Kiwako Ito and Shari R. Speer

Using interactive tasks to elicit natural dialogue

1. Introduction

Basic research into the relationship between intonation and speaker’s intentions about syntax and information structure addresses whether, when, and how speakers use prosodic information to signal linguistic and paralinguistic meaning. Speakers use prosody for a range of functions in communication: to mark the difference between immediately relevant vs. background information; to express contrast, contradiction, and correction; and to indicate the intended syntax of ambiguous utterances. In this paper, we review the methodologies used to examine intonation and sentence and discourse structure in language production studies. We focus on design conflicts that arise when experimental techniques need to generate a sufficient number of items for analysis, while simultaneously eliciting utterances that are representative of natural, conversational speech. We discuss the advantages and disadvantages inherent in (1) using spontaneous speech vs. various text-based elicitation techniques, (2) the effect of different levels of detail in instructions to speakers, and (3) the effects of manipulating the experimental environment. In the final sections, we present details of the method used in our own ‘tree decoration project’ as well as some preliminary cross-linguistic speech production data. These data include evidence for regularities in the realization of pitch accent placement and type in English discourse structure, and phonetic evidence for the intonational marking of discourse information status for accented and unaccented lexical items in Japanese.

1.1 Methodological issues

Typically, research on the production of prosody has relied on the comparison of carefully selected utterances that involve some intonational feature contrasting minimally in different conditions. Speakers are required to pronounce an utterance in some manner that will reveal their use of the relevant prosodic information to indicate the intended interpretation. Often, speakers are trained in a practice session with reference to examples from expert speakers (often the experimenters or their colleagues) who carefully produce the appropriate contrasts. While the results of these experiments generally have shown that speakers do use available prosodic contrasts in their production, the wider relevance of such research has been challenged by claims that the investigated contrasts are a product of laboratory conditions and
are not readily found in everyday situations. Thus, it is argued that speakers’ production of salient prosodic contrasts is dependent upon the extent to which their attention is drawn to the different possible interpretations of an utterance that can be indicated by prosodic structure.

A recent claim from studies of the prosodic resolution of syntactic ambiguity is