary according to the size of the syntactic material that precedes a phonological phrase boundary (Recovery) and the size of the material that follows a phonological phrase boundary (Planning). We evaluated the claims of the LRB and some related hypotheses by designing materials which allowed us to compare three different versions of Recovery and one version of Planning. Specifically, we gathered naïve speakers' productions of sentences which varied the length of three post-verbal arguments using a speaker/listener conversation task. The location of intonational boundaries in the sentences was determined according to ToBI-labeled break indices, and by acoustic measures of duration of phonological phrase-final words and following silence. The predictions of the LRB were evaluated in a series of analyses, which will be described below and referred back to in a discussion of the findings of the experiment.

The results of the experiment support a model of intonational boundary production based on two factors: Recovery and Planning. Specifically, the probability of a boundary at a specific location in a sentence increases with the size of the material a speaker has recently completed and with the size of the material the speaker is going to produce. Moreover, the accumulated results support a version of Recovery based on the size of the most recently completed syntactic constituent (Semantic Grouping) over versions which quantify Recovery in terms of Integration Distance or the distance back to the last boundary (Incremental). Semantic Grouping was supported by analyses of variance conducted at critical points across the sentences, and by regression analyses. Targeted analyses failed to support specific predictions of Integration Distance, and regression models that used Integration Distance did not account for speaker behavior as well as Semantic Grouping did. Finally, the Incremental version made some successful predictions when boundaries were quantified in one way, but failed to correctly predict speaker behavior when boundaries were quantified differently, and failed to account for as much variance as Semantic Grouping in a series of regression analyses. These results support Watson and Gibson's original formulation of the LRB, which we restate in (7), with Recovery changed to Semantic Grouping.

(9) Left Constituent / Right Constituent Boundary (LRB) Hypothesis: Two independent factors are summed to predict the likelihood of an intonational boundary at a phonological phrase boundary in production.

1) Semantic Grouping (LHS): The number of phonological phrases in the largest syntactic constituent that has just been completed. A constituent is completed if it has no rightward dependents.

2) Planning (RHS): The number of phonological phrases in the largest upcoming syntactic constituent if it is not an obligatory argument of the most recently processed constituent.

Summary of analyses

The first set of analyses consisted of a series of analyses of variance where the dependent measure was the proportion of boundaries which speakers produced at phonological phrase boundaries in the sentences. Boundaries in this set of analyses were quantified in two ways:

As places where the ToBI coders indicated boundaries of strength '3' (intermediate boundary) or '4' (full intonational boundary).

As the sum of the duration of the phonological phrase-final word and any following silence.

The second set of analyses consisted of a series of analyses of variance where the dependent measure was the proportion of boundaries which speakers produced at phonological phrase boundary in the sentences. Boundaries in this set of analyses were quantified in two ways:

As places where the ToBI coders indicated boundaries of strength '4' (full intonational boundary).

As the duration of any silence which followed a phonological phrase-final word.

The third set of analyses consisted of a series of regressions, in which we compared the fit of the data to models which consisted of the planning component paired with one of the four versions of recovery: (1)semantic grouping, (2) integration distance, (3) incremental based on the labels of ToBI labeler 1. or (4) incremental based on the labels of ToBI labeler 2. That is, we compared how often speakers placed boundaries in the locations that each of the models predict. In addition, we computed the model fit on a trial-by-trial basis, as well as over the means of the conditions. Boundaries for these analyses were once again quantified as places where the ToBI coders indicated boundaries of strength '3' (intermediate boundary) or '4' (full intonational boundary).

The fourth set of analyses consisted of a second set of regression models in which we again compared the fit of the data to models which consisted of the planning component paired with one of the four versions of recovery. In this set of analyses, however, boundaries were quantified as places where the ToBI coders indicated boundaries of strength '4' (full intonational boundary).

The fifth and final set of analyses was a series of regression equations where we compared the fit of the data to models which consisted of the planning component paired with one of two versions of recovery (semantic grouping or integration distance). Here, however, the regression model was no longer predicting ToBI labels but rather the duration of phonological phrase-final words and any following silence.

Summary of findings

Support for Recovery:

Evidence consistent with all recovery theories:

1. Analysis 1a demonstrated that speakers placed more boundaries after long direct objects than short direct objects and that speakers placed more boundaries after long indirect objects than short indirect objects.
2. Analysis 1b demonstrated that phrase-final words which completed long direct objects were longer, and more likely to be followed by silence, than words which completed short direct objects. In addition, this analysis demonstrated that phrase-final words which completed long indirect objects were longer and more likely to be followed by silence than words which completed short indirect objects.
3. Analysis 2a demonstrated that speaker placed more boundaries after long direct objects than short direct objects, and that speakers were more likely to place boundaries after long indirect objects than short indirect objects
4. Analysis 2b demonstrated that longer silences followed long direct objects than short direct objects.

Evidence that helps decide among recovery theories:

Semantic Grouping

The Semantic Grouping version of Recovery was supported by the main effects of length of direct object, indirect object, and modifier demonstrated in analyses 1a, 1b, 2a, and 2b.

Semantic Grouping was also supported over the other two versions of Recovery by the regression models tested in analyses 3, 4, and 5. In each of these analyses, the models based on Semantic Grouping accounted for more variance in the speaker production data than the other two models. This result held when the models were run on the data on a trial-by-trial basis and when the models were run on the means of each condition.

Incremental Distance

Analysis 1a provided support for the Incremental Distance version of Recovery in two ways: by a two-way interaction between the length of the direct object and the length of the indirect object, and by a three-way interaction between length of direct object, length of indirect object, and length of modifier. The two-way interaction suggests that the placement of a boundary at an early point in the sentence influences the probability of a boundary later in the sentence. Specifically, if a speaker had not placed a boundary after a short direct object, s/he would be more like to place a boundary after a short indirect object, a result that is in contrast to the predictions of both the Semantic Grouping and Integration Distance versions of Recovery. The three-way interaction suggests that, when the modifier is long enough, it will be preceded by a boundary, regardless of whether or not a boundary has recently been produced.

Integration Distance:

In analysis 1a, the Integration Distance version of Recovery failed in two important ways, as evidenced by a null effect in one critical comparison, when an effect was predicted, and a strong effect in another comparison where a null effect was predicted. Specifically,

speakers placed more boundaries at the syntactic boundary between the indirect object and the modifier when integration distance was matched, but the length of the immediately preceding constituent (i.e. the indirect object) was varied. Moreover, when integration distance was varied, but the length of immediately preceding constituent (i.e. the indirect object) was held constant, the incidence of boundaries was the same.

In analysis 2a the Integration Distance version of Recovery was again not supported in two ways: Speakers' boundary production did not change when the distance back to the verb was manipulated but the size of the most recently-completed constituent was the same; and speakers' boundary production increased when the size of the completed constituent increased but integration distance back to the verb was held constant.

Support for Planning:

1. Results from analysis 1a indicated that speakers were somewhat more likely to place boundaries before long indirect objects than short indirect objects. In addition, they were more likely to place boundaries before long modifiers than short modifiers.
2. Results from analysis 1b demonstrated that speakers produced significantly longer phrase-final words and measurable silence preceding long modifiers than short modifiers.
3. Results from analysis 2a indicated that speakers were more likely to place boundaries before long indirect objects than short indirect objects and that speakers were more likely to place boundaries before long modifiers than short modifiers.
4. Results from analysis 2b indicated that speakers placed longer silences before long modifiers than short modifiers.
5. Results from analyses 3, 4, & 5 indicated that in every model tested, the Planning component of the model accounted for a significant amount of variance in speaker boundary placement.

In addition to offering insight into the processes that underlie spoken language production, the results of the current study motivate us to reiterate a point made by Watson & Gibson which may be able to explain some of the discrepancies between previous investigations of the relationship between syntactic structure and speech planning (cf. Watson & Gibson, 2004 for similar arguments). For example, although Kraljic and Brennan (2005) found evidence that speakers will prosodically disambiguate a syntactically-ambiguous sentence whether or not they are aware of the ambiguity, Snedeker and Trueswell (2003) found that speakers only prosodically disambiguated sentences if they were aware of the ambiguity. We suggest that the important difference between the two studies, which Kraljic and Brennan point out, is the length of the stimuli used in each. Whereas the sentences that Snedeker and Trueswell used were comprised of three phonological phrases (e.g. "Tap the frog with the feather"), the sentences in Kraljic and Brennan's studies were comprised of four phonological phrases (e.g. "Put the dog in the basket on the star").

Both the modifier attachment production and the goal attachment production of "Put the dog in the basket on the star" result in places in the sentence where the LRB weight is high enough to induce speakers to place boundaries as a result of constraints on speech production. When the modifier attachment of "on the star" is intended, the

semantic grouping component of the LRB has a weight of one ("the dog"), while the *planning* component of the LRB has a weight of two ("in the basket on the star"), resulting in a total LRB weight of three. Similarly, when the goal attachment of "on the star" is intended, the *semantic grouping* component of the LRB has a weight of two ("the dog in the basket"), while the *planning* component of the LRB has a weight of one ("on the star"), resulting in a total LRB weight of three. In contrast, both the modifier and goal attachment productions of "Tap the frog with the feather," result in LRB weights of only two. When the modifier attachment of "with the feather" is intended, a boundary would be expected between "Tap" and "the frog," but the LRB weight of two may not be large enough to induce the speaker to place a boundary in this location. Similarly, when the instrument attachment of "with the feather" is intended, a boundary would be expected to produced between "the frog," and "with the feather." Once again, the total LRB weight at this location may not be enough to induce the speaker to insert a boundary.

However, the additional pressure of having a listener understand an ambiguous sentence was enough to induce the speakers to place disambiguating boundaries in the sentence. That is, the speakers' knowledge of the ambiguity, and their need to communicate the ambiguity to their listener, can induce them to place a boundary at a point in the sentence even if the LRB weight is lower than three phonological phrases. The assertion that there is a threshold LRB weight at which a speaker must place a boundary, and that the desire to communicate to a listener can lower this threshold and induce boundaries, can be tested in future work using sentences in which the same types of constituents differ systematically in length.

A final important contribution of this study is the demonstrated relationship between the quantitative and qualitative evaluations of the speakers' data. That is, we observed high correlations between the acoustic measures of word duration and/or silence and the perceptual ToBI labels, indicating, as previously noted, that phrase-final lengthening and silence are strong cues to the presence of an intonational boundary. These correlations are especially noteworthy considering that the words which are being compared have very different segmental characteristics and are produced by different speakers, due to the fully between-subject design of the experiment . Although the acoustic results did not always mirror the ToBI codings, it is plausible that we would not observe all the effects in the acoustic data that we observed for the perceptual data, because the duration and silence are only two of the many acoustic cues that gives rise to the perception of a boundary. The relationship that we did observe, however, suggests that with more research into the acoustic cues that give rise to the perception of intonational boundaries, future investigations of phrasing and syntactic structure may be undertaken without necessitating the involvement of multiple, highly-trained prosodic annotators.

Chapter 4

Chapter 2 demonstrated that coders achieved very high agreement on the application of binary prosodic categories to a large corpus of speech. These results suggest that the binary categories of both prosodic annotation systems (ToBI and RaP) are good reflections of the prosodic features that speakers and listeners are using to communicate. However, labeler agreement decreases substantially for categories with more than two levels, suggesting that these multi-level categories may not be readily applied to the normal speech system, and that prosodic annotation systems are not accurately capturing important categories of prosody.

The goal of the set of studies described in this chapter was to undertake a methodologically sound exploration of the relationship between the acoustics of words and their information structure. Some of the specific questions we will address are:

1. What are the acoustic features that indicate an entity as accented?
2. Do the acoustic features of entities which are widely focused differ from entities which are deaccented?
3. Are accents acoustically differentiated by speakers on the basis of the type of information status of the accented entity, specifically, for entities which are new or contrastive in the discourse?
4. Are certain types of information status differentiated more readily than others?
5. If (4) is true, under what circumstances will speakers differentiate information status?
6. Is there a reliable relationship between the extent to which a speaker acoustically encodes given/new or new/contrastive distinction and a listener's ability to resolve the information structure?

The first section of this chapter will describe a series of studies which demonstrate that listeners are sensitive to information structure as indicated by prosodic features. Next, we will describe early attempts at characterizing the acoustic characteristics of different information status categories.

Interpretation of pitch accents

This section will recount studies which have shown how naïve listeners are able to recognize and utilize accents to decide which information is new or given in a sentence, and that this information is available to listeners immediately.

Most and Saltz (1979) used a question-choice paradigm to investigate whether listeners interpret stress as signaling new information. Listeners heard sentences like *The PITCHER threw the ball* or *The pitcher threw the BALL*, where words in caps indicate words associated with pitch accents, and were asked to write a question for which the sentence would be an appropriate reply. In 68% of responses, the listeners' questions semantically focused the stressed element (e.g. *Who threw the ball?* in response to *The PITCHER threw the ball*), demonstrating that listeners are interpreting accent placement as indicating the new information in an answer. Although these results are suggestive, no quantitative definition of accent in given in this paper, and consequently, there is no way to know what acoustic or prosodic cues the listeners were interpreting as signaling accent.

Bock and Mazzella (1983) also asked whether accents signal new information, but also investigated the consequences of accent placement for on-line comprehension. Listeners heard a context sentence followed by a target sentence, and were instructed to press a button when they understood what the target sentence meant. The four conditions are presented in Table I. The critical manipulations were whether a) the context sentence semantically focused an element by the placement of an accent, and b) whether the accent in the target sentence was on the item which was semantically focused by the context sentence.

Condition	Context Sentences	Target Sentences
1. Appropriate accent	ARNOLD didn't fix the radio.	DORIS fixed the radio.
2. Inapproriate accent	Arnold didn't FIX the radio.	DORIS fixed the radio.
3. No context accent	Arnold didn't fix the radio.	DORIS fixed the radio.
4. Control	Arnold didn't fix the radio.	Doris fixed the radio.

Table 1: Sentence Set from Bock and Mazzella (1983) Experiment 1

The results demonstrated that listeners understoond the target sentence most quickly when it had been preceded by a context sentence with an appropriately placed accent (condition 1). The fact that condition 1 resulted in faster comprehension times than the "No context accent" condition (4) demonstrates that the semantic focusing of the element that will be contrasted in the target sentence facilitates listeners' comprehension of the target sentence. Once again, there is no discussion of the quantification of accent in this experiment, save the agreement of the accent producer and one other listener. Therefore, this paper leaves open the question of how accents are in fact realized such that they can facilitate listener comprehension of contrastive information.

Most and Saltz (1979) and Bock and Mazzella (1983) demonstrated that the *presence* of accents can facilitate listeners' comprehension. There is also evidence from Terken and Noteboom (1987) that listeners' comprehension can be facilitated by the *absence* of accents. In their experiment, listeners heard spoken descriptions of changes that were taking place in a visually presented configuration of letters. The listeners' task was to verify the spoken descriptions, like "the p is on the right of the k," as quickly as possible. Previous descriptions were manipulated so that both the 'p' and the 'k' were either given or new in the target description. In accordance with prior results, Terken and Noteboom found that listeners were faster to verify new information that was accented than new information that was deaccented. In addition, they demonstrated that given information that was deaccented was more quickly verified than given information that was accented, indicating that listeners are able to use deaccentuation as a cue to the discourse status of the deaccented element.

Birch and Clifton (1995) reported results similar to those of Terken and Noteboom. In a set of four experiments, Birch and Clifton operationalized accents according to the ToBI system of prosodic annotation. They had subjects make speeded judgments of the appropriateness of answers to questions. They found that listeners were faster to indicate that a sentence was an appropriate response to a question when what was new in the sentence was accented or what was given in the sentence was deaccented.

Although the studies described above by Bock & Mazzella (1983), Terken and Noteboom (1983), and Birch and Clifton (1995) demonstrate that the appropriateness of accents placement can have an early effect on listeners' comprehension, they leave open

the question of *when* this referential information is available to listeners. Dahan, Tanenhaus, & Chambers (2002) conducted an eye-tracking study which showed that listeners can use accent information immediately to decide the identity of a temporarily ambiguous word by determining whether it was given or new in the discourse. The researchers presented subjects with a visual display containing four objects and four geometric shapes. The critical objects in the test trials were two nouns that shared the same first syllable (e.g. candle, candy) or onset and nucleus (e.g. bell, bed). Listeners followed spoken instructions like the following to move these items around the display:

1 a. Put the candle below the triangle.

b. Put the candy below the triangle.

2 a. Now put the CANDLE above the square

b. Now put the candle ABOVE THE SQUARE.

The eye-tracking data demonstrated that when subjects heard sentence 1a followed by 2a, they were more likely to look at the competitor (candy) than the target (candle) than when sentence 1a was followed by 2b. That is, subjects interpreted the accent on "candle" in sentence 2a as indicating that the word with which it was associated was new to the discourse. Conversely, when subjects first heard sentence 1b, they were more likely to look to the competitor (candle) when it was unaccented (2b) than when it was accented (2a). In this case, subjects interpreted the deaccenting of "candle" as indicating that it was given in the discourse.

The studies described above all demonstrate that both the presence and absence of accents on semantically focused sentence elements affect listeners' comprehension of the discourse. One question left open by the studies described above, however, is whether different types of accents have different meanings. Most and Saltz (1979) and Bock and Mazzella (1983), and Terken and Noteboom (1983), provide no description of the phonological or acoustic realization of the accents in their studies. Birch and Clifton (1995) indicate that all accents in their four experiments were realized with a L+H* accent, according to the ToBI conventions, which will be described below; however, in the ToBI framework, the L+H* accent is hypothesized to indicate contrastive information (Pierrehumbert & Hirschberg, 1990), and in Birch and Clifton's experiments, all accents indicated new information. Dahan, et al. (2002) also report the phonological categories of their accents, but did not control for the type of pitch accent that was associated with the target noun in their study. In fact, the ToBI codings they report indicate that out of 24 cases, the accent was H* in 15 cases, and L+H* in the remaining 9. The following section will discuss hypothesized semantic categories of accents, and experimental work designed to investigate the reality of these categories, both acoustically and perceptually.

Categories of pitch accents

There has been extensive debate throughout the history of research on intonation about the extent to which different accents signal different types of information. There are two major camps in this debate: those who argue that different pitch accents indicate different semantic categories (Pierrehumbert & Hirschberg, 1990; Gussenhoven, 1983), and those who believe that a single accent can indicate one (or more) meanings depending on the context in which it occurs (Bolinger, 1961; Cutler, 1977).

The most attention has been paid to the question of whether different categories of accent signal new and contrastive information. Bolinger (1961) argues that the meaning of a particular accent is defined by the context in which it occurs; that there is no unique

contrastive intonation, as defined by pitch. Cutler (1977), likewise, claims that "the attempt to extract from [intonation contours] an element of commonality valid in all contexts must be reckoned a futile endeavor" (p. 106).

In contrast to the position advocated by Bolinger and Cutler, however, is that taken by Gussenhoven (1983) and Pierrehumbert & Hirschberg (1990). Gussenhoven, having defined accents as *falls, rises,* and *fall-rises,* argues that a *fall* means that the speaker is adding the accented item to the discourse, whereas the *fall-rise* indicates that the speaker is selecting the accented item from those already in the discourse. Gussenhoven's position is similar, then, to Pierrehumbert and Hirschberg (1990), who propose a compositional semantics of intonational meaning where pitch accents, phrase accents, and boundary tones combine to form the meaning of a sentence. With regard to pitch accents, they claim that a high tone on a stressed syllable (H*) signals new information and a high accent preceded by a low (L+H*) signals contrastive or corrective information They suggest that a speakers' use of H* indicates that the proposition including the H* should be added the listener's beliefs. In contrast, a speaker's use of L+H* indicates that the s/he intends that "the accented item—and not some alternative related item—should be believed" (p. 296).

What is missing from the above debate about whether or not different accents can convey meaning differences irrespective of the context in which they occur is empirical investigation into the acoustic realization of accents associated with new information and those indicating contrast.

Acoustic realization of accents

This section will discuss work aimed at defining accents acoustically, and, furthermore, whether there are acoustic differences between accents that have been proposed to have different meanings.

Previous work has suggested that a large variety of acoustic measures indicate the presence of accents in English. The list of proposed factors includes: pitch (Lieberman, 1960; Cooper, Eady & Mueller, 1985; Eady and Cooper, 1986), duration (Fry, 1954), and energy (Bolinger, 1958; Kochanski, Grabe, Coleman, & Rosner, 2005).

Early studies on the acoustic correlates of stress focused on lexical stress. Both Fry (1954) and Lieberman (1960) investigated this question by measuring acoustic features of words whose part of speech varies with stress. For example, con-TRAST is a verb, while CON-trast is a noun. In both production and perception studies, Fry found that intensity and duration of the vowel of the stressed syllable contributed most strongly to the noun-verb disambiguation, such that stressed vowels were produced with a longer duration and a greater intensity than non-stressed vowels. Lieberman replicated Fry's results by demonstrating that stressed syllables had higher amplitudes and longer durations than their non-stressed counterparts. Moreover, he demonstrated that the fundamental frequency of stressed syllables was higher than non-stressed syllables. Finally, Lieberman proposes a decision algorithm that can discern, with 99.2% accuracy, which syllable of a two-syllable word is stressed by combining information about F0, amplitude, and duration.

Studies investigating acoustic features of phrasal stress have investigated similar variables. Cooper, et al. (1985), for example investigated the role of duration and fundamental frequency in speakers' production of focus. In this study, they manipulated the position of focus in a target sentence like: *Chuck liked the present that Shirley sent to*

her sister. They did this by manipulating which element in the sentence contrasted with prior discourse, as the example in (2) demonstrates:

(2) A. Did William or Chuck like the present that Shirley sent to her sister?

B. Did Chuck like the letter or the present that Shirley sent to her sister?

C. Did Chuck like the present that Melanie sent to her sister or the one that Shirley sent?

D. Did Chuck like the present that Shirley sent to her sister or the one she sent to her brother?

With regard to duration, Cooper, et al. demonstrated that the duration of the focused word was significantly greater than when it was not focused. The last word of the sentence (e.g. *sister*) showed a significantly smaller increase in duration due to focus, but this may be because all sentence-final words are lengthened, thus washing out duration differences due solely to focus. There was no effect of deaccenting on duration, in that words which occurred *after* the sentence focus were not shortened relative to non-focused words in the same position.

The effects of fundamental frequency (F0) were similar to those of duration in that all but the sentence-initial focused word (e.g. *Chuck*) exhibited a higher F0 than non-focused words in the same position. In contrast to the duration results, however, there was evidence of deaccentuation in the F0 results, such that words which occurred after the sentence focus were produced with a lower mean F0 than non-focused words in the same sentence position.

In a second study, Eady & Cooper (1986) again investigated the role of F0 and duration in the production of stress in sentences like: *The ship is departing from France on Sunday.* Contrary to the previous experiment, however, they defined the focused element of a sentence in terms of what was focused by a wh-question, as in (3). In this way, the words which were stressed in this experiment would be stressed by virtue of being new information, and not be virtue of being contrastive, as they had been in the previous study.

(3) A. What is happening?

B. What is departing from France on Sunday?

C. On what day is the ship departing from France?

The results from this experiment were quite similar to the previous experiment in which speakers produced stress on contrastive elements. First, stressed elements were produced with longer durations in both initial and final positions. Second, focused words were realized with higher F0 than non-focused words. Again, there was evidence of deaccentuation such that post-focal words were realized with a lower F0 than neutral words.

From the results of these two experiments, Eady and Cooper (1986) conclude that "the type of sentence focus used [in the second study] has essentially the same effect on acoustical attributes as does contrastive stress" (p.408), suggesting that there is no acoustic difference between accents realizing new and contrastive stress. However, this claim is hard to evaluate for two reasons: First, these results come from two independently conducted experiments, and it is not possible to compare speaker behavior across experiments; second, the authors limited their analyses to only select acoustic measurements, and differences would have been observed if they had looked at additional

parameters. The next section will report results of experiments in which new and contrastive stress are compared within the same experiments.

One caveat of both studies conducted by Eady and Cooper is that, although they were investigating the productions of multiple speakers, they did not include all speakers in their analyses. They selected speakers for analysis based on the extent to which the speakers produced accents in the appropriate places, as determined by one of the authors and another listener.

Although all of the studies described in this section thus far have found evidence of F0 differences between stressed and unstressed words, a recent study has called into question the importance of F0 in determining the perception of accents. Kochanski, et al. (2005) trained a classifier to recognize pitch accents which had been hand-labeled in a corpus of read and spontaneous speech, using five acoustic measures: loudness, duration, aperiodicity (a measure of the periodicity of the waveform ranging from 0 to 1), spectral slope (an acoustic correlate of the 'breathiness' of the speech), and f0. They found that loudness was the best predictor of accents, regardless of speaking style or dialect. Moreover, F0 was the worst of the five predictors in determining whether a syllable was accented.

The studies described in this section provide evidence that F0, intensity (loudness), and duration all contribute to some degree to the perception of accent. The following section will explore a series of experimental studies intended to explore whether these acoustic features are different depending on whether the accent they define is signaling "new" or "contrastive" information.

New vs. contrastive accents

The first part of this section will review recent work exploring whether new and contrastive stress are perceived categorically. The second section will review work exploring whether these accents are produced categorically.

Perception

Bartels and Kingston (1994) synthesized a continuum of stimuli intended to vary between H* and L+H* by independently varying four acoustic characteristics of the target accent. They then presented sentences containing these synthesized accents to naïve listeners and asked them to make a meaning judgment, in order to test the hypothesis that listeners would interpret entities accompanied by an H* as new to the discourse but interpret entities accompanied by L+H* as contrastive with information already in the discourse (Pierrehumbert & Hirschberg, 1990). From their results, they conclude that the most salient cue to contrastiveness is the height of the peak of a high accent, such that a higher peak indicates greater contrast. Two secondary cues to contrastivensss were (a) the depth of the dip preceding the peak, such that deeper dips indicated greater contrastiveness and (b) the timing of the peak, such that early peaks were interpreted as more contrastive than late peaks. Importantly, there was no clear evidence of a categorical boundary between L+H* and H* in terms of meaning differences. That is, it was not the case that L+H* consistently signaled contrastive information while H* consistently signaled new information. Bartels and Kingston suggest that the main finding that a higher peak indicates greater contrastiveness supports the possibility that contrastive accents are merely more salient versions of new accents.

Watson, Tanenhaus, & Gunlogson (2004) used an eye-tracking paradigm similar to that utilized by Dahan, et al. (2002) to explore whether listeners immediately interpret

H* and L+H* categorically, with H* indicating new material and L+H* indicating contrastive material. Listeners heard the directions in (1), while interacting with the items in a computerized display. The target item (e.g. camel/candle) in sentence 1c was crossed with the type of accent aligned with it in a 2x2 design.

(1)

a. Click on the camel and the dog.

b. Move the dog to the right of the square.

c. Now, move the *camel/candle* below the triangle.

L+H*/H*

Eye-tracking results demonstrated that listeners were more likely to look to the contrastive referent (the camel) when they heard the L+H* accent than to the new referent (the candle), suggesting that listeners quickly interpreted the L+H* accent as contrastive. In contrast, listeners did not look more quickly at the new referent (the candle) when it was indicated with a H* accent, indicating that H* is not preferentially treated as signaling new information. Although these results suggest that L+H* is preferentially interpreted as interpreting contrastive information, they do not indicate that H* is preferentially treated as signaling new information. They support the notion, rather, that L+H* has preferentially meaning, but that H* could indicate either meaning, and, therefore, that there does not exist a categorical distinction between the two accents.

Taken together, these perception studies suggest that listeners can perceive differences between accents depending on the intended meaning. However, the difference between accents that signal new and contrastive meanings does not appear to be categorical, such that one accent preferentially signals new information and another signals contrastive information. These data also suggest that differences between new and contrastive accents are determined by the height of the pitch peak on the accented syllable, rather than a difference in the overall shape of the accent.

Production

Ito, Speer, and Beckman (2004) developed a novel paradigm for eliciting spontaneous productions which would vary in their information structure, including not only elements that were new or given with respect to previous context, but also items which were contrastive with previous discourse. They found that given adjectives often received a pitch accent, whereas given nouns were less likely to be accented. Conversely, both adjectives and nouns which were used contrastively (e.g. "*green* candy" preceded by "*beige* candy" or "green *candy*" preceded by "green *house*") were more likely to be produced with a L+H* accent than when used non-contrastively. These data give some suggestion that speakers use different accents to indicate items which are new vs. items which are contrastive. One advantage of this study over other production studies is the inclusion of data from sixteen speakers, ensuring that observed differences are not merely due to idiosyncrasies of individual speakers; however, the study is limited in two important ways: First, there is no explanation of the acoustics of the accents; only the ToBI labels are reported; second, no statistical measures of the reported differences are provided.

Krahmer & Swerts (2001) also investigated the shape of contrastive accents and how they differ (if at all) from new accents. To do this, they had eight subjects engage in a game task, where they produced noun phrases (adjective + noun) in one of four

discourse contexts: (1) both new (NN), (2) both contrastive (CC), (3) contrastive adjective / given noun (CG), (4) given adjective / contrastive noun (GC). Two raters identified the location of accents and found that in cases where both adjective and noun were new or contrastive, speakers tended to place accents on both adjective and noun. When only one of the two was contrastive, the contrastive element tended to receive a pitch accent while the non-contrastive element was deaccented.

To investigate whether contrastive accents differ from new accents, Krahmer and Swerts presented the utterances of two speakers from the production study to listeners. They played the productions in one of two contexts: either the adjective and noun together, or each in isolation. The listeners' task was to say which item in a pair was more prominent. The results indicated that listeners perceived contrastive accents as more prominent than new accents when the accents were presented in context; when presented in isolation, there is no clear pattern of difference between contrastive and new accents. Krahmer and Swerts interpret these results as indicating that, although there is no categorical difference between new and contrastive accents, the difference between new and contrastive accents can readily be determined from the context in which these accents are produced. Unfortunately, Krahmer & Swerts present results from only two of eight speakers, and include no explanation for the selection of those two speakers.

Calhoun (2005), like Krahmer and Swerts, also addressed the possibility the differences between new and contrastive accents can be realized non-locally; that is, that the acoustic features of words in the local domain of an accented word can vary systematically when the type of accent on the focused word varies even when the acoustics of the focused word do not differ. She explored this hypothesis in a pilot study where she compared version of the phrase "when that moves the square" where the information status of *square* varied between given, new and contrastive with respect to previous discourse. She first demonstrated that local acoustic cues differ significantly depending on discourse status in that a regression model demonstrated that the mean F0 of *the*, the mean F0 of *square*, and the duration of *square* all significantly predicted the information status of *square*. Second, she found that the non-local context of *square* could also predict it's discourse status, such that the F0 and intensity of the phrase *that moves* were also significant predictors of the topic status of *square*. Finally, she demonstrated that the F0 and intensity of *square* relative to the F0 and intensity of *that* and *moves* could also predict the discourse status of *square*. Taken together, these results suggest that the topic status of a word can be reflected in the acoustics of the word's entire phrase, and not just by acoustic measures of the word itself.

Finally, Calhoun (2003) also attempted to discern a reliable difference between H* and L+H*. In a production study, she had one speaker produce sentences in which the two accents had been disambiguated by context. She first identified where the speaker had placed accents, and, for each accent, measured a series of acoustic features. The results of this portion of the experiment demonstrated that the two accents were most strongly differentiated by (a) the alignment of the pitch minimum preceding the high and (b) the height of the high maximum. A subsequent perception study, in which listeners chose their preferred rendition of an accent in a semantic context, demonstrated that only the height of the pitch maximum mattered to the disambiguation of new and contrastive accents.

As in the perception studies, the production studies reported here also suggest that speakers can differentiate accents on the basis of meaning. They provide additional

support for the hypothesis contrastive accents are indicated by a higher pitch peak, and that contrastive accents are realized with greater intensity than new accents.

Current Study aims

We designed the current set of studies to address limitations of the studies described above and to investigate the function of acoustic features in the determination of information status.

Limitations of previous studies

We first identified several methodological limitations of previous investigations of the relationship between acoustic measures and information status. With regard to the perception of prosody, most studies use trained speakers to produce their materials (Calhoun, 2003; Birch & Clifton, 1995, Most & Saltz, 1979; Bock & Mazzella, 1983). These speakers are presumably fully aware of the experimental aims, and therefore aim to produce maximally different prosody for different semantic conditions. This approach is problematic, because the obtained results from these studies may not generalize to all speakers or all perceivers. Even when production studies do employ naïve speakers, they do not use all of the speakers. Several previous experiments have excluded speakers' data from analysis for not producing accents consistently (Eady & Cooper, 1986), or with no explanation (Krahmer & Swerts, 2001).

The current studies attempt to account for these stated limitations in several ways: First, the speakers in the current studies were not trained speakers, so any differences in their productions should be naturally occurring. Second, we did not exclude speakers from our experiments on the basis of whether or not they behaved as we hypothesized they should (i.e. placed accents in particular places). We excluded only subjects who were not native English speakers, who did not take the task seriously, or who were not recorded well.

A limitation of previous studies of the perception of prosody is that, rather than report acoustic measures, they often report only the ToBI annotations of their materials (Birch & Clifton, 1995; Ito, et al., 2004). Investigations of inter-annotator agreement in ToBI suggests that H* and L+H* are the most often confused in the ToBI system (Syrdal & McGory, 2000) and are often collapsed in these studies (Pitrelli, et al., 1994; Yoon, et al., 2004). Therefore, it is difficult to interpret results of studies which are based on the difference between H* and L+H* without any reporting of the acoustic differences between these categories. To avoid the confusion inherent in ToBI labels, we report acoustic measures which differ across conditions, and not just ToBI labels, in order to avoid confusion about what these labels might mean.

Finally, the current studies attempt to improve on the task demands of previous studies. Previous studies asked listeners to make judgments about which of two stimuli was more prominent (Krahmer & Swerts, 2001), or what accent is acceptable in a particular context (Birch & Clifton, 1995). Our producers and perceivers were engaged in a meaning task; the perceivers were trying to communicate a particular meaning of a sentence, and our dependent measure was the perceiver's semantic interpretation of the sentence. This approach is an improvement over previous methods because it relates differences in accent types to differences in meaning, rather than simply to perceptual differences.

The details of the current method for Experiments 1-3 were as follows: Each speaker in the experiment produced the same sentence under seven different information status conditions, which were manipulated by the setup question that the speaker was answering with each question. These conditions were produced by putting either new or contrastive focus on either the subject, verb, or object. From the speaker's production, a Listener chose from seven possible sentences which one s/he though the speaker was answering. We performed acoustic analyses on the productions that speakers were able to correctly categorize, assuming that those productions bore the correct cues to information status and focus location.

Experiment 1

Method

Participants

We recorded 16 pairs of subjects for this study. Six speakers were excluded from analysis for the following reasons: One speaker was not a native American English speaker, two speakers were too quiet to be acoustically analyzed, and three did not take the task seriously, often laughing during their productions. Each subject received ten dollars/hour for his/her participation in both experimental roles (speaker and listener).

Materials

Each trial consisted of a set-up question and a target sentence. The target sentence could plausibly answer any one of the seven set-up questions, which served to focus different constituents on the sentence. Two factors were manipulated: (1) the constituent in the target sentence that was focused by the question (subject, verb, object); and (2) the discourse status of the focused constituent (new, contrastive). In addition, we added an additional condition which focused the entire sentence (i.e. *What happened?*), for a total of seven conditions. We included this condition in order to see whether accents in any the narrow-focus conditions differed from those in the wide-focus condition.

The words in the target sentences were chosen so that they could be compared across items and to aid in the extraction of acoustic features. To this end, all subject names and object noun phrases (NPs) were two-syllable names/words with first-syllable stress, comprised of sonorant phonemes, such as "Damon" and "omelet." All verbs were one-syllable, comprised mostly of sonorant phonemes, such as "fried."

We constructed 14 sets of all 7 conditions, resulting in 98 experimental items. Each subject pair was presented with every item, resulting in a full within-subjects design. A complete item is presented in Table 1. All materials can be found in Appendix A.

Condition	Status	Focused Argument	Setup Question	Target
1	New	wide	What happened yesterday?	Damon fried an omelet yesterday.
2	New	S	Who fried an omelet yesterday?	Damon fried an omelet yesterday.
3	New	V	What did Damon do to an omelet yesterday?	Damon fried an omelet yesterday.
4	New	O	What did Damon fry yesterday?	Damon fried an omelet yesterday.
5	Contrastive	S	Did Harry fry an omelet yesterday?	No, Damon fried an

				omelet yesterday.
6	Contrastive	V	Did Damon bake an omelet yesterday?	No, Damon fried an omelet yesterday.
7	Contrastive	O	Did Damon fry a chicken yesterday?	No, Damon fried an omelet yesterday.

Table 1: Example item from Experiment 1

Procedure

The experiment was conducted using Linger (2.92), a software platform for language processing experiments. Linger was written and designed by Doug Rohde, and can be downloaded at http://tedlab.mit.edu/~dr/Linger/. Two participants were included in each trial, and sat at computers in the same room such that neither could see the others' screen. One participant was the speaker, and the other was the listener. The speakers were instructed that they would be producing answers to questions out loud for their partners (the listeners), and that the listeners would be required to choose which question the speaker was answering from a set of seven choices.

Each trial began with the speaker being presented with the question on the computer screen to read silently until s/he understood it. The speaker then saw the answer to the question, accompanied by a reminder that s/he would only be producing the answer aloud. Following this, the speaker had one more chance to read the question and answer, and then was instructed to press a key to begin recording, and another key to stop recording.

The listener sat at another computer, and pressed a key to see the seven questions that s/he would have to choose his/her answer from. When s/he felt familiar with the questions, s/he told the Speaker s/he was ready. After the speaker produced a sentence out loud for the listener, the listener chose the question s/he thought the speaker was answering.

In early pilots in which there was no feedback for incorrect responses, we observed that Listeners were at chance to choose the correct question, and that Speakers were not perceptibly disambiguating the answers for the Listener. In order to remedy this situation, we introduced feedback for both the Speaker and the Listener such that when the Listener answered incorrectly, his/her computer emitted an audible buzz. Listeners who received this feedback were well above chance levels in choosing the correct answer, presumably because the Speakers were explicitly aware when they had not provided accurate prosodic cues to the correct answer.

Data Analysis

We report results based on productions of all speakers in the experiment except those trials that were not recorded or cutoff for technical reasons, trials that were too quiet to contribute useful acoustic data, and trials in which the speaker was disfluent, or used different words from those presented on the screen. Overall, these exclusions results in 105 of the 980 total trials (9%).

Acoustic Factors

Based on previous investigations of the acoustic correlates of prosodic features, we chose a series of acoustic measures which we believed would reflect accentuation. For each word, we obtained measures of the following, using the Praat program (Boersma & Weenink, 2006) :

1. **duration**: duration, in ms, of each word, excluding any silence before or after the word.
2. **silence**: duration, in ms, of any measurable silence following the word, which was not due to stop closure.
3. **duration + silence**: the sum of the duration of the word and any following silence, in ms.
4. **mean pitch**: The mean F0, in Hz, of the entire word
5. **maximum pitch**: the maximum F0 value (in Hz) across the entire word
6. **pitch peak location**: a measure between 0 and 1 indicating the proportion of the way through the word where the maximum F0 occurs.
7. **minimum pitch**: the minimum F0 (in Hz) across the entire word
8. **pitch valley location**: a measure between 0 and 1 indicating the proportion of the way through the word where the minimum F0 occurs.
9. **initial pitch**: the mean F0 value of the initial 5% of the word
10. **early pitch**: the mean F0 value (in Hz) of 5% of the word centered at the point 25% of the way through the word
11. **center pitch**: the mean F0 value (in Hz) of 5% of the word centered on the midpoint of the word
12. **late pitch**: the mean F0 of 5% of the word centered on a point 75% of the way through the word
13. **final pitch**: the mean F0 of the last 5% of the word
14. **mean intensity**: mean intensity (in dB) of the word
15. **maximum intensity**: the highest dB level in the word
16. **minimum intensity**: the lowest dB level in the word
17. **intensity peak location**: a measure between 0 and 1 indicating the proportion of the way through the word where the maximum intensity (in dB) occurs
18. **intensity valley location**: a measure between 0 and 1 indicating the proportion of the way through the word where the minimum intensity (in dB) occurs
19. **maximum amplitude**: the maximum amplitude (sound pressure in Pascal) across the word
20. **energy**: the square of the amplitude multiplied by the duration of the word
21. **1st quarter pitch**: The difference, in Hz, between initial pitch and early pitch.
22. **2nd quarter pitch**: The difference, in Hz, between early pitch and center pitch.
23. **3rd quarter pitch**: The difference, in Hz, between center pitch and late pitch.
24. **4th quarter pitch**: The difference, in Hz, between late pitch and final pitch.

Results – Production

Important acoustic factors – ANOVAs

In order to determine which acoustic factors differed across conditions, we performed a series of 4 x 3 analyses of variance on each acoustic measure. The first factor, *information status*, had four levels, corresponding to (1) given, (2) wide focus, (3) narrow new focus, and (4) narrow contrastive focus. The second factor, *sentence position*, had three levels: (1) Subject, (2) verb, and (3) object.

(1) Given words were any words which were not focused in one of the sentences where either the subject, verb, or object were focused. For example, when the setup question focused the subject, as in *Who fried an omelet yesterday?*, or *Did Harry fry an omelet yesterday?*, the given words in the answer were *fried* and *omelet.*

(2) Words with wide focus were either the subject, verb, and object when the setup question focused the entire sentence, as in *What happened yesterday?.*

(3) Words with narrow new focus were either the subject, verb, or object when the setup question imposed new focus on that constituent. For example, if the setup question was *What did Damon fry yesterday?*, *omelet* received new focus in the answer.

(4) Words with narrow contrastive focus were either the subject, verb, or object when the setup question imposed contrastive focus on that constituent. For example, if the setup question was *Did Damon fry a chicken yesterday?*, *omelet* received contrastive focus in the answer.

We performed each ANOVA using both subjects and item as random factors. The results from the entire series of ANOVAs are presented in Table 1, which indicates for which of the four information status levels each acoustic measure varied significantly.

Acoustic measure	Information Status differentiated
duration	1, 2, 34
silence	1, 24, 3
duration + silence	1, 2, 34
mean pitch	1, 2, 3, 4
maximum pitch	1, 2, 34
pitch peak location	1, 234
pitch valley location	12, 34
early pitch	1, 2, 34
center pitch	1, 2, 3, 4
maximum intensity	1, 2, 34
maximum amplitude	1, 2, 34
energy	1, 2, 34
1st quarter pitch	1, 234
2nd quarter pitch	123, 4
3rd quarter pitch	12, 3, 4
4th quarter pitch	12, 34

Table 2: Information status differentiated by individual acoustic features across the subject, verb, and object.

Acoustic features used in disambiguation

Although the results of the analyses of variance provide some evidence as to which acoustic features speakers use to disambiguate both the presence or absence of accents, and the type of information status intended by a particular type of accent, they cannot answer the question of which acoustic factors were ultimately responsible for Listeners' condition choices, for two reasons. First, the ANOVAs only indicate acoustic features which differ between conditions in single sentence locations; they do not provide any indication of differences across the sentences. Listeners hear each word in the context of a complete sentence, and presumably interpret the acoustics of each word in the sentence with reference to the other words. As such, the absolute difference between words in the same position across different conditions may not matter as much to the Listener as the relationship between a word of interest (focused or non-focused) and the other words in the sentence. Second, the ANOVAs do not provide any indication about the relative importance for the different acoustic features in determining differences in Listener behavior. In other words, the finding of significant differences between one acoustic factor across conditions does not necessarily indicate that Speakers and Listeners are using that acoustic feature to disambiguate between conditions.

Therefore, in order to determine which of the candidate acoustic features actually mediated the difference between different accent locations or different levels of information status, we conducted a series of stepwise discriminant analyses on the data from Experiment 1, where we entered all acoustic features for all words as predictors. Across all analyses, the acoustic features which consistently resulted in the best classification of conditions were (1) duration + silence, (2) mean pitch, (3) maximum pitch, and (4) maximum intensity. From this result, we used only these four factors in the discriminant analyses that we performed on all subsequent data sets in this chapter.

Following the identification of the critical acoustic measures through the stepwise discriminant analyses, three subsequent discriminant analyses were conducted to determine whether the measures of (1) duration + silence, (2) maximum pitch, (3) mean pitch, and (4) maximum intensity on the three critical words in the sentence could predict (a) accent location, (b) information status, and (c) new vs. contrastive status.

Accent Location – correct trials

The overall Wilks's lambda was significant, $\Lambda = .071$, $\chi^2(24) = 1102.85$, $p < .001$, indicating that the acoustic measures could differentiate among the three accent locations. In addition, the residual Wilks's lambda was significant, $\Lambda = .283$, $\chi^2(11) = 536.01$, $p < .001$, indicating that some predictors could still differentiate accent location after partialling out the effects of the first function. Figure 1 indicates a separation of the focus locations on the discriminant functions.

When we tried to predict focus location, we were able to correctly classify 94.6% of the sentences in our sample. To assess how well the classification would perform on a new sample, we performed a leave-one-out classification and correctly classified 93.4% of our sentences. At individual sentence locations, the discriminant function was able to correctly classify subject focus 98% of the time, verb focus 93% of the time, and object focus 94% of the time.

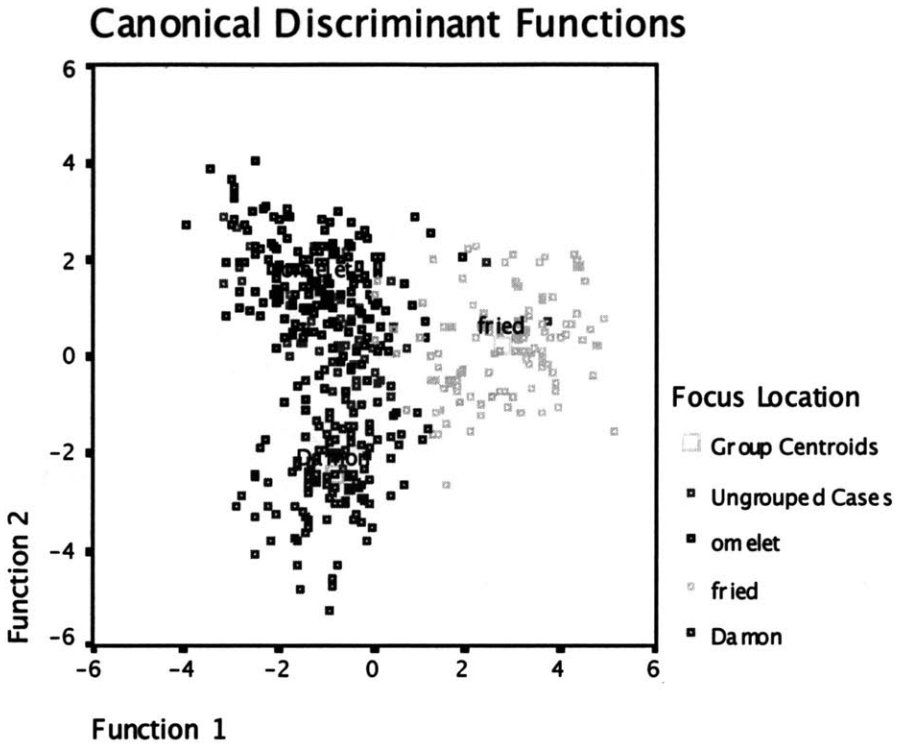


Figure 1: Separation of focus locations on two discriminant functions in Experiment 1

Information Status – correct trials

The overall Wilks's lambda was significant, $\Lambda = .846$, $\chi^2(24) = 78.37$, $p < .001$, indicating that the acoustic measures could differentiate among the information status types. In addition, the residual Wilks's lambda was significant, $\Lambda = .950$, $\chi^2(11) = 24.06$, $p < .05$, indicating that some predictors could still differentiate information status after partialling out the effects of the first function. Figure 2 indicates a separation of the information status on the discriminant functions.

When we tried to predict information status, we were able to correctly classify 50.0% of the sentences in our sample. To assess how well the classification would perform on a new sample, we performed a leave-one-out classification and correctly classified 45.8% of our sentences. For individual levels of information status, the discrim function was able to correctly classify new focus 54% of the time, contrastive focus 41% of the time, and wide focus 71% of the time.

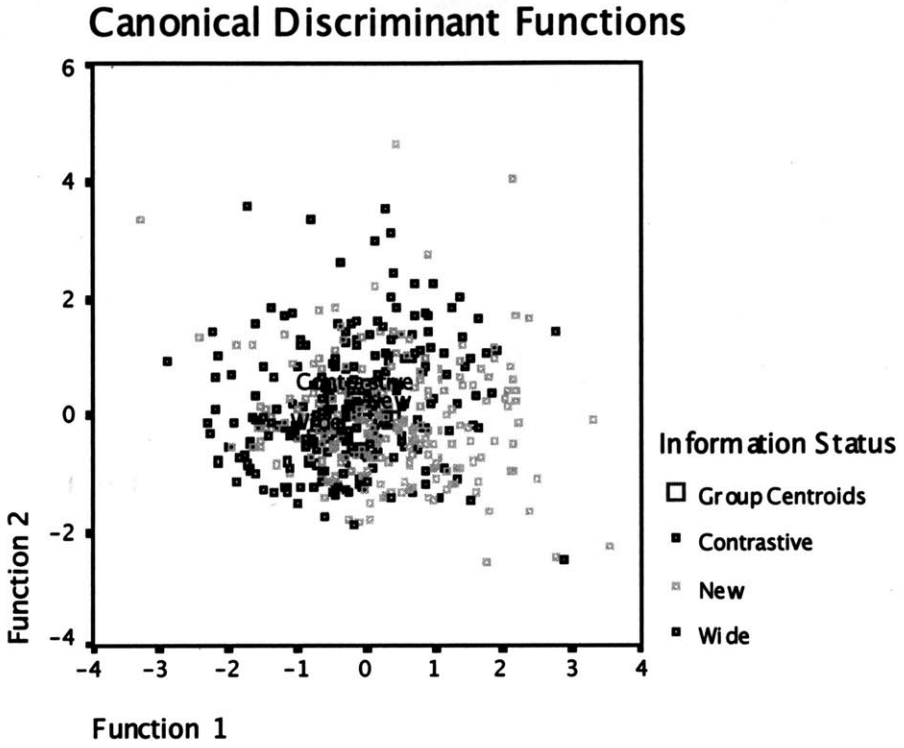
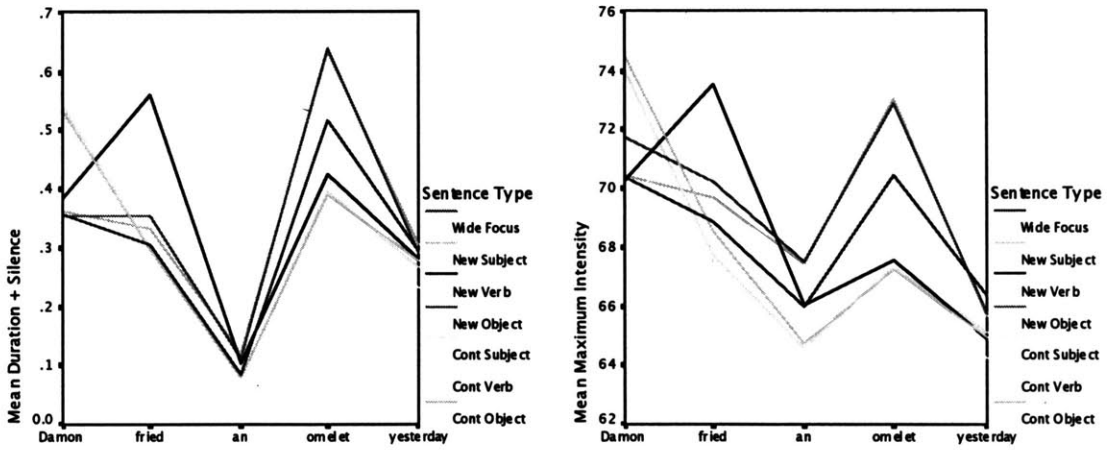


Figure 2: Separation of information status on two discriminant functions for Experiment 1.

New vs. Contrastive – correct trials

The overall Wilks's lambda was significant, $\Lambda = .914$, $\chi^2(12) = 37.47$, $p < .001$, indicating that the acoustic measures could differentiate between the new and contrastive conditions. When we tried to predict new vs. contrastive status, we were able to correctly classify 65.2% of the sentences in our sample. To assess how well the classification would perform on a new sample, we performed a leave-one-out classification and correctly classified 59.5% of our sentences.



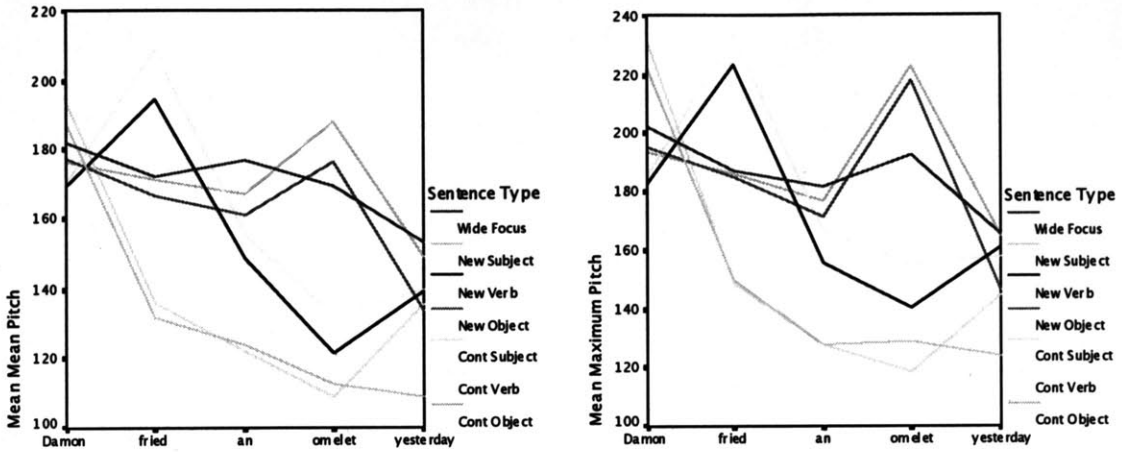


Figure 3: Average value of four acoustic features across every word and every condition in Experiment 1. The top left graph indicates the average of the sum of the duration of each word and the duration of any following silence, in seconds. The top right graph indicates the average maximum intensity in decibels. The bottom left graph indicates the average mean pitch, in Hertz. The bottom right graph indicates the average maximum pitch, in Hertz.

Results – Perception

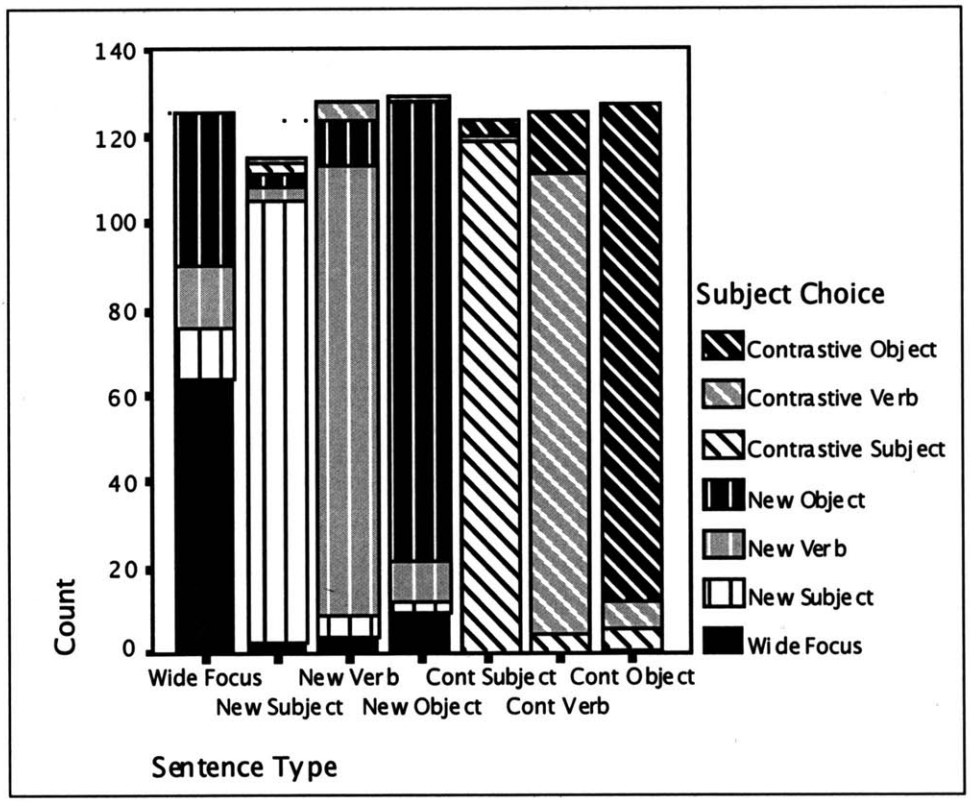


Figure 4: Total count of Listeners' condition choice by sentence type

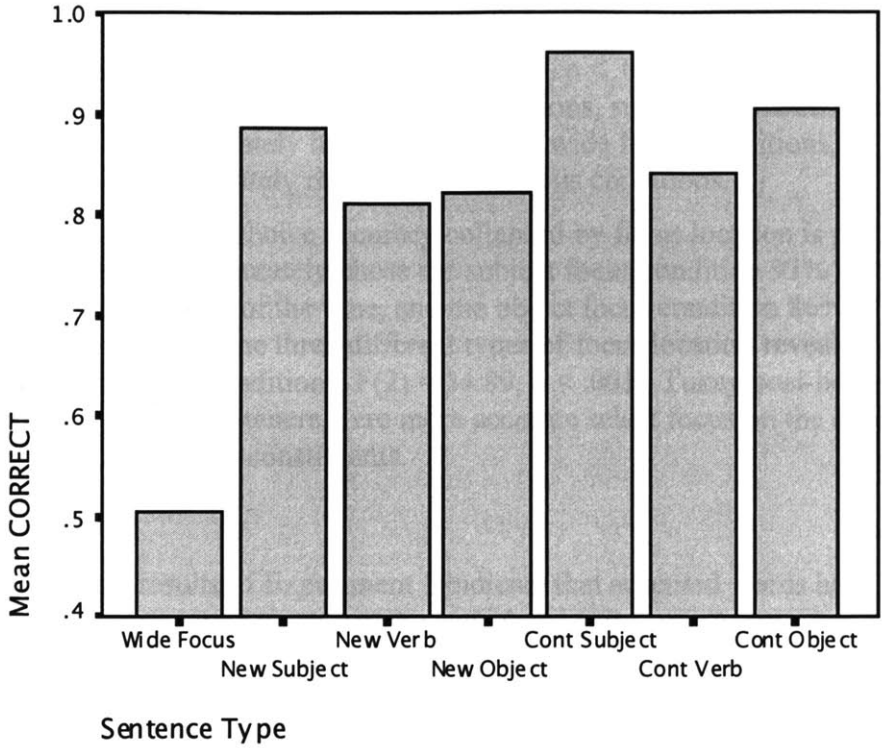


Figure 5: Mean Listener accuracy by condition

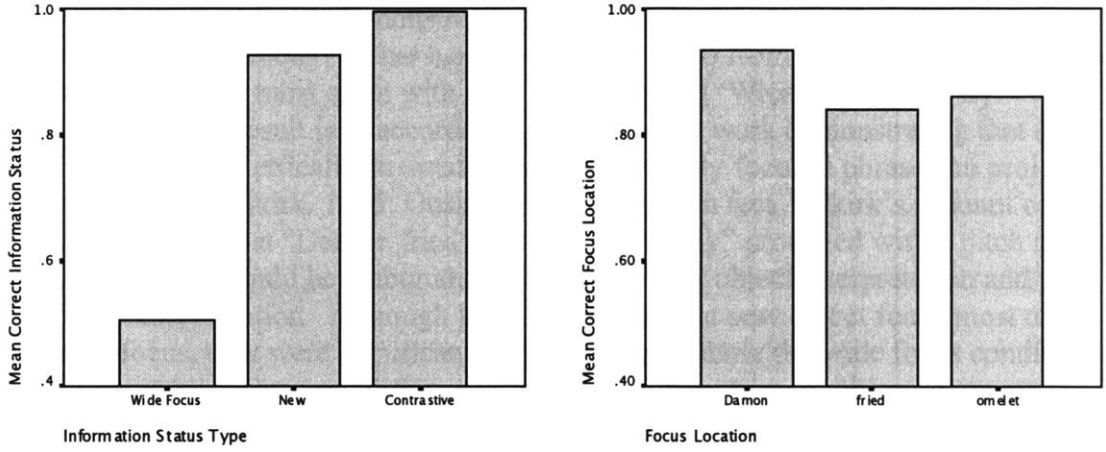


Figure 6: Mean Listener accuracy collapsed by Information Status (left) and focus location (right).

Listeners' choices of question sorted by the intended question are plotted in Figure 4, and their overall accuracy percentage by condition is plotted in Figure 5. Listeners' overall accuracy was 82%. An omnibus ANOVA on accuracy means by condition demonstrated a significant effect of condition, such that some conditions were answered more accurately than others, $F(6) = 124.24$, $p < .001$. Tukey post-hoc comparisons revealed that accuracies for the Wide focus condition (51%) was significantly lower than all of the other conditions. Individual subject accuracy ranged from 52-97%, and there were significant differences between listeners, $F(9) = 85.07$, $p<.001$. There were no significant differences in accuracy across items, $F(13) = 1.42$, $p = .143$.

Listeners' condition choice accuracy collapsed by information status is plotted in Figure 6 (left). Listeners accurately chose the wide focus condition 51% of the time, the new

focus condition 93% of the time, and the contrastive focus condition 99% of the time. An overall ANOVA on the three different types of information status revealed significant differences across conditions, $F(2) = 1077.47$, $p < .001$. Tukey post-hoc comparisons revealed differences between all three conditions, such that constrastive conditions were answered more accurately than either new or wide focus conditions, and new conditions were answered accurately more than wide focus conditions.

Listeners' condition choice accuracy collapsed by focus location is plotted in Figure 6 (right). Listeners accurately chose the subject focus condition 93% of the time, the verb focus condition 84% of the time, and the object focus condition 86% of the time. An overall ANOVA on the three different types of focus location revealed significant differences across conditions, $F(2) = 34.89$, $p < .001$. Tukey post-hoc comparisons demonstrated that Listeners were more accurate select focus on the subject ("Damon"), than on the other two constituents.

Discussion

The acoustic results of Experiment 1 indicate that accented words have longer durations than their de-accented counterparts, incur larger pitch excursions, are more likely to be followed by silence, and are produced with greater intensity. This result replicates previous studies which showed that accents are realized by a variety of acoustic factors. The perception results reflect the acoustic results well, in that listeners were highly successful in discriminating between the three narrow focus conditions which focused the Subject, Verb, and Object respectively.

The perception results also demonstrated that listeners were least successful in discriminating wide focus ("What happened yesterday?") from the other conditions, and confused wide focus most often with New Object focus ("What did Damon fry yesterday?"). This result is in accordance with previous work demonstrating that an accent on the final metrically stressed syllable of a widely focused phrase can project to the entire phrase (Selkirk, 1995; Gussenhoven, 1999). In fact, Selkirk's account of focus projection suggest that "Damon fried an omelet yesterday" produced with a pitch accent only on "omelet" should be ambiguous between the new object interpretation and the wide focus interpretation. Although listeners did confuse new object focus most often with wide focus, they were significantly more likely to chose the wide focus condition than the new object focus condition, suggesting that these two conditions were not produced with the same acoustics. It may be the case that, in order to produce wide focus, speakers were not simply producing only accent on the object.

Despite revealing multiple reliable acoustic differences between focused and non-focused elements, the discriminant analyses performed on the productions showed only minimal differences between focus associated with new information and focus associated with contrastive information. Specifically, the discriminant results demonstrate only fair classification accuracy for the information status of the focused element.

The production results, as indicated by the results of the discriminant function analysis, suggest that Speakers disambiguated focus location with a combination of word duration and silence, intensity, mean F0, and maximum F0. The results of the perception study suggest that Listeners were highly accurate in determining the location of intended focus from the Speaker's use of these acoustic cues. On the other hand, however, the production results as indicated by the discriminant function results, suggest that the Speakers did not successfully disambiguate information status with acoustic cues.

However, the perception results suggest that Listeners were able to successfully disambiguate information status from the Speakers' productions.

The reason that the Listeners were successful in determining information status when the discriminant function was not successful was probably that the Listeners had an additional cue to the disambiguate contrastive focus from either new or wide focus. Specifically, contrastive focus conditions were always begun with a 'No,' which presumably served to disambiguate the information status without the need for acoustic information. We tested this possibility in a second perception study on the productions from Experiment 1.

Experiment 1A

In a second listening experiment, we spliced all of the "No"s out of the contrastive answers so that the Listener would not have this explicit cue to the contrastive status of the focused element. We also spliced out the beginning of the non-contrastive answers so that any residual silence in these answers would not be a cue for the listener. The resulting sentences then all had the form: "Damon fried an omelet yesterday." Although we intended to have several new naïve listeners do the question-choice task with these spliced answers, we abandoned that plan after the author piloted this new task. Her results, illustrated in Figure X, indicate that, although she demonstrated high accuracy in deciding which constituent was being focused, she was at chance to decide whether the accent was new or contrastive.

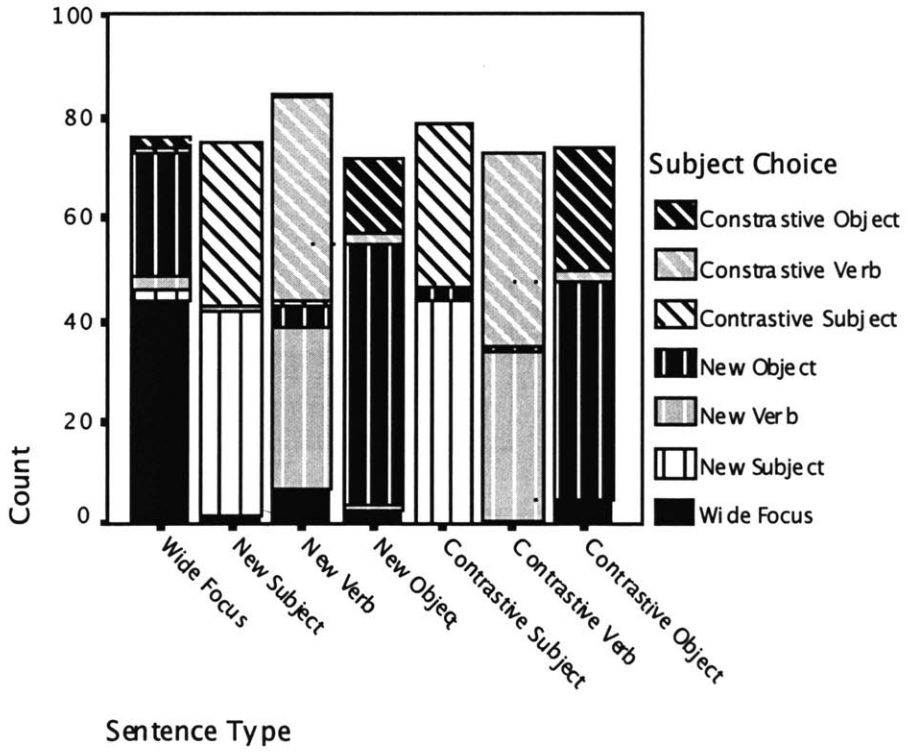


Figure 7: Author's condition choice by correct condition for Experiment 1A productions with "No""s removed.

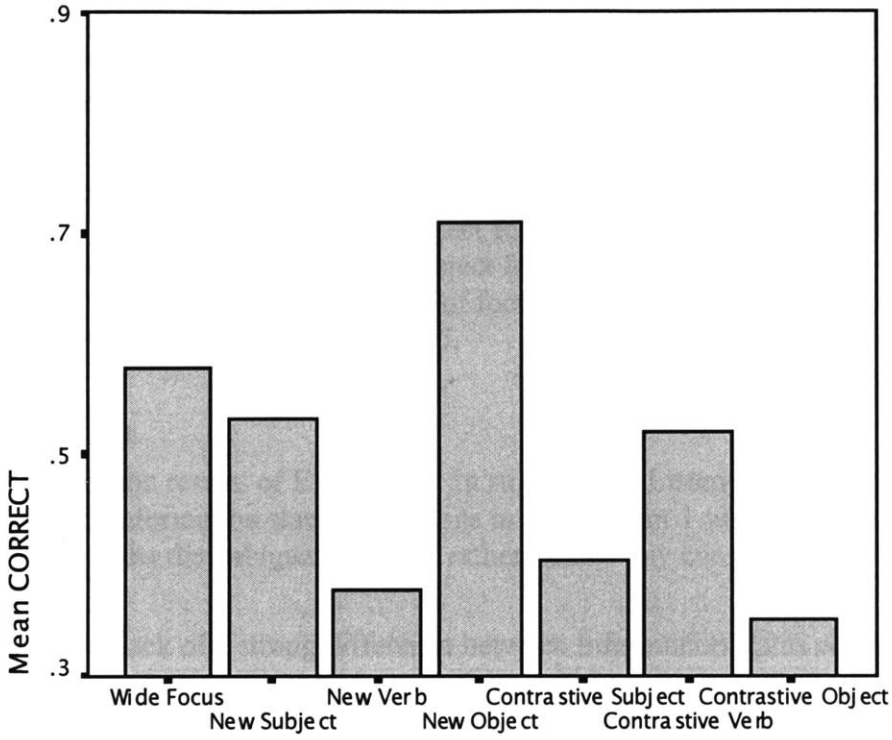


Figure 8: Author's accuracy by condition for Experiment 1A.

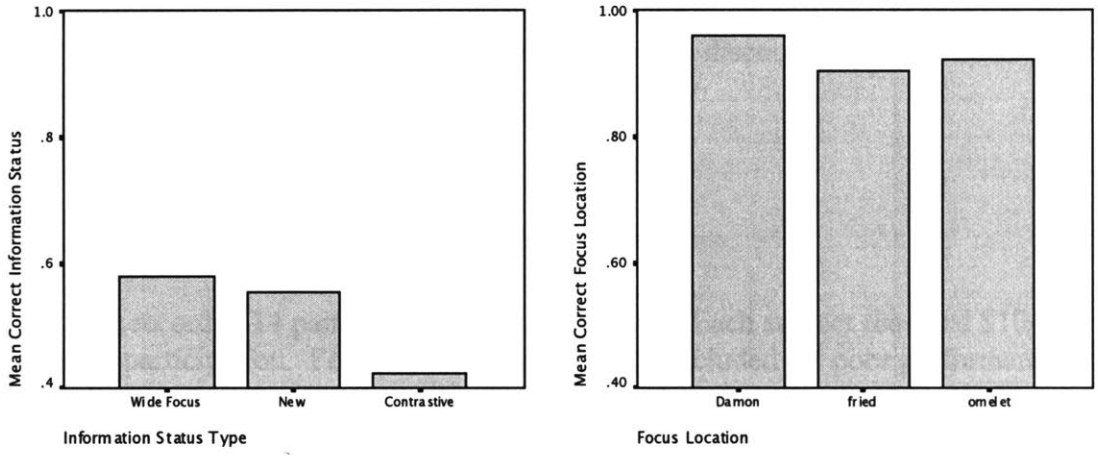


Figure 9: Mean author accuracy collapsed by Information Status (left) and focus location (right).

Results

The author's choice of question sorted by the intended question is plotted in Figure 7. The author's overall accuracy was 48%. An omnibus ANOVA on accuracy means by condition demonstrated a significant effect of condition, such that some conditions were answered more accurately than others, $F(6) = 5.10$, $p < .001$. Tukey post-hoc comparisons revealed that accuracy for the new object condition was significantly higher than for the new verb, contrastive subject, or contrastive object conditions.

The author's condition choice accuracy collapsed by information status is plotted in Figure 9 (left). She accurately chose the wide focus condition 58% of the time, the new

focus condition 56% of the time, and the contrastive focus condition 42% of the time. An overall ANOVA on the three different types of information status revealed significant differences across conditions, $F(2) = 5.01$, $p < .001$. Tukey post-hoc comparisons revealed differences between all three conditions, such that constrastive conditions were answered less accurately than either new or wide focus conditions.

The author's condition choice accuracy collapsed by focus location is plotted in Figure 9 (right). She accurately chose the subject focus condition 96% of the time, the verb focus condition 90% of the time, and the object focus condition 92% of the time. An overall ANOVA on the three different types of focus location revealed no significant difference across conditions, $F(2) = 1.93$, $p = .15$.

Discussion

The perception results of Experiment 1a suggest that Listeners' high accuracy in selecting the correct information status conditions in Experiment 1 was probably due to the presence of the disambiguating 'No,' rather than to any cues in the prosody of the speaker.

The lack of a strong difference between information status conditions could have resulted from one of two reasons. First, it could be the case that speakers do not acoustically differentiate between new and contrastive accents. Alternatively, however, it could mean that speakers do not provide prosodic cues to information that is predictable from the context, as contrastive status was in this experiment, from the inclusion of the 'No' in the answer. A growing body of literature suggests that speakers are less likely to produce intonational cues for sentence structures with more predictable meanings from the contexts (Snedeker & Trueswell, 2004). We tested this possibility in Experiment 2, where we specifically removed the additional cue to the discourse status of the focused constituent ("No").

Experiment 2

Method

Participants

We recorded 14 pairs of subjects for this study. Each subject received $10/hour for his/her participation. Four pairs of subjects were excluded for poor performance, in that the Listeners in the pair chose the correct sentence 20% of the time or less.

Materials

The materials for Experiment Two were identical to those from Experiment 1 described above with the exception that the word "No" was excluded from the contrastive conditions and the words "I heard that" were added to each of the seven conditions. An example item is presented in Table 3.

Condition	Status	Focused Argument	SetupQuestion	Target
1	New	wide	What happened yesterday?	I heard that Damon fried an omelet yesterday.
2	New	S	Who fried an omelet yesterday?	I heard that Damon fried an omelet yesterday.
3	New	V	What did Damon do to an omelet yesterday?	I heard that Damon fried an omelet yesterday.

4	New	O	What did Damon fry yesterday?	I heard that Damon fried an omelet yesterday.
5	Contrastive	S	Did Harry fry an omelet yesterday?	I heard that Damon fried an omelet yesterday.
6	Contrastive	V	Did Damon bake an omelet yesterday?	I heard that Damon fried an omelet yesterday.
7	Contrastive	O	Did Damon fry a chicken yesterday?	I heard that Damon fried an omelet yesterday..

Table 3: Example item from Experiment 2

Procedure

The procedure for Experiment 2 was the same as that described for Experiment 1.

Results – Production

We tested the acoustic features we had identified in Experiment 1 on the new productions in Experiment 2. Once again, we conducted three discriminant analyses to determine whether the measures of (1) duration + silence, (2) maximum pitch, (3) mean pitch, and (4) maximum intensity on the three critical words in the sentence could predict (a) accent location, (b) information status, and (c) new vs. contrastive conditions.

Accent Location – correct trials

The overall Wilks's lambda was significant, $\Lambda = .061$, $\chi^2(24) = 1480.62$, $p < .001$, indicating that the acoustic measures could differentiate among the three accent locations. In addition, the residual Wilks's lambda was significant, $\Lambda = .279$, $\chi^2(11) = 674.66$, $p < .001$, indicating that some predictors could still differentiate accent location after partialling out the effects of the first function. Figure 10 indicates a separation of the focus locations on the discriminant functions.

When we tried to predict focus location, we were able to correctly classify 97.0% of the sentences in our sample. To assess how well the classification would perform on a new sample, we performed a leave-one-out classification and correctly classified 96.5% of our sentences. At individual sentence locations, the discriminant function was able to correctly classify subject focus 93% of the time, verb focus 94% of the time, and object focus 87% of the time.

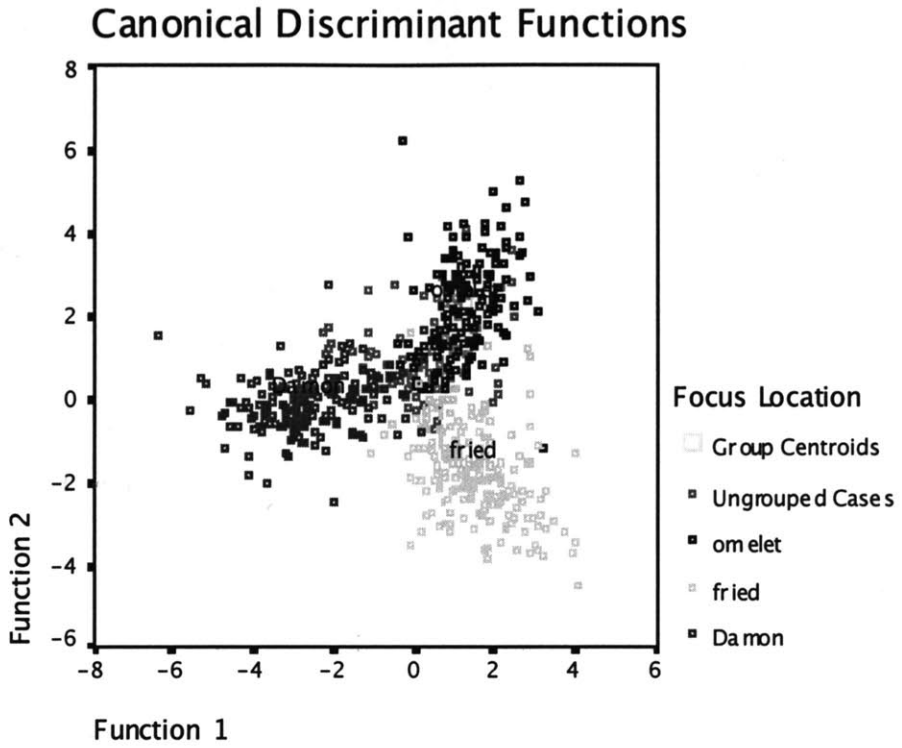


Figure 10: Separation of focus locations on two discriminant functions

Information Status – correct trials

The overall Wilks's lambda was significant, $\Lambda = .705$, $\chi^2(24) = 216.19$, $p < .001$, indicating that the acoustic measures could differentiate among the three information status conditions. In addition, the residual Wilks's lambda was significant, $\Lambda = .963$, $\chi^2(11) = 23.08$, $p < .05$, indicating that some predictors could still differentiate information status after partialling out the effects of the first function. Figure 11 indicates a separation of the information status on the discriminant functions.

When we tried to predict information status, we were able to correctly classify 50.9% of the sentences in our sample. To assess how well the classification would perform on a new sample, we performed a leave-one-out classification and correctly classified 48.3% of our sentences. For individual levels of information status, the discriminant function was able to correctly classify new focus 52% of the time, contrastive focus 44% of the time, and wide focus 70% of the time.

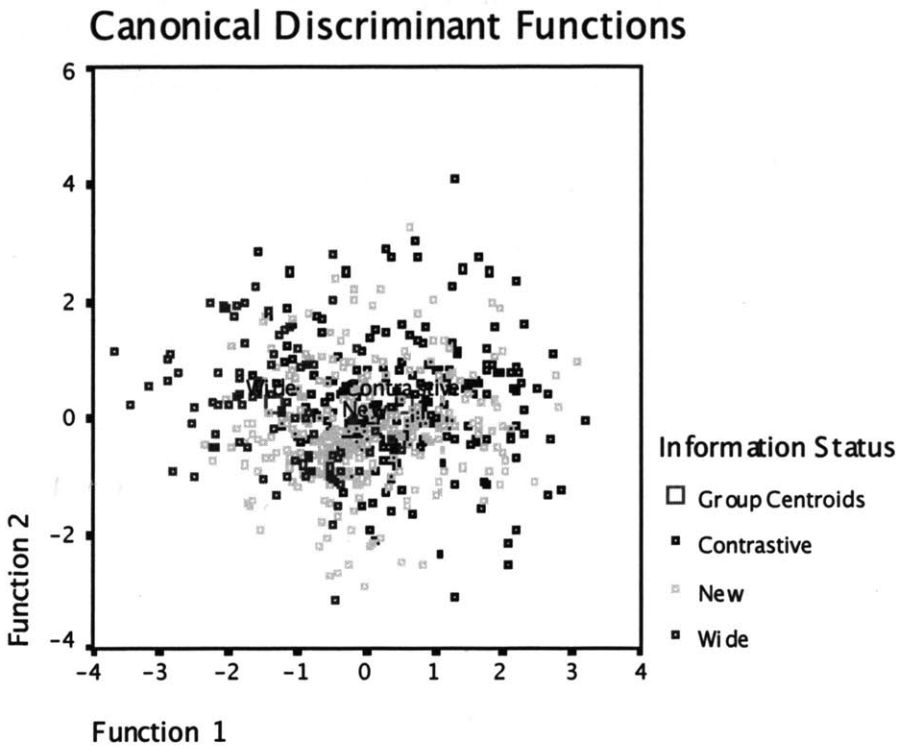


Figure 11: Separation of information status types on two discriminant functions

New vs. Contrastive – correct trials

The overall Wilks's lambda was significant, $\Lambda = .908$, $\chi^2(24) = 51.17$, $p < .001$, indicating that the acoustic measures could differentiate among new and contrastive conditions. When we tried to predict new vs. contrastive information status, we were able to correctly classify 59.8% of the sentences in our sample. To assess how well the classification would perform on a new sample, we performed a leave-one-out classification and correctly classified 57.7% of our sentences.

Information Status – including "I"

We also wanted to determine how important the prosody of the "I heard that" was to the differentiation of the information status of the sentences in Experiment 2. To investigate this question, we performed a stepwise discriminant function analysis which included as predictors the four acoustic factors we had identified in Experiment 1 (duration + silence, mean pitch, maximum pitch, maximum intensity) for the subject ("Damon"), verb ("fried"), and object ("omelet"), as well as for the first three words of the sentence ("I heard that"). The analysis revealed that the most important variables for the differentiation of information status were (1) the duration + silence of "I", (2) the maximum pitch of "I", and (3) the Maximum intensity of "I". Following that, we conducted another analysis in which we included as predictors the duration + silence, mean pitch, maximum pitch, and maximum intensity of the subject, verb, object, and "I."

The overall Wilks's lambda was significant, $\Lambda = .469$, $\chi^2(32) = 454.79$, $p < .001$, indicating that the acoustic measures could differentiate among the three information status conditions. In addition, the residual Wilks's lambda was significant, $\Lambda = .815$, $\chi^2(15) = 122.76$, $p < .001$, indicating that some predictors could still differentiate information status after partialling out the effects of the first function. Figure 12 indicates a separation of the information status on the discriminant functions.

When we tried to predict information status, we were able to correctly classify 69.2% of the sentences in our sample. To assess how well the classification would perform on a new sample, we performed a leave-one-out classification and correctly classified 67.4% of our sentences. For individual levels of information status, the discriminant function was able to correctly classify new focus 72% of the time, contrastive focus 64% of the time, and wide focus 75% of the time.

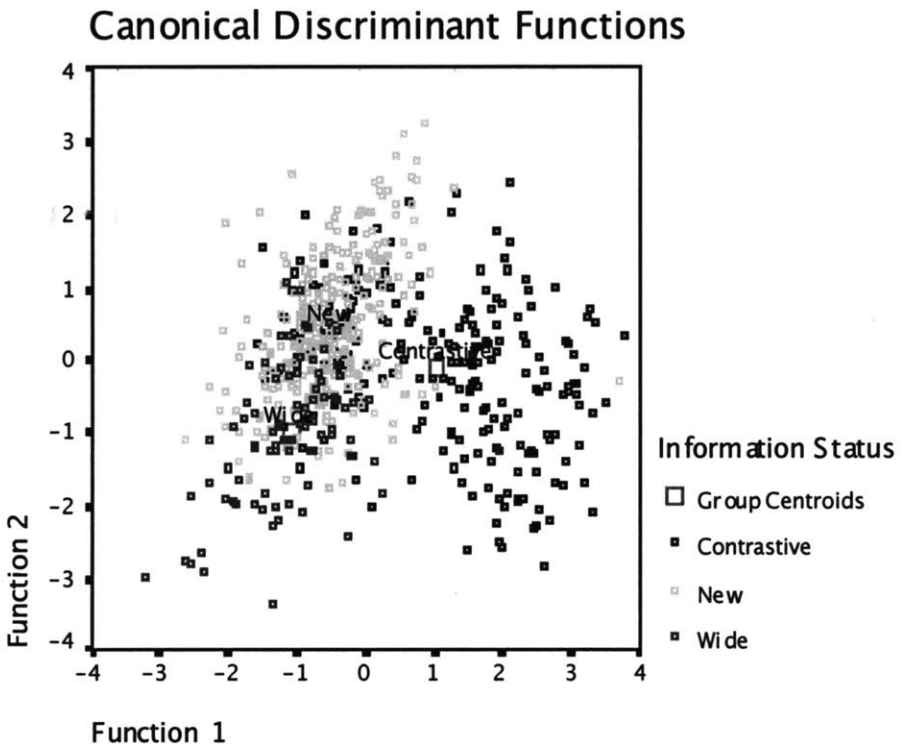


Figure 12: Separation of information status types on two discriminant functions

New vs. Contrastive – including "I"

The overall Wilks's lambda was significant, , $\Lambda = .606$, $\chi^2(16) = 258.37$, $p < .001$, indicating that the acoustic measures could differentiate among new and contrastive conditions. When we tried to predict new vs. contrastive conditions, we were able to correctly classify 79.3% of the sentences in our sample. To assess how well the classification would perform on a new sample, we performed a leave-one-out classification and correctly classified 78.9% of our sentences. This classification accuracy is much higher than that achieved using the acoustics of the subject, verb, and object alone, indicating that speakers were primarily using the prosody of "I" to encode the difference between information status type.

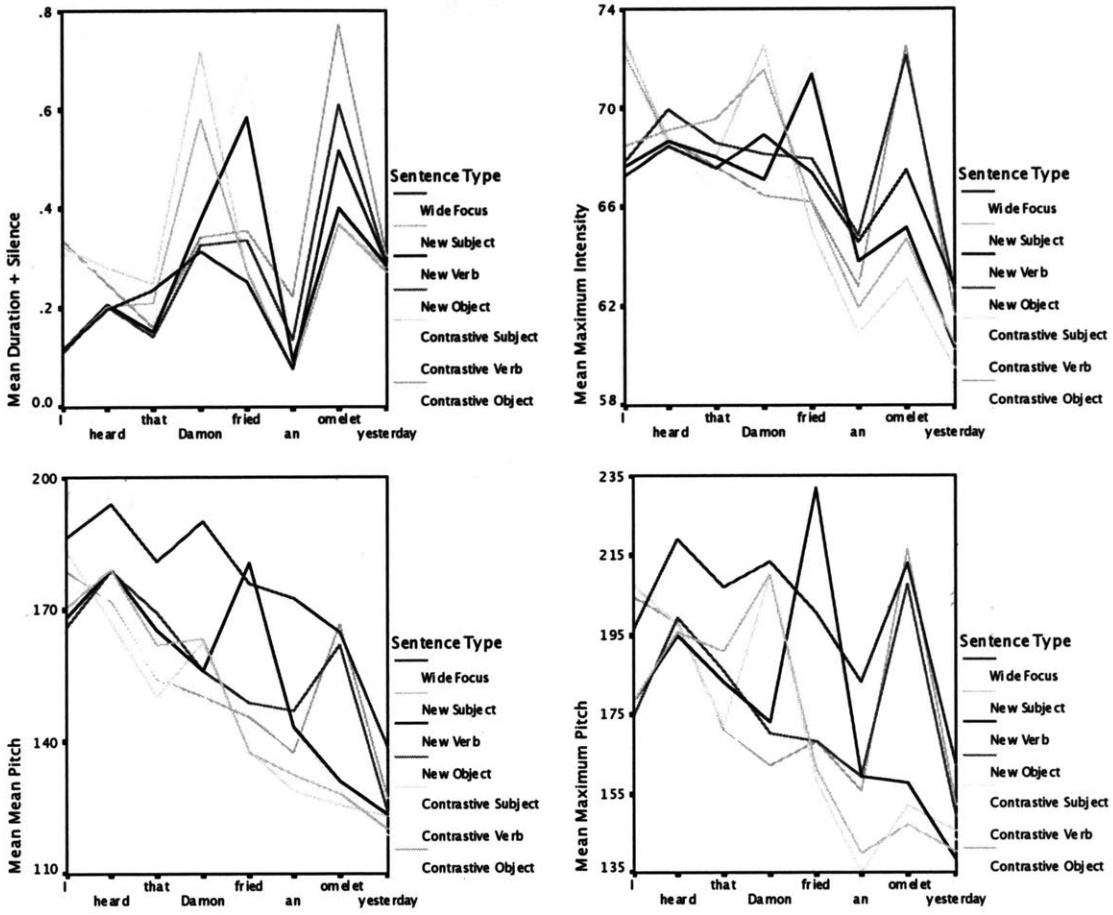


Figure 13: Average value of four acoustic features across every word and every condition in Experiment 2. The top left graph indicates the average of the sum of the duration of each word and the duration of any following silence, in seconds. The top right graph indicates the average maximum intensity in decibels. The bottom left graph indicates the average mean pitch, in Hertz. The bottom right graph indicates the average maximum pitch, in Hertz.

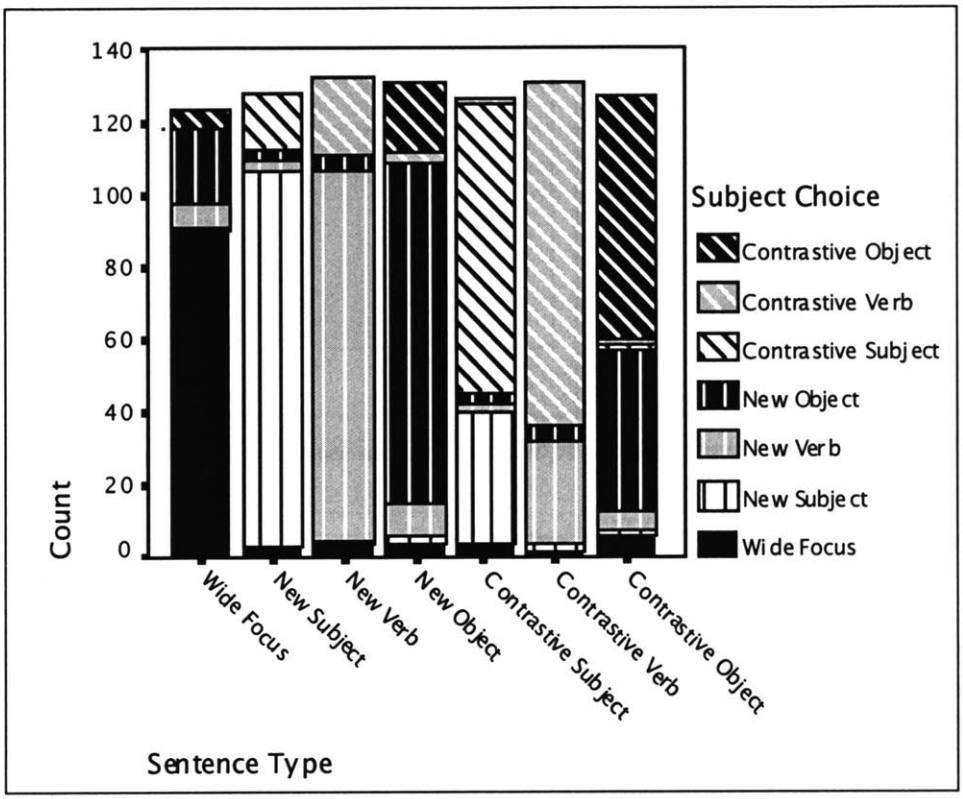


Figure 14: Total count of Listeners' condition choice by sentence type

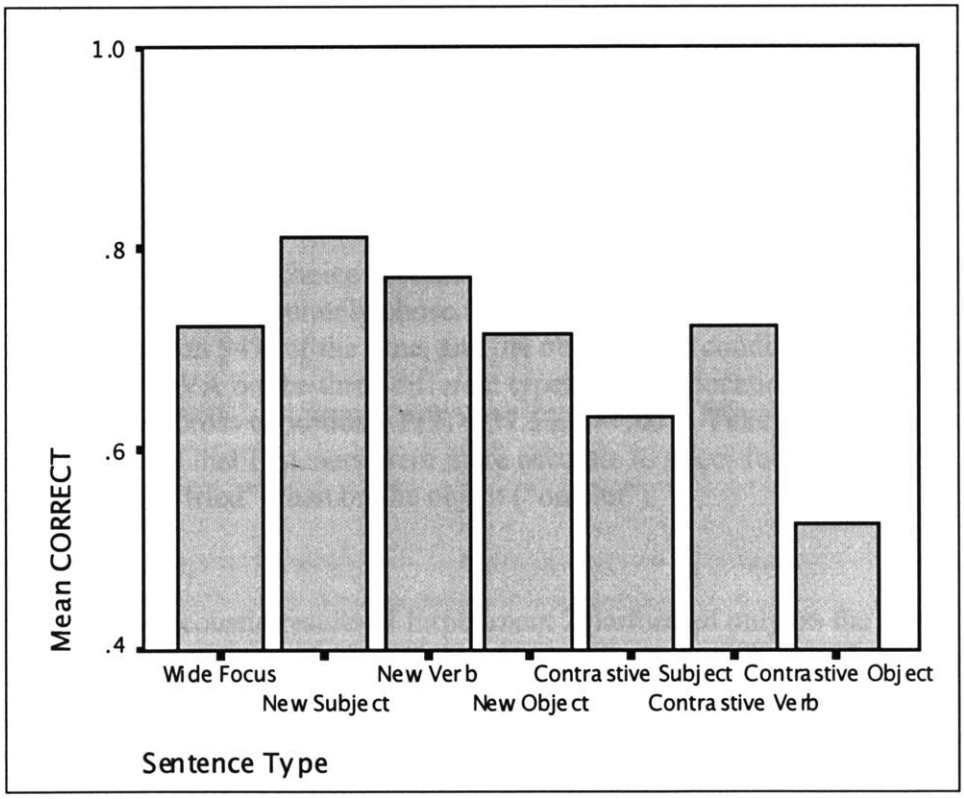


Figure 15: Mean Listener accuracy by condition

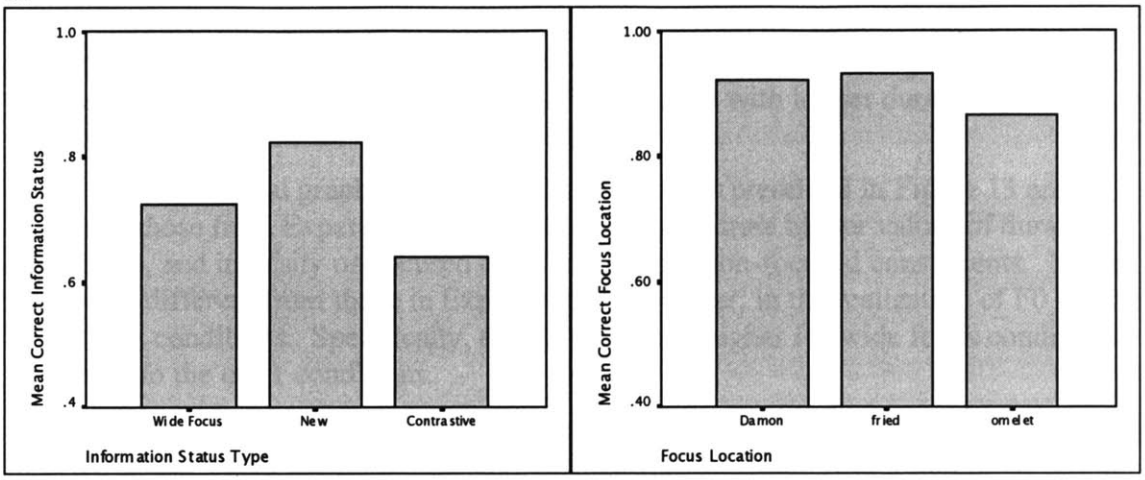


Figure 16: Mean Listener accuracy collapsed by Information Status (left) and focus location (right).

Listeners' choices of question sorted by the intended question are plotted in Figure 14, and their overall accuracy percentage by condition is plotted in Figure 15. Listeners' overall accuracy was 70%. An omnibus ANOVA on accuracy means by condition demonstrated a significant effect of condition, such that some conditions were answered more accurately than others, $F(6) = 8.39$, $p < .001$Individual subject accuracy ranged from 58-80%, and there were significant differences between listeners, $F(9) = 26.00$, $p<.001$. There were no significant differences in accuracy across items, $F(13) = 1.1$.

Listeners' condition choice accuracy collapsed by information status is plotted in Figure 16 (left). Listeners accurately chose the wide focus condition 73% of the time, the new focus condition 82% of the time, and the contrastive focus condition 64% of the time. An overall ANOVA on the three different types of information status revealed significant differences across conditions, $F(2) = 154.39$, $p < .001$. Tukey post-hoc comparisons revealed differences between all three conditions, such that constrastive conditions were answered more accurately than either new or wide focus conditions, and new conditions were answered accurately more than wide focus conditions.

Listeners' condition choice accuracy collapsed by focus location is plotted in Figure 16 (right). Listeners accurately chose the subject focus condition 93% of the time, the verb focus condition 94% of the time, and the object focus condition 87% of the time. An overall ANOVA on the three different types of focus location revealed significant differences across conditions, $F(2) = 37.53$, $p < .001$. Tukey post-hoc comparisons demonstrated that Listeners were more accurate to select focus on the subject ("Damon"), or the verb ("fried") than on the object ("omelet").

Discussion

The acoustic results of Experiment 2 performed only on the subject, verb, and object, once again demonstrated that speakers were consistent and successful in disambiguating the intended position of focus. Similar to Experiment 1, focused words were indicated with longer durations, a greater probability of post-word silence, higher intensity, higher mean F0, and higher maximum F0. Also, as demonstrated in Experiment 1, speakers did not systematically disambiguate information status with their productions, as evidenced by the discrim results of classification of only 50%.

The acoustic results of Experiment 2 performed on the subject, verb, object, and "I" indicated more systematic disambiguation of information status. Specifically, contrastive conditions were indicated by "I"'s produced with longer durations, higher intensity, and higher mean F0 and maximum F0.

The individual graphs of single acoustic features presented in Figure 13 are similar to those from Experiment 1, in that they demonstrate higher values of duration + silence, F0, and intensity on focused constituents than non-focused constituents. These results are different from those in Experiment 1, however, in the realization of F0 for the wide focus conditions. Specifically, the F0 values are higher for wide focus conditions compared to the other conditions.

Once again, the perception results demonstrated that Listeners were very accurate in determining which constituent of the sentence was focused. They were, however, not as accurate at determining the information status of the sentence as they were in the first study. This second result is to be expected since the Listeners no longer had the explicit cue of the "No" in the contrastive conditions to signal the difference between new and contrastive meanings.

Although we designed Experiment Two with the same words in each condition in order to encourage speakers to disambiguate between the new and contrastive readings of the focused constituents with the accent they placed on those constituents, it was also possible for speakers to disambiguate the discourse status of the focused constituent with how they prosodified the first three words of the sentence (i.e. "I heard that"). In fact, a discriminant function analysis demonstrated that the prosody of the first word ("I") of each sentence was the strongest predictors of the difference between new and contrastive information status.

Why might speakers have disambiguated information status most strongly with their production of "I heard that"? One possibility is that emphasizing the "I," which the discriminant results suggest that the speakers did, serves to signal pragmatically that the speaker means to contrast the information in the sentence that follows "I heard that" with what his/her questioner assumes. Whether or not this is a something the speakers would normally do to preface contrastive information is an open question, however, given that their task was explicitly to induce their listeners to correctly choose the new or contrastive condition.

The results of Experiment 2 once again do not provide evidence as to whether the lack of a difference between the acoustic realization of information status is a result of a general lack of such disambiguation in normal speech or whether, similar to Experiment 1, speakers did not disambiguate information status on the focused words because they did so on the "I heard that." This latter possibility was explored in Experiment 3, where speakers had only the sentence "Damon fried an omelet yesterday" available to realize acoustic differences between information statuses.

Experiment 3

Experiment Three was designed to be the strongest test of speakers' ability to disambiguate information status with prosody. That is, given the sentences they are required to speak, speakers may only disambiguate the discourse status of sentence constituents by producing different accents on those constituents. If speakers are truly able to disambiguate between "new" and "contrastive" information with different accents, we will see that reflected in their productions in Experiment 3.

Method

Participants

We recorded 17 pairs of subjects for this experiment. As before, we excluded speaker and listener pairs in which the Listener did not achieve accuracy greater than 20%. This resulted in the exclusion of two pairs. We also excluded one pair in which the speaker was not a native speaker of English. Finally, we excluded two pairs of subjects who produced unnatural prosody to effect a disambiguation between new and contrastive conditions. Specifically, they produced contrastive accents with unnaturally emphatic accents. These exclusions resulted in a total of 13 subjects whose productions and perceptions were analyzed.

Materials

The materials for Experiment 3 are identical to those from Experiment 2 described above save for the exclusion of "I heard that" from all conditions. An example item is presented in Table 4.

Condition	Status	Focused Argument	SetupQuestion	Target
1	New	wide	What happened yesterday?	Damon fried an omelet yesterday.
2	New	S	Who fried an omelet yesterday?	Damon fried an omelet yesterday.
3	New	V	What did Damon do to an omelet yesterday?	Damon fried an omelet yesterday.
4	New	O	What did Damon fry yesterday?	Damon fried an omelet yesterday.
5	Contrastive	S	Did Harry fry an omelet yesterday?	Damon fried an omelet yesterday.
6	Contrastive	V	Did Damon bake an omelet yesterday?	Damon fried an omelet yesterday.
7	Contrastive	O	Did Damon fry a chicken yesterday?	Damon fried an omelet yesterday..

Table 4: Example item from Experiment 3.

Procedure

The procedure for Experiment 3 was the same as that described for Experiment 1.

Results – Production

We tested the acoustic features we had identified in Experiment 1 on the new productions in Experiment 3. Once again, we conducted three discriminant analyses to determine whether the measures of (1) duration + silence, (2) maximum pitch, (3) mean pitch, and (4) maximum intensity on the three critical words in the sentence could predict (a) accent location, (b) information status, and (c) new vs. contrastive conditions.

Accent Location – correct trials

The overall Wilks's lambda was significant, $\Lambda = .095$, $\chi^2(24) = 1274.05$, $p < .001$, indicating that the acoustic measures could differentiate among the three accent locations. In addition, the residual Wilks's lambda was significant, $\Lambda = .329$, $\chi^2(11) = 602.50$, $p < .001$, indicating that some predictors could still differentiate accent location after partialling out the effects of the first function. Figure X indicates a separation of the focus locations on the discriminant functions.

When we tried to predict focus location, we were able to correctly classify 92.5% of the sentences in our sample. To assess how well the classification would perform on a new sample, we performed a leave-one-out classification and correctly classified 92.0% of our sentences.

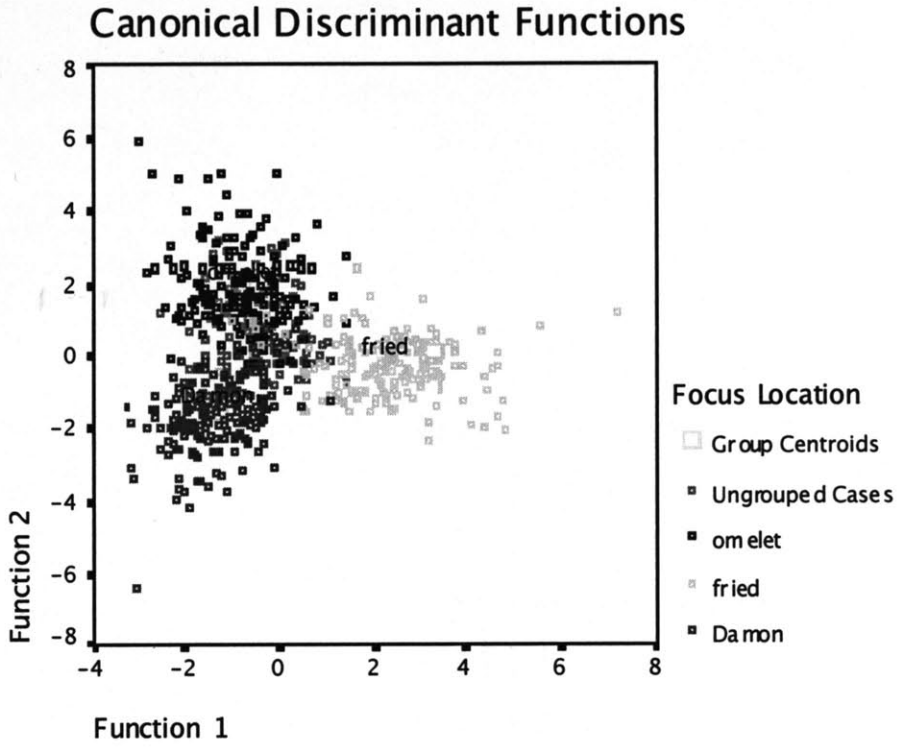


Figure 17: Separation of focus locations on two discriminant functions

Information Status – correct trials

The overall Wilks's lambda was significant, $\Lambda = .759$, $\chi^2(24) = 178.29$, $p < .001$, indicating that the acoustic measures could differentiate among the three information status conditions. In addition, the residual Wilks's lambda was significant, $\Lambda = .947$, $\chi^2(11) = 35.35$, $p < .001$, indicating that some predictors could still differentiate information status after partialling out the effects of the first function. Figure X indicates a separation of information status on the discriminant functions.

When we tried to predict information status, we were able to correctly classify 52.0% of the sentences in our sample. To assess how well the classification would perform on a new sample, we performed a leave-one-out classification and correctly classified 50.0% of our sentences.

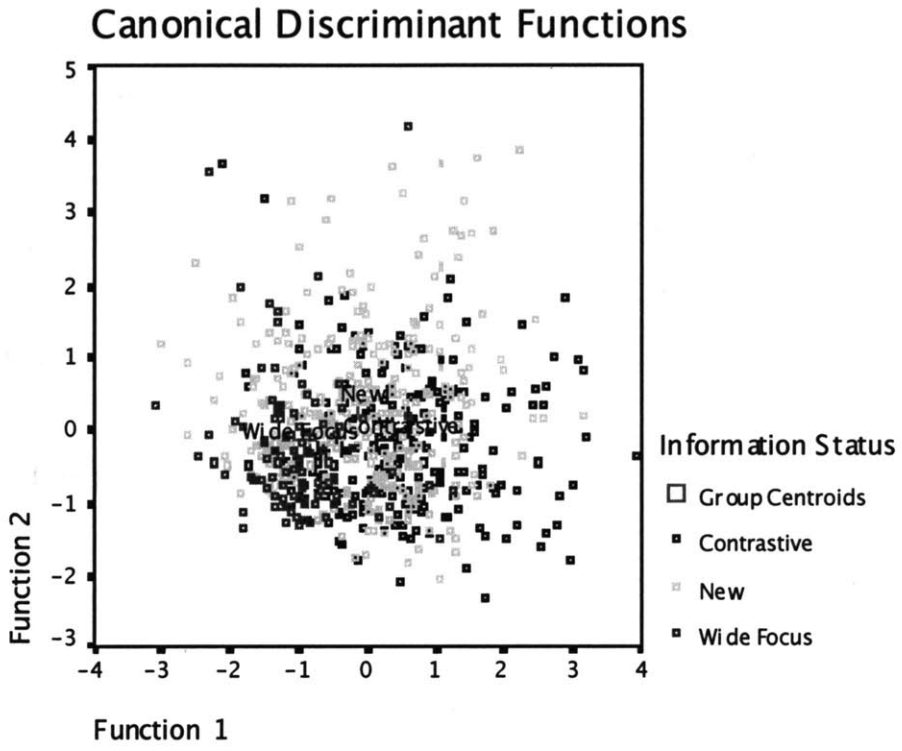


Figure 18: Separation of information status types on two discriminant functions

New vs. Contrastive – correct trials

The overall Wilks's lambda was significant, $\Lambda = .892$, $\chi^2(24) = 62.08$, $p < .001$, indicating that the acoustic measures could differentiate among new and contrastive conditions. When we tried to predict new vs. contrastive information status, we were able to correctly classify 64.2% of the sentences in our sample. To assess how well the classification would perform on a new sample, we performed a leave-one-out classification and correctly classified 62.4% of our sentences.

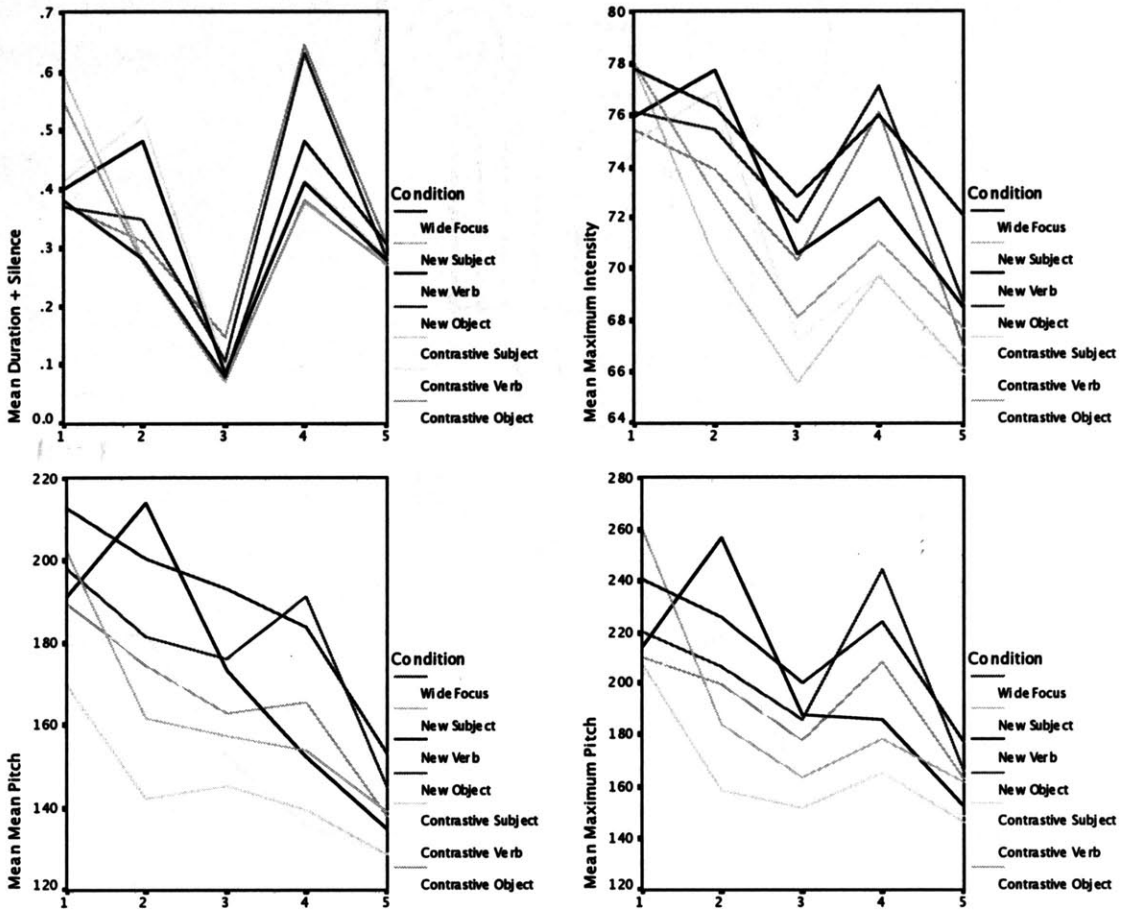


Figure 19: Average value of four acoustic features across every word and every condition in Experiment 3. The top left graph indicates the average of the sum of the duration of each word and the duration of any following silence, in seconds. The top right graph indicates the average maximum intensity in decibels. The bottom left graph indicates the average mean pitch, in Hertz. The bottom right graph indicates the average maximum pitch, in Hertz.

Results – Perception

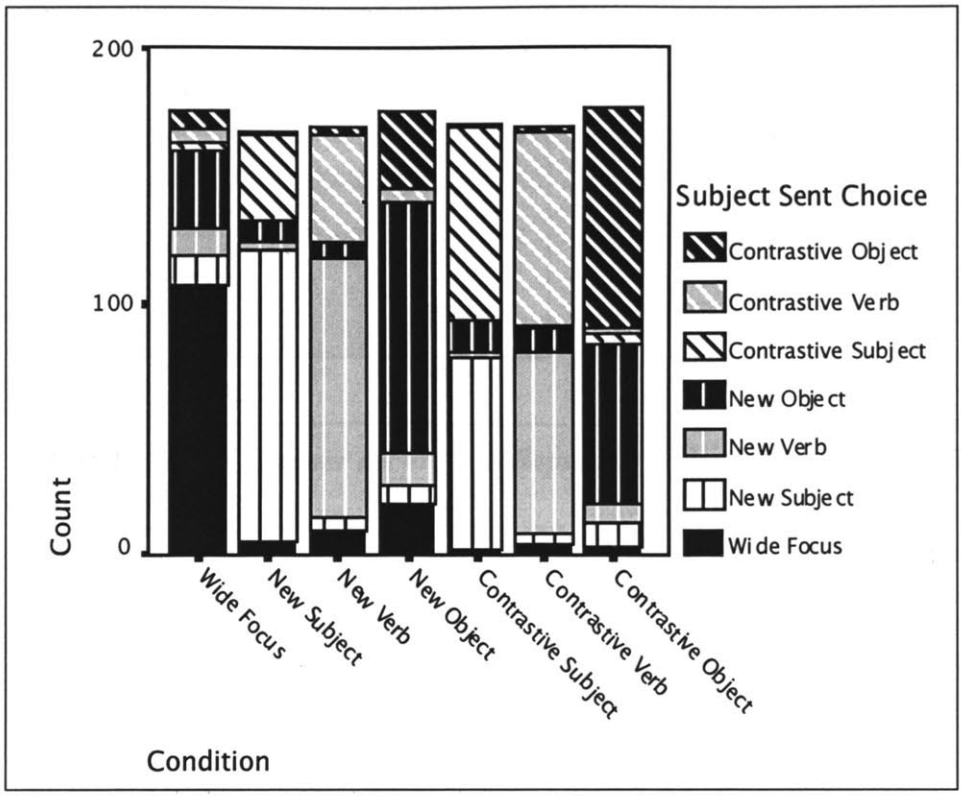


Figure 20: Total count of Listeners' condition choice by sentence type

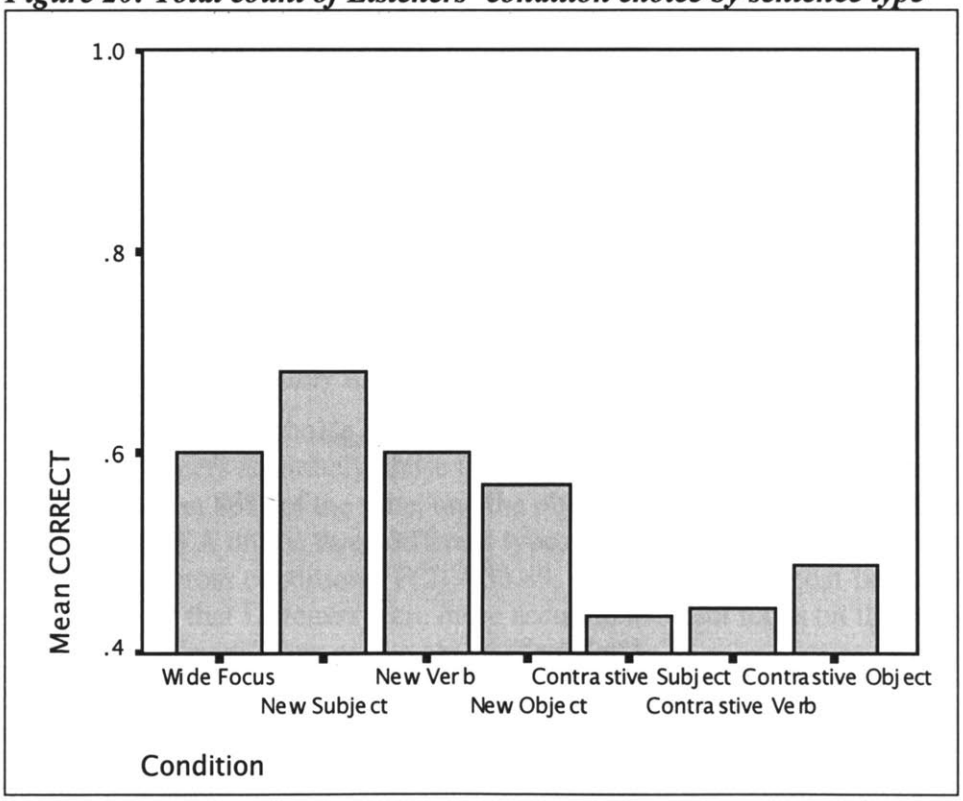


Figure 21: Mean Listener accuracy by condition

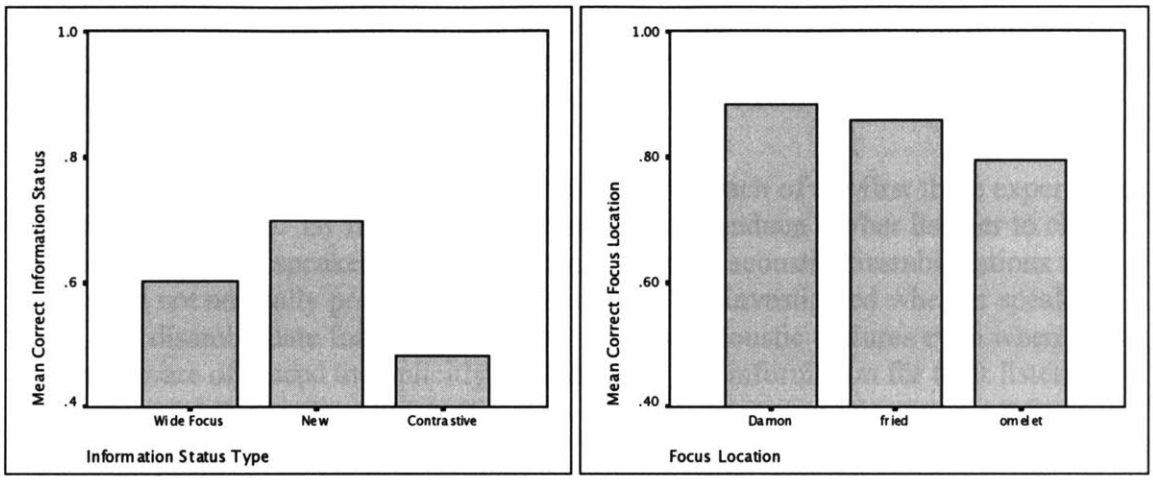


Figure 22: Mean Listener accuracy collapsed by Information Status (left) and focus location (right).

Listeners' choices of question sorted by the intended question are plotted in Figure 20, and their overall accuracy percentage by condition is plotted in Figure 21. Listeners' overall accuracy was 55%. An omnibus ANOVA on accuracy means by condition demonstrated a significant effect of condition, such that some conditions were answered more accurately than others, $F(6) = 4.90$, $p < .001$. Tukey post-hoc comparisons revealed that Listener accuracies for the three contrastive conditions were all lower than the the New subject condition; no other comparisons reached significance. Individual subject accuracy ranged from 39-96%, and there were significant differences between listeners, $F(9) = 9.9$, $p<.001$. There were no significant differences in accuracy across items, $F(13) < 1$.

Listeners' condition choice accuracy collapsed by information status is plotted in Figure 22 (left). Listeners accurately chose the wide focus condition 60% of the time, the new focus condition 70% of the time, and the contrastive focus condition 48% of the time. An overall ANOVA on the three different types of information status revealed significant differences across conditions, $F(2) = 157.82$ $p < .001$. Tukey post-hoc comparisons revealed differences between all three conditions, such that constrastive conditions were answered more accurately than either new or wide focus conditions, and new conditions were answered accurately more than wide focus conditions.

Listeners' condition choice accuracy collapsed by focus location is plotted in Figure 22 (right). Listeners accurately chose the subject focus condition 88% of the time, the verb focus condition 86% of the time, and the object focus condition 80% of the time. An overall ANOVA on the three different types of focus location revealed significant differences across conditions, $F(2) = 33.49$, $p < .001$. Tukey post-hoc comparisons demonstrated that Listeners were more accurate to select focus on the subject ("Damon"), or the verb ("fried") than on the object ("omelet").

Discussion

Focus location

In the third experiment, we observed the same high accuracy in determining the focus location of the accent in the sentence by both listeners and by a discriminant function analysis using duration + silence, maximum pitch, mean pitch, and maximum intensity. These results suggest, once again, that speakers are consistently indicating

focus location using this set of acoustic features, and that listeners are able to interpret these cues as indicating the location of sentence focus. Moreover, these results suggest that different speakers are using the same set of features across multiple experiments to indicate focus location.

It must be noted, however, that the task used in each of the first three experiments was not a natural one. By making the speaker's goal to induce his/her listener to choose the correct question, speakers may have been producing acoustic disambiguations that they would not normally produce. In Experiment 4, we investigated whether speakers would still disambiguate focus location with the same acoustic features even when they were not aware of a need to explicitly disambiguate this information for their listener.

Information status

The results of Experiment 3 once again demonstrate that speakers were not systematically disambiguating information status with their productions. In fact, across Experiments 1, 2, and 3, speakers overall did not produce acoustic cues to information status disambiguation which allowed the discriminant function analysis to classify information status with better than 50-52% accuracy across all experiments. These results suggest that speakers did not alter the way they produced the subject, verb, and object of the sentence whether or not there were other ways to disambiguate the sentence (e.g. with "No" in Experiment 1 or with "I heard that" in Experiment 2).

These results, however, do not lead to the conclusion that speakers do not acoustically disambiguate information status. As previously mentioned, the task may not have elicited speakers' normal production behavior. Specifically, the fact that the speakers' task was to induce his/her listener to choose the correct question, speakers may have settled on a strategy—any strategy—that would allow their listeners to successfully choose the right question. Different speakers may have pursued very different strategies which, although allowing their individual listener to perform accurately, did not overall result in systematic acoustic disambiguation of information status.

We investigated this possibility by performing a discriminant function analysis on the two best Speakers from Experiment 3. The best Speakers were those whose Listeners performed with the highest accuracy, averaging 83% accuracy. The classification accuracy achieved by the discriminant analysis is presented in Figure 21. The discriminant function analysis was able to correctly classify new focus 76% of the time, contrastive focus 90% of the time, and wide focus 89% of the time. The associated Listener classification accuracy is presented in Figure 22. Listeners of the two best Speaker were able to correctly classify new focus 77% of the time, contrastive focus 95% of the time, and wide focus 64% of the time.

The results of analyses performed on the best speakers from Experiment 3 suggest that, under certain circumstances, speakers do systematically disambiguate information status acoustically. It may be the case that this disambiguation is not consistent across speakers, and so results collapsed over multiple speakers washes out the disambiguating cues utilized by that individual speakers.

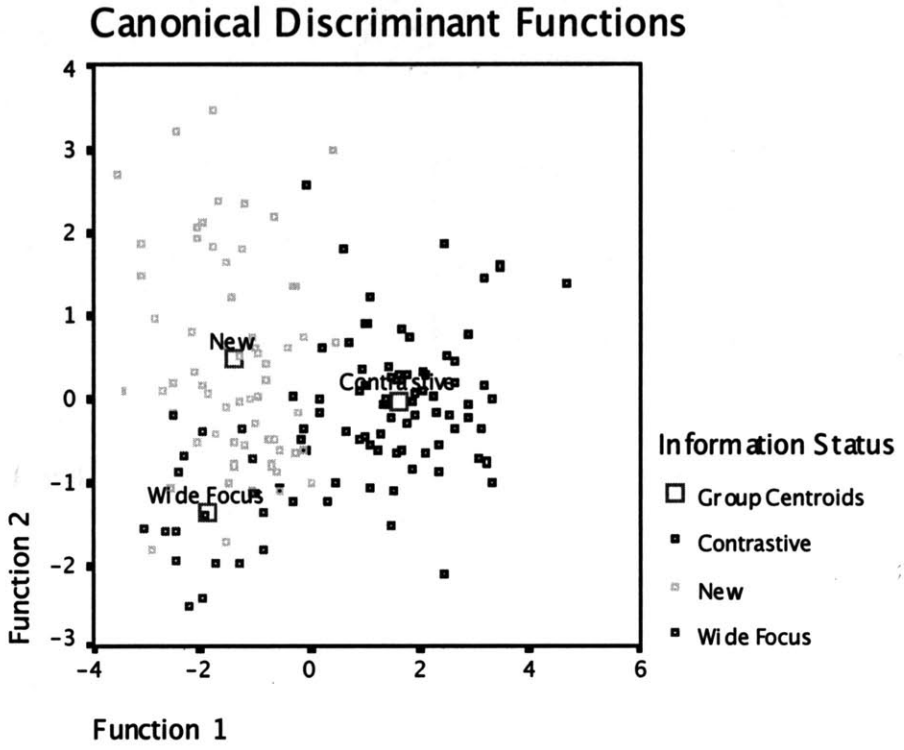


Figure 23: Separation of information status productions of the two best speakers from Experiment 3 on two discriminant functions

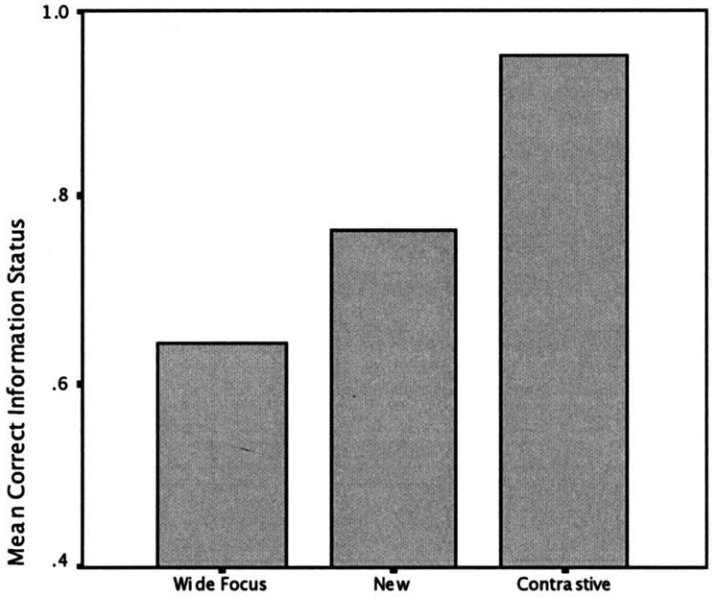


Figure 24: Listener accuracy on information status conditions from the best two Speakers from Experiment 3.

Experiment 4

The first three experiments that we conducted are valuable in that they constitute several major methodological advantages over previous work investigating the production and comprehension of accents in English. First, they constitute one of the first attempts to investigate whether and how naïve speakers disambiguate focus and discourse status of

sentence constituents, as well as whether naïve listeners are sensitive to such disambiguation. Second, they represent one of the first systematic evaluations of the acoustic features of English accents produced by naïve speakers.

It is important to note, however, that the productions from speakers in the first three experiments are not naturally, spontaneously produced. In order to maximize the amount of data we could gather from a small number of subjects, we chose to utilize a within-subjects design and to provide the words that we intended the speakers to produce. Both of these decisions could be argued to take away from the applicability of the results. Specifically, the fact that speakers were aware of the seven possible interpretations of each sentence, and of their goal of providing their listener with means to chose the correct question, encouraged them to maximally differentiate their productions to the best of their ability. Furthermore, the speakers always knew the words they were going to produce before they produced them.

To account for these limitations of the first three experiments, the fourth experiment in this sequence was designed to investigate accent production in more naturalistic productions. To the extent that the speakers' behavior in Experiment 4 mimics that of the speakers in Experiments 1-3, we can consider that behavior to be spontaneous, and not the result of unnatural communication pressures we placed on our subjects.

Method

Participants

We recorded seven pairs of subjects in Experiment 4, for a total of 14 speakers. Two speakers had to be replaced, because they failed to produce the correct words on more than 25% of trials. Subjects were paid $10/hour for their participation.

Materials

Condition	Status	Focused Argument	SetupQuestion	Target
1	New	wide	What happened last night?	Lena fried an omelet last night.
2	New	S	Who fried an omelet last night?	Lena fried an omelet last night.
3	New	V	What did Lena do to an omelet last night?	Lena fried an omelet last night.
4	New	O	What did Lena fry last night?	Lena fried an omelet last night.
5	Contrastive	S	Did Harry fry an omelet last night?	Lena fried an omelet last night.
6	Contrastive	V	Did Lena bake an omelet last night?	Lena fried an omelet last night.
7	Contrastive	O	Did Lena fry a chicken last night?	Lena fried an omelet last night.

Table 5: Example Item from Experiment 4.

Procedure

The experiment was conducted in two parts. The first part was a training session, where participants learned the correct names for pictures of people, actions, and objects. In the second part of the experiment, the two participants took turns serving as the Questioner and the Speaker. The Questioner was produced question for the Speaker, and then the Speaker produced the answer to the question which was indicated by pictures

presented on his/her screen. The next sections will describe each portion of the experiment in more detail.

Picture Naming Training

In a preliminary training session, both participants learned the mapping between 96 pictures and names, so that they could produce the names from memory during the main experiment. In a power point presentation, each picture, corresponding to a person, an action, or an object, was presented with its intended name. In order to facilitate memorization, the pictures were presented in alphabetical order. Participants were instructed to go through the power point at their own pace, with the goal of learning the mappings.

When they felt they had learned the mappings, participants waere given a picture-naming test, which consisted of 27 items from the full list of 96. Participants were told of their mistakes, and, if they incorrectly named four or more of the items on the test, they were instructed to go back through the power point to improve their memory of the picture-name mappings.

Question-Answer Experiment

After both members of the pair had successfully learned the picture-name mappings, the members were randomly assigned to the Questioner and Speaker roles. They sat at computers in the same room such that neither could see the other's screen.

The Questioner saw a question on his/her screen which he produced aloud for the Speaker. The Speaker then used the pictures on his/her screen to answer the Questioner.

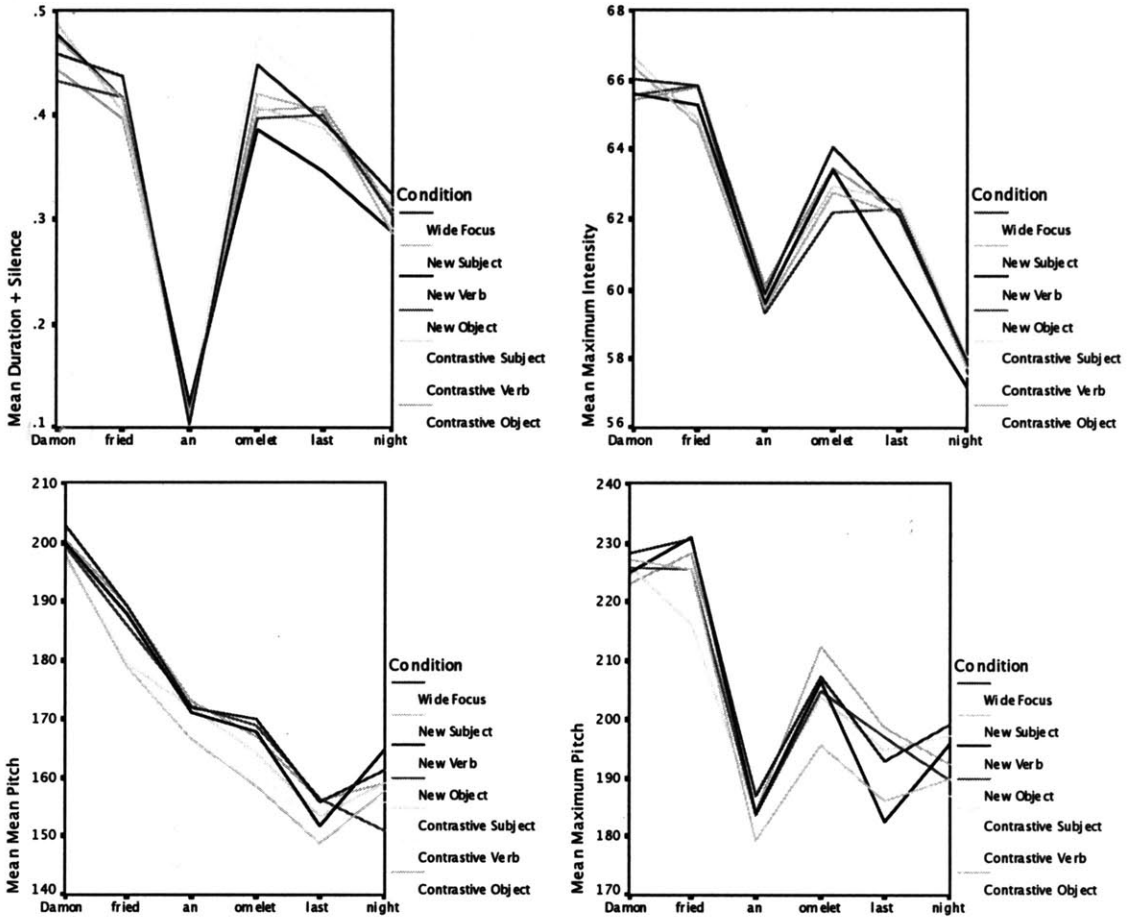


Figure 25: Average value of four acoustic features across every word and every condition in Experiment 3. The top left graph indicates the average of the sum of the duration of each word and the duration of any following silence, in seconds. The top right graph indicates the average maximum intensity in decibels. The bottom left graph indicates the average mean pitch, in Hertz. The bottom right graph indicates the average maximum pitch, in Hertz.

As in the previous experiments, three discriminant function analyses were conducted to determine whether the measures of (1) duration + silence, (2) maximum pitch, (3) mean pitch, and (4) maximum intensity on the three critical words in the sentence could predict (a) accent location, (b) information status, and (c) new vs. contrastive.

Accent Location

The overall Wilks's lambda was significant, $\Lambda = .67$, $\chi^2(24) = 165.44$, $p < .001$, indicating that the acoustic measures could differentiated among the three accent locations. In addition, the residual Wilks's lambda was significant, $\Lambda = .92$, $\chi^2(11) = 35.76$, $p < .001$, indicating that some predictors could still differentiate accent location after partialling out the effects of the first function. Figure X indicates a separation of the focus locations on the discriminant functions.

When we tried to predict focus location, we were able to correctly classify 60% of the sentences in our sample. The kappa value of .40 indicated moderate accuracy in classification performance. To assess how well the classification would perform on a

new sample, we performed a leave-one-out classification and correctly classified 58% of our sentences. At individual sentence locations, the discrim function was able to correctly classify subject focus 69% of the time, verb focus 51% of the time, and object focus 60% of the time.

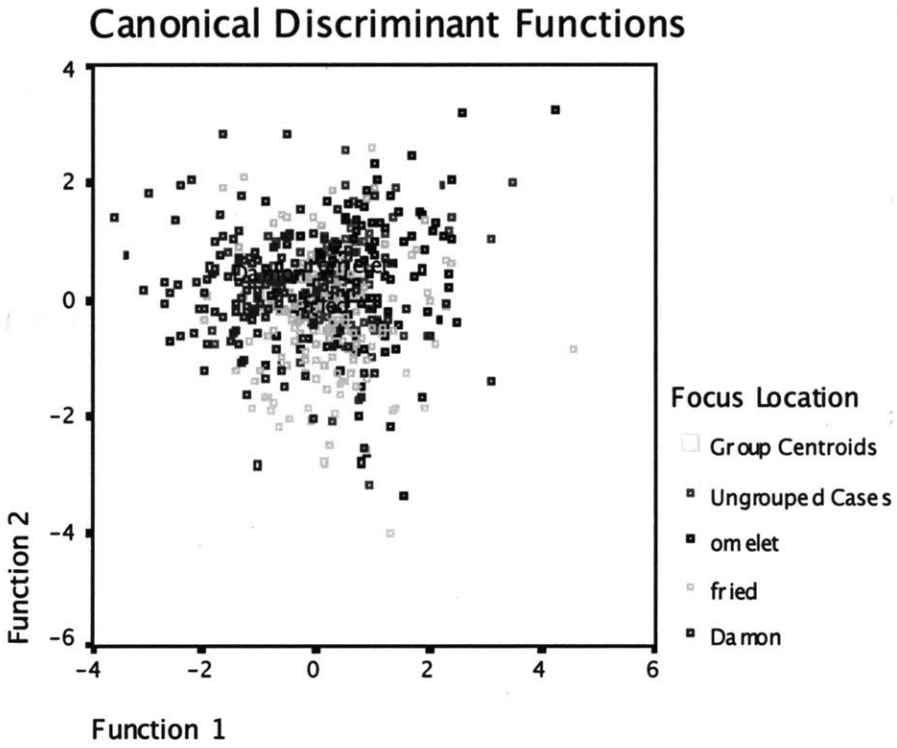


Figure 26: Separation of focus locations on two discriminant functions in Experiment 4.

Information Status

The overall Wilks's Lambda was significant, $\Lambda = .93$, χ^2 (24) = 37.47, $p < .05$, indicating that the acoustic features could discriminate between the three information status conditions. Figure 27 indicates the separation of information status groups on the two discriminant functions.

When we tried to classify information status, we were able to correctly classify 42% of cases. The kappa value of -.11, however, indicated that agreement was lower than what would be expected by chance, and therefore not reliable. In fact, when we estimated the percent of new cases that would be correctly classified, using a leave-one-out technique, we correctly classified only 36% of cases. For individual levels of information status, the discrim function was able to correctly classify new focus 52% of the time, contrastive focus 27% of the time, and wide focus 60% of the time.

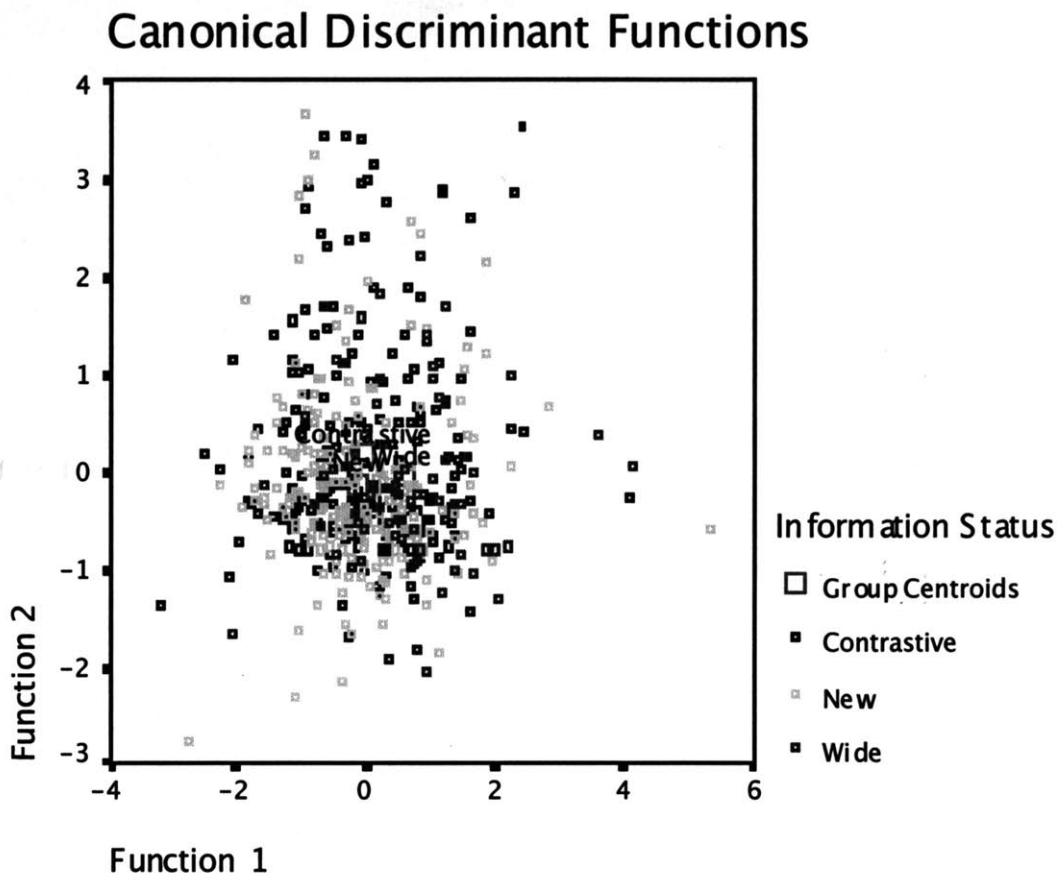


Figure 27: Separation of information status on two discriminant functions in Experiment 4.

New vs. Contrastive

The overall Wilks's Lambda was not significant, $\Lambda = .98$, $\chi^2\ (12) = 10.61$, $p = .56$, indicating that the acoustic features selected could not discriminate between new and contrastive sentences.

Discussion

Focus Location

As in the previous experiments, speakers did consistently provide acoustic cues which served to disambiguate focus location for their listeners. In addition, they indicated focus with the same cues they had used in previous experiments: increased duration, higher intensity, higher mean F0, and higher maximum F0. These results are noteworthy in that, unlike in the previous experiments, the speakers in the current experiment were not explicitly trying to induce their listeners to choose the correct meaning. Therefore, these results suggest that the acoustic cues we identified are those which are normally used by speakers in natural conversation to indicate the focused material in a sentence.

Information Status

The results from Experiment 4 indicate, once again, that speakers do not systematically differentiate between information status conditions. Specifically, a

discriminant function analysis could not classify speakers' productions by information status any better than chance would predict. There are at least two possible explanations for this result. The first, as previously suggested, is that speakers do not systematically differentiate information status with prosody. As previously stated, it could be the case that different speakers have different ways of encoding information status with their prosody that, when averaged with the prosody of other speakers, can no longer effectively disambiguate between conditions. The second possible explanation is that the task did not mimic natural communication pressures that would induce speakers to produce differences in prosody.

It should be noted that the data from Experiment 4 were not pre-screened in the same way as those productions from the first three experiments. Specifically, because there was no listener engaged in a meaning task in Experiment 4, there is no way to select (a) only those speakers who were engaged in the task and were prosodically differentiating their productions, or (b) those individual productions which bear meaning. Future analyses of these data should be performed on only those trials which are shown to convey information about focus location or information status to a new set of Listeners.

General Discussion

The results from the four experiments presented in this chapter represent important contributions to basic research on the information that is conveyed by prosody. Specifically, these data demonstrate that speakers do systematically provide cues to the location of focused material with prosody, and that they use a combination of duration, intensity, and pitch indicate this information. Furthermore, speakers provide cues to focus location whether or not the task explicitly demands it.

The second important result from this series of experiments is that that speakers do not systematically provide cues to information status with prosody. That is, speakers do not indicate the new or contrastive status of focused words consistently. In addition, it appears that speakers are not any more or less likely to provide acoustic cues to information status when the context in which the information occurs contains additional cues to information status. Specifically, whether or not the contrastive sentences were preceded with a "No" did not influence the overall usefulness of acoustic cues to contrastive focus.

In addition, these experiments represent a significant improvement over previous investigations of the relationship between prosody and meaning for the following reasons: First, we did not exclude speakers based on our perceptions of their productions. Speakers who were excluded were either not providing information to their Listeners, or were obviously not taking the task seriously. Second, we investigated the acoustic realization of different sentence positions, and multiple items, to ensure that differences we observed were not limited to certain syntactic positions or to a limited set of materials. Third, we utilized multiple, untrained speakers to ensure that our results are generalizeable to all speakers and are not due to speakers prior beliefs about what type of accent signals a particular type of information. Finally, we elicited and selected for analysis productions using a meaning task, rather than basing differences on perceptual differentiability or on ratings of the appropriateness of certain prosodic contours for particular purposes.

These studies open the door to future investigations of the ways in which information status is realized in speech and to whether different accent categories indicate any differences in meaning. Are there any circumstances under which speakers will produce systematic cues to new and contrastive status? If so, are these different information types realized with consistent acoustic cues, or consistent intonational shapes? If, on the other hand, there are no consistent differences between accents that mark new and contrastive information status, is there support for any categories of accents, or do accents vary continuously depending on multiple discourse, speaker, and context factors? Accents may be only one of many cues to information status that cannot be interpreted in the absence of their full linguistic context.

References

Bartels, C., Kingston, J., (1994). Salient Pitch Cues in the Perception of Contrastive Focus. *The Journal of the Acoustical Society of America*, Volume 95, Issue 5, 2973-

Beaver, D., Clark, B., Flemming, E., Jaeger, T.F., & Maria Wolters, 2007. When Semantics meets Phonetics: Acoustical studies of second occurrence focus. *Language*, 83(2), 245-276.

Beckman, M. and Hirschberg, J. (1994) The ToBI annotation conventions. Technical report, Ohio State University, Columbus.

Beckman, M. and Pierrehumbert, J. (1986) Intonational structure in Japanese and English. *Phonology Yearbook*, 3, 255-309.

Beckman, M. E., Hirschberg, J., & Shattuck-Hufnagel, S. (2005). The original ToBI system and the evolution of the ToBI framework. In S.-A. Jun (ed.) Prosodic Typology -- The Phonology of Intonation and Phrasing, Oxford University Press, Chapter 2: 9-54.

Beckman, M., and Ayers-Elam, G. (1994) Guidelines for ToBI Labelling. v. 2. Ohio State University. Available at www.ling.ohio-state.edu/research/ phonetics/E_ToBI/.

Birch, S. and Clifton, C. Jr. (1995) Focus, accent, and argument structure: effects on language comprehension. *Language and Speech,* 38 (4), 365-391.

Bock & Mazzella (1983) Intonational marking of given and new information: Some consequences for comprehension. *Memory and Cognition*, 11, 64-76.

Boersma, Paul & Weenink, David (2006). Praat: doing phonetics by computer (Version 4.3.10) [Computer program]. Retrieved June 3, 2005, from http://www.praat.org/

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

Bolinger, D. (1961). Contrastive accent and contrastive stress. *Language,* 37, 83-96.

Brugos, A., Shattuck-Hufnagel, S, & Veilleux, N. (2006) Transcribing Prosodic Structure of Spoken Utterances with ToBI, January (IAP) 2006, chapter 2.12 http://ocw.mit.edu/OcwWeb/Electrical-Engineering-and-Computer-Science/6-911January--IAP--2006/CourseHome/index.htm

Calhoun, S. (2004). Phonetic dimensions of intonational categories: the case of L+H* and H*. In *Proceedings of the International Conference on Spoken Language Processing*, Nara: Japan.

Calhoun, S. (2005). It's the difference that matters: An argument for contextually-grounded acoustic intonational phonology. In *Linguistics Society of America Annual Meeting*, Oakland, California.

Carlson, K., Clifton, C., Jr., & Frazier, L. (2001). Prosodic boundaries in adjunct attachment. *Journal of Memory and Language*, 45, 58–81.

Choi, J., Hasegawa-Johnson, M., & Cole, J. (2005). Finding intonational boundaries using acoustic cues related to the voice source. *Journal of Acoustical Society of America*, 188 (4), 2579-2587.

Cooper, W., Eady, S. & Mueller, P. (1985) Acoustical aspects of contrastive stress in question-answer contexts. *Journal of Acoustical Society of America*, 77(6), 2142-2156.

Cooper, W. E., & Paccia-Cooper, J. (1980). Syntax and speech. Cambridge, MA: Harvard University Press.

Cutler, A. (1977). The Context-Independence of "Intonational Meaning". *Chicago Linguistic Society (CLS 13),* 104-115.

Dahan, D., Tanenhaus, M. K., & Chambers, C. G. (2002). Accent and reference resolution in spoken-language comprehension. *Journal of Memory & Language Language*, 47, 292-314.

Dilley, L. (2005) The phonetics and phonology of tonal systems, PhD thesis, MIT.

Dilley, L. and Brown, M. (2005) The RaP Labeling System, v. 1.0. Available at http://faculty.psy.ohio-state.edu/pitt/dilley/rap-system.htm.

Dilley, L., Breen, M., Gibson, E., Bolivar, M., and Kraemer, J. (2006) A comparison of inter-coder reliability for two systems of prosodic transcriptions: RaP (Rhythm and Pitch) and ToBI (Tones and Break Indices). *Proceedings of the International Conference on Spoken Language Processing*, Pittsburgh, PA.

Eady, S. J., & Cooper, W. E. (1986). Speech intonation and focus location in matched statements and questions. Journal of the Acoustical Society of America, 80,402-415.

Ferreira, F. (1988). Planning and timing in sentence production: The syntax-to-phonology conversion. Unpublished dissertation, University of Massachusetts, Amherst, MA.

Ferreira, F. (1991). Effects of length and syntactic complexity on initiation times for prepared utterances. *Journal of Memory and Language*, 30, 210–233.

Fox Tree, J. (1995). The effect of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language*, 34, 709-738.

Gee, J. P., & Grosjean, F. (1983). Performance structures. A psycholinguistic and linguistic appraisal. *Cognitive Psychology*, 15, 411–458.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68, 1-76.

Gordon, P. C., Hendrick, R, & Johnson, M (2001). Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27, 1411-1423.

Greenberg, S., Carvey, H., & Hitchcock, L. (2002). The relation between stress accent and pronunciation variation in spontaneous American English discourse. In *Proceedings of the ISCA Workshop on Prosody and Speech Processing.*

Grodner, D. and Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentential complexity. *Cognitive Science*, 29, 261-290.

Grodner, D. J., Gibson, E., & Watson, D. (2005). The influence of contextual contrast on syntactic processing: Evidence for strong-interaction on sentence comprehension. *Cognition*, 95, 275-296.

Gussenhoven, C. (1983).Testing the reality of focus domains. *Language and Speech*, 26, 61–80.

Gussenhoven, C. (1999). On the limits of focus projection in English. In P. Bosch & R. van der Sandt (Eds.), *Focus: Linguistic, cognitive, and computational perspectives* (pp.

43 –55). Cambridge, U.K.: Cambridge University Press.

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

Hirschberg, J. (1995) Prosodic and acoustic cues to spontaneous and read speech. In *Proceedings of International Congress on Phonetic Sciences,* 2, 36-43.

Ito, K. Speer, S. R. and Beckman, M. E. (2004) Informational status and pitch accent distribution in spontaneous dialogues in English, In *Proceedings of the International Conference on Spoken Language Processing*, Nara: Japan, 279-282.

Landis, R. & Koch, G. (1977) The Measurement of Observer Agreement for Categorical Data. *Biometrics*, Vol. 33, No. 1., pp. 159-174.

Kjelgaard, M., & Speer, S. (1999). Prosodic facilitation and interference in the resolution of temporary syntactic ambiguity. *Journal of Memory and Language*, 40, 153–194.

Kochanski, G., Grabe, E., Coleman, J., & Rosner, B. (2005) Loudness predicts prominence: fundamental frequency lends little. *The Journal of the Acoustical Society of America*, 118 (2), 1038-1054.

Krahmer, E., & Swerts, M. (1998). Reconciling two competing views on contrastiveness. *Speech Communication*, 34, 391-405.

Kraljic, T. and S. E. Brennan (2005). Prosodic disambiguation of syntactic structure: For the speaker or for the addressee? *Cognitive Psychology* 50: 194-231.

Ladd, D. R., Schepman, A. (2003). Sagging transitions between high accent peaks in English: experimental evidence. *Journal of Phonetics*, 31, 81-112.

Lieberman, P. (1960). Some acoustic correlates of word stress in American English. *The Journal of the Acoustical Society of America*, 32(4), 451-454.

Linguistic Data Consortium (1997) CALLHOME American English Speech. http://www.ldc.upenn.edu.

Most, R. B. & Saltz, E. (1979) Information structure in sentences: New information. *Language and Speech*, 22, 89-95.

Nespor, M., & Vogel, I. (1986). Prosodic phonology. Dordrecht: Foris Publications.

Ostendorf, M.F., Price P. J., Shattuck-Hufnagel S. (1995) The Boston University Radio News Corpus. Technical Report No. ECS-95-001, Boston University.

Pierrehumbert, J., & Beckman, M. (1988). *Japanese Tone Structure*. Cambridge, MA: MIT Press.

Pierrehumbert, J. & Hirschberg, J. (1990). The Meaning of Intonational Contours in the Interpretation of Discourse. In: P. R. Cohen & J. Morgan & M. E. Pollack (eds.). *Intentions in Communication.* Cambridge/MA: MIT Press, 271-311.

Pierrehumbert, J.B. (1980). *The phonology and phonetics of English intonation.* Unpublished dissertation, MIT.

Pitrelli, J., Beckman, M. & Hirschberg, J. (1994) Evaluation of prosodic transcription labeling reliability in the ToBI framework. In *Proceedings of the International Conference on Spoken Language Processing*, 123-126.

Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S., & Fong, C. (1991). The use of prosody in syntactic disambiguation. *Journal of Acoustical Society of America*, 90, 2956–2970.

Price, Patti, Mari Ostendorf, Stefanie Shattuck-Hufnagel and C. Fong. 1991. The use of prosody in syntactic disambiguation. *Journal of the Acoustical Society of America,* 90: 2956-2970.

Schafer, A., Speer, S., Warren, P., & White, S. D. (2000). Intonational disambiguation in sentence production and comprehension. *Journal of Psycholinguistic Research,* 29(2): 169-182.

Selkirk, E. O. (1984). *Phonology and syntax: The relation between sound and structure.* Cambridge, MA: MIT Press.

Selkirk, E. (1995). Sentence Prosody: Intonation, Stress, and Phrasing. In: J.Goldsmith (ed.). *The Handbook of Phonological Theory.* Oxford: Blackwell, 550-569.

Silverman, K., Beckman, M., Pitrelli, J. Ostendorf, M. Wightman, C. Price, P., Pierrehumbert, J. & Hirschberg, J. (1992) ToBI: A standard for labeling English prosody. In Proceedings of the International Conference on Spoken Language Processing, 867-870.

Snedeker, J., & Trueswell, J. (2003). Using prosody to avoid ambiguity: Effects of speaker awareness and referential contest. *Journal of Memory and Language*, 48, 103–130.

Speer, S. R., Kjelgaard, M. M., & Dobroth, K. M. (1996). The influence of prosodic structure on the resolution of temporary syntactic closure ambiguities. *Journal of Psycholinguistic Research*, 25, 247–268.

Syrdal, A. and McGory, J. (2000) Inter-transcriber reliability of ToBI prosodic labeling. In *Proceedings of the International Conference on Spoken Language Processing*, Beijing: China, 235-238.

Taylor, P. (2000). Analysis and synthesis of intonation using the tilt model. *Journal of the Acoustical Society of America.* 107(3):1697-1714.

Terken and Noteboom (1987) Opposite Effects of Accentuation and Deaccentuation on Verification Latencies for Given and New Information. *Language and Cognitive Processes* 2(3/4), 145-163.

Watson, D. & Gibson, E. (2004). The relationship between intonational phrasing and syntactic structure in language production. *Language and Cognitive Processes*, 19, 713-755.

Watson, D., Breen, M., & Gibson, E. (2006). The role of syntactic obligatoriness in the production of intonational boundaries. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 32(5), 1045-1056.

Watson, D., Tanenhaus, M. & Gunlogson, C. (2004). Processing pitch accents: Interpreting H* and L+H* Presented at *the 17th Annual CUNY Conference on Human Sentence Processing*, Cambridge, MA.

Wheeldon, L. & Lahiri, A. (1997). Prosodic units in speech production. *Journal of Memory and Language*, 37, 356–381.

Wightman, C. (2002) "ToBI or not ToBI?" In Proceedings of the International Conference on Speech Prosody, Aix-en-Provence: France, 2002.

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. (1992) Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of Acoustical Society of America*, 91(3), 1707-1717.

Yoon, T., Chavarria, S., Cole, J., & Hasegawa-Johnson, M. (2004) Intertranscriber reliability of prosodic labeling on telephone conversation using ToBI. In *Proceedings of the International Conference on Spoken Language Processing*., Nara: Japan, 2729-2732.

Appendix A

1. The mobster paid the bounty (of thirty diamonds) to the gangster (with burly henchmen) quickly / before the crime was committed.
2. The caterer brought the pastries (with lemon filling) to the party (for Oscar winners) early / before the guests had arrived.
3. The gigolo sent a bouquet (of sixty roses) to the showgirl (from Hello Dolly) on Sunday / before the performance last night.
4. The colonel assigned the mission (of killing Castro) to the soldier (with sniper training) last night / last night at the Pentagon.
5. The wizard granted the powers (of magic healing) to the suitor (of England's princess) last night / after being threatened with death.
6. The matriarch left the necklace (with sapphire inlay) to the daughter (of peasant parents) secretly / before the family found out.
7. The director offered the payment (of thirty million) to the actor (of poignant dramas) yesterday / after filming had already begun.
8. The academy presented the award (of greatest import) to the actor (of little renown) on Sunday / last week in Los Angeles.
9. The executive delivered the statement (of corrupt actions) to the judges (of business conduct) regretfully / before a ruling was issued.
10. The professor assigned the chapter (on local history) to the students (of social science) yesterday / after the first midterm exam.
11. The writer pitched the story (of happy orphans) to the chairman (of Disney Studios) at lunch / over several drinks after lunch.
12. The student gave the basket (of chocolate brownies) to the teacher (of ancient history) today / before the start of vacation.
13. The lieutenant evacuated the soldiers (of several platoons) to a region (with unarmed locals) yesterday / after the mysterious phone call.
14. The girl attached the posters (of missing children) to the windows (of local buildings) today / after her shopping trip downtown.
15. The priest delivered the turkeys (with homemade stuffing) to the homeless (at local shelters) on Thursday / before people arrived for dinner.
16. The socialite donated the suitcase (of lovely dresses) to the woman (in dirty clothing) yesterday / after meeting her at church.
17. The lawyer left the duties (of mindless errands) to the partner (with lower status) this morning / after the lengthy conference call.
18. The girl lent the booklet (of practice exams) to the classmate (from second period) on Friday / before the test on Friday.
19. The gentleman sent the bouquet (of gorgeous roses) to the woman (with shiny lipstick) on Monday / after spotting her from afar.

20. The millionaire assigned a chauffeur (with little patience) to his mistress (in Southern Europe) today / after a quarrel on Friday.
21. The station offered the ballad (with minor changes) to the public (in nearby cities) last week / after the debate last week.
22. The grandmother gave the necklace (of twenty pearls) to the grandson (from Kansas City) on Sunday / at the annual family reunion.
23. The architect placed the statue (of Roger Sherman) in the courtyard (with pretty flowers) carefully / with tremendous pride and satisfaction.
24. The son put his backpack (with heavy textbooks) in the kitchen (with seven people) last night / without stopping to eat dinner.
25. The critic handed the letter (for Steven Spielberg) to the postman (with curly sideburns) personally / in the sunshine of morning.
26. The committee allocated the money (from Tuesday's auction) to the members (from Costa Rica) yesterday / after numerous hours of discussion.
27. The bride put the favors (of mini bouquets) on the tables (of several guests) happily / before the wedding reception began.
28. The spy told the secrets (of deadly weapons) to the leaders (of foreign nations) quietly / through a network of operatives.
29. The salesman conveyed his advice (on buying vases) to the clients (from rural Texas) on Friday / after a meeting on Friday.
30. The professor assigned a project (on Asian Studies) to his students (with heavy workloads) yesterday / without regard for other classes.
31. The tycoon lent the limo (with leather seating) to his buddies (from Swarthmore College) often / for several days last month.
32. The referee explained the format (of soccer contests) to the players (from Amherst College) on Friday / before the big tournament began.

Appendix B

1a. Context: What happened yesterday?
1b. Context: Who fried an omelet yesterday?
1c. Context: What did Damon do to an omelet yesterday?
1d. Context: What did Damon fry yesterday?
1e. Context: Did Harry fry an omelet yesterday?
1f. Context: Did Damon bake an omelet yesterday?
1g. Context: Did Damon fry a chicken yesterday?
Target: No, Damon fried an omelet yesterday.

2a. Context: What happened yesterday?
2b. Context: Who sold her diamond yesterday?
2c. Context: What did Megan do with her diamond yesterday?
2d. Context: What did Megan sell yesterday?
2e. Context: Did Jodi sell her diamond yesterday?
2f. Context: Did Megan lose her diamond yesterday?
2g. Context: Did Megan sell her sapphire yesterday?
Target: No, Megan sold her diamond yesterday.

3a. Context: What happened last night?
3b. Context: Who dried a platter last night?
3c. Context: What did Mother do to a platter last night?
3d. Context: What did Mother dry last night?
3e. Context: Did Daddy dry a platter last night?
3f. Context: Did Mother wash a platter last night?
3g. Context: Did Mother dry a bowl last night?
Target: No, Mother dried a platter last night.

4a. Context: What happened last night?
4b. Context: Who read an email last night?
4c. Context: What did Norman do with an email last night?
4d. Context: What did Norman read last night?
4e. Context: Did Kelly read an email last night?
4f. Context: Did Norman write an email last night?
4g. Context: Did Norman read a letter last night?
Target: No, Norman read an email last night.

5a. Context: What happened this morning?
5b. Context: Who poured a smoothie this morning?
5c. Context: What did Lauren do with a smoothie this morning?
5d. Context: What did Lauren pour this morning?
5e. Context: Did Judy pour a smoothie this morning?
5f. Context: Did Lauren drink a smoothie this morning?
5g. Context: Did Lauren pour a cocktail this morning?
Target: No, Lauren poured a smoothie this morning.

6a. Context: What happened this morning?
6b. Context: Who sewed her dolly this morning?

6c. Context: What did Nora do to her dolly this morning?
6d. Context: What did Nora sew this morning?
6e. Context: Did Jenny sew her dolly this morning?
6f. Context: Did Nora rip her dolly this morning?
6g. Context: Did Nora sew her blanket this morning?
Target: No, Nora sewed her dolly this morning.

7a. Context: What happened on Tuesday?
7b. Context: Who trimmed her eyebrows on Tuesday?
7c. Context: What did Molly do to her eyebrows on Tuesday?
7d. Context: What did Molly trim on Tuesday?
7e. Context: Did Sarah trim her eyebrows on Tuesday?
7f. Context: Did Molly wax her eyebrows on Tuesday?
7g. Context: Did Molly trim her hair on Tuesday?
Target: No, Molly trimmed her eyebrows on Tuesday.

8a. Context: What happened on Tuesday?
8b. Context: Who burned a candle on Tuesday?
8c. Context: What did Nolan do to a candle on Tuesday?
8d. Context: What did Nolan burn on Tuesday?
8e. Context: Did Steven burn a candle on Tuesday?
8f. Context: Did Norman break a candle on Tuesday?
8g. Context: Did Nolan burn a log on Tuesday?
Target: No, Nolan burned a candle on Tuesday.

9a. Context: What happened last week?
9b. Context: Who killed a termite last week?
9c. Context: What did Logan do to a termite last week?
9d. Context: What did Logan kill last week?
9e. Context: Did Billy kill a termite last week?
9f. Context: Did Logan trap a termite last week?
9g. Context: Did Logan kill a cockroach last week?
Target: No, Logan killed a termite last week.

10a. Context: What happened last week?
10b. Context: Who caught a bunny last week?
10c. Context: What did Radar do to a bunny last week?
10d. Context: What did Radar catch last week?
10e. Context: Did Fido catch a bunny last week?
10f. Context: Did Radar lick a bunny last week?
10g. Context: Did Radar catch a squirrel last week?
Target: No, Radar caught a bunny last week.

11a. Context: What happened on Sunday?
11b. Context: Who pulled a stroller on Sunday?
11c. Context: What did Darren do to a stroller on Sunday?
11d. Context: What did Darren pull on Sunday?
11e. Context: Did Maggie pull a stroller on Sunday?
11f. Context: Did Darren push a stroller on Sunday?
11g. Context: Did Darren pull a sled on Sunday?

Target: No, Darren pulled a stroller on Sunday.

12a. Context: What happened on Sunday?
12b. Context: Who peeled a carrot on Sunday?
12c. Context: What did Brandon do to a carrot on Sunday?
12d. Context: What did Brandon peel on Sunday?
12e. Context: Did Tommy peel a carrot on Sunday?
12f. Context: Did Brandon eat a carrot on Sunday?
12g. Context: Did Brandon peel a potato on Sunday?
Target: No, Brandon peeled a carrot on Sunday.

13a. Context: What happened on Friday?
13b. Context: Who cleaned a pillow on Friday?
13c. Context: What did Maren do to a pillow on Friday?
13d. Context: What did Maren clean on Friday?
13e. Context: Did Debbie clean a pillow on Friday?
13f. Context: Did Maren buy a pillow on Friday?
13g. Context: Did Maren clean a rug on Friday?
Target: No, Maren cleaned a pillow on Friday.

14a. Context: What happened on Friday?
14b. Context: Who fooled a bully on Friday?
14c. Context: What did Lindon do to a bully on Friday?
14d. Context: Who did Lindon fool on Friday?
14e. Context: Did Kelly fool a bully on Friday?
14f. Context: Did Lindon fight a bully on Friday?
14g. Context: Did Lindon fool a teacher on Friday?
Target: No, Lindon fooled a bully on Friday.

Appendix C

1a. Question: What happened last night?
1b. Question: Who fed a bunny last night?
1c. Question: What did Damon do to a bunny last night?
1d. Question: What did Damon feed last night?
1e. Question: Did Jenny feed a bunny last night?
1f. Question: Did Damon pet a bunny last night?
1g. Question: Did Damon feed a baby last night?
Response: Damon fed a bunny last night.

2a. Question: What happened last night?
2b. Question: Who caught a bunny last night?
2c. Question: What did Damon do to a bunny last night?
2d. Question: What did Damon catch last night?
2e. Question: Did Lauren catch a bunny last night?
2g. Question: Did Damon catch a squirrel last night?
Response: Damon caught a bunny last night.

3a. Question: What happened last night?
3b. Question: Who burned a candle last night?
3c. Question: What did Damon do to a candle last night?
3d. Question: What did Damon burn last night?
3e. Question: Did Molly burn a candle last night?
3f. Question: Did Damon break a candle last night?
3g. Question: Did Damon burn a log last night?
Response: Damon burned a candle last night.

4a. Question: What happened last night?
4b. Question: Who cleaned a carrot last night?
4c. Question: What did Darren do to a carrot last night?
4d. Question: What did Darren clean last night?
4e. Question: Did Lauren clean a carrot last night?
4f. Question: Did Darren eat a carrot last night?
4g. Question: Did Darren clean a chicken last night?
Response: Darren cleaned a carrot last night.

5a. Question: What happened last night?
5b. Question: Who peeled a carrot last night?
5c. Question: What did Darren do to a carrot last night?
5d. Question: What did Darren peel last night?
5e. Question: Did Molly peel a carrot last night?
5f. Question: Did Darren eat a carrot last night?
5g. Question: Did Darren peel a potato last night?
Response: Darren peeled a carrot last night.

6a. Question: What happened last night?
6b. Question: Who found a diamond last night?
6c. Question: What did Darren do to a diamond last night?

6d. Question: What did Darren find last night?
6e. Question: Did Nora find a diamond last night?
6f. Question: Did Darren buy a diamond last night?
6g. Question: Did Darren find a ring last night?
Response: Darren found a diamond last night.

7a. Question: What happened last night?
7b. Question: Who sold a diamond last night?
7c. Question: What did Darren do to a diamond last night?
7d. Question: What did Darren sell last night?
7e. Question: Did Jenny sell a diamond last night?
7f. Question: Did Darren lose a diamond last night?
7g. Question: Did Darren sell a sapphire last night?
Response: Darren sold a diamond last night.

8a. Question: What happened last night?
8b. Question: Who found a dollar last night?
8c. Question: What did Jenny do to a dollar last night?
8d. Question: What did Jenny find last night?
8e. Question: Did Damon find a dollar last night?
8f. Question: Did Jenny lose a dollar last night?
8g. Question: Did Jenny find a quarter last night?
Response: Jenny found a dollar last night.

9a. Question: What happened last night?
9b. Question: Who sewed a dolly last night?
9c. Question: What did Jenny do to a dolly last night?
9d. Question: What did Jenny sew last night?
9e. Question: Did Darren sew a dolly last night?
9f. Question: Did Jenny rip a dolly last night?
9g. Question: Did Jenny sew a blanket last night?
Response: Jenny sewed a dolly last night.

10a. Question: What happened last night?
10b. Question: Who read an email last night?
10c. Question: What did Jenny do to an email last night?
10d. Question: What did Jenny read last night?
10e. Question: Did Logan read an email last night?
10f. Question: Did Jenny open an email last night?
10g. Question: Did Jenny read a letter last night?
Response: Jenny read an email last night.

11a. Question: What happened last night?
11b. Question: Who smelled a flower last night?
11c. Question: What did Jenny do to a flower last night?
11d. Question: What did Jenny smell last night?
11e. Question: Did Nolan smell a flower last night?
11f. Question: Did Jenny plant a flower last night?
11g. Question: Did Jenny smell a skunk last night?
Response: Jenny smelled a flower last night.

12a. Question: What happened last night?
12b. Question: Who burned a letter last night?
12c. Question: What did Lauren do to a letter last night?
12d. Question: What did Lauren burn last night?
12e. Question: Did Darren burn a letter last night?
12f. Question: Did Lauren write a letter last night?
12g. Question: Did Lauren burn a magazine last night?
Response: Lauren burned a letter last night.

13a. Question: What happened last night?
13b. Question: Who mailed a letter last night?
13c. Question: What did Lauren do to a letter last night?
13d. Question: What did Lauren mail last night?
13e. Question: Did Logan mail a letter last night?
13f. Question: Did Lauren open a letter last night?
13g. Question: Did Lauren mail a package last night?
Response: Lauren mailed a letter last night.

14a. Question: What happened last night?
14b. Question: Who read a novel last night?
14c. Question: What did Lauren do to a novel last night?
14d. Question: What did Lauren read last night?
14e. Question: Did Nolan read a novel last night?
14f. Question: Did Lauren write a novel last night?
14g. Question: Did Lauren read a newspaper last night?
Response: Lauren read a novel last night.

15a. Question: What happened last night?
15b. Question: Who fried an omelet last night?
15c. Question: What did Lauren do to an omelet last night?
15d. Question: What did Lauren fry last night?
15e. Question: Did Damon fry an omelet last night?
15f. Question: Did Lauren bake an omelet last night?
15g. Question: Did Lauren fry a chicken last night?
Response: Lauren fried an omelet last night.

16a. Question: What happened last night?
16b. Question: Who peeled an onion last night?
16c. Question: What did Logan do to an onion last night?
16d. Question: What did Logan peel last night?
16e. Question: Did Molly peel an onion last night?
16f. Question: Did Logan chop an onion last night?
16g. Question: Did Logan peel an apple last night?
Response: Logan peeled an onion last night.

17a. Question: What happened last night?
17b. Question: Who fried an onion last night?
17c. Question: What did Logan do to an onion last night?
17d. Question: What did Logan fry last night?

17e. Question: Did Nora fry an onion last night?
17f. Question: Did Logan chop an onion last night?
17g. Question: Did Logan fry a potato last night?
Response: Logan fried an onion last night.

18a. Question: What happened last night?
18b. Question: Who cleaned a pillow last night?
18c. Question: What did Logan do to a pillow last night?
18d. Question: What did Logan clean last night?
18e. Question: Did Jenny clean a pillow last night?
18f. Question: Did Logan buy a pillow last night?
18g. Question: Did Logan clean a rug last night?
Response: Logan cleaned a pillow last night.

19a. Question: What happened last night?
19b. Question: Who dried a platter last night?
19c. Question: What did Molly do to a platter last night?
19d. Question: What did Molly dry last night?
19e. Question: Did Logan dry a platter last night?
19f. Question: Did Molly wash a platter last night?
19g. Question: Did Molly dry a bowl last night?
Response: Molly dried a platter last night.

20a. Question: What happened last night?
20b. Question: Who sold a platter last night?
20c. Question: What did Molly do to a platter last night?
20d. Question: What did Molly sell last night?
20e. Question: Did Nolan sell a platter last night?
20f. Question: Did Molly find a platter last night?
20g. Question: Did Molly sell a vase last night?
Response: Molly sold a platter last night.

21a. Question: What happened last night?
21b. Question: Who poured a smoothie last night?
21c. Question: What did Molly do to a smoothie last night?
21d. Question: What did Molly pour last night?
21e. Question: Did Damon pour a smoothie last night?
21f. Question: Did Molly drink a smoothie last night?
21g. Question: Did Molly pour a cocktail last night?
Response: Molly poured a smoothie last night.

22a. Question: What happened last night?
22b. Question: Who pulled a stroller last night?
22c. Question: What did Nolan do to a stroller last night?
22d. Question: What did Nolan pull last night?
22e. Question: Did Nora pull a stroller last night?
22f. Question: Did Nolan push a stroller last night?
22g. Question: Did Nolan pull a sled last night?
Response: Nolan pulled a stroller last night.
23a. Question: What happened last night?

23b. Question: Who bought a stroller last night?
23c. Question: What did Nolan do to a stroller last night?
23d. Question: What did Nolan buy last night?
23e. Question: Did Jenny buy a stroller last night?
23f. Question: Did Nolan sell a stroller last night?
23g. Question: Did Nolan buy a wheelbarrow last night?
Response: Nolan bought a stroller last night.

24a. Question: What happened last night?
24b. Question: Who sewed a sweater last night?
24c. Question: What did Nolan do to a sweater last night?
24d. Question: What did Nolan sew last night?
24e. Question: Did Lauren sew a sweater last night?
24f. Question: Did Nolan knit a sweater last night?
24g. Question: Did Nolan sew a quilt last night?
Response: Nolan sewed a sweater last night.

25a. Question: What happened last night?
25b. Question: Who killed a termite last night?
25c. Question: What did Nora do to a termite last night?
25d. Question: What did Nora kill last night?
25e. Question: Did Nolan kill a termite last night?
25f. Question: Did Nora trap a termite last night?
25g. Question: Did Nora kill a cockroach last night?
Response: Nora killed a termite last night.

26a. Question: What happened last night?
26b. Question: Who changed a toddler last night?
26c. Question: What did Nora do to a toddler last night?
26d. Question: What did Nora change last night?
26e. Question: Did Damon change a toddler last night?
26f. Question: Did Nora wash a toddler last night?
26g. Question: Did Nora change a baby last night?
Response: Nora changed a toddler last night.

27a. Question: What happened last night?
27b. Question: Who fed a toddler last night?
27c. Question: What did Nora do to a toddler last night?
27d. Question: What did Nora feed last night?
27e. Question: Did Darren feed a toddler last night?
27f. Question: Did Nora dress a toddler last night?
27g. Question: Did Nora feed a bunny last night?
Response: Nora fed a toddler last night.

28a. Question: What happened last night?
28b. Question: Who pulled a wagon last night?
28c. Question: What did Nora do to a wagon last night?
28d. Question: What did Nora pull last night?
28e. Question: Did Logan pull a wagon last night?
28f. Question: Did Nora push a wagon last night?

28g. Question: Did Nora pull a wheelbarrow last night?
Response: Nora pulled a wagon last night.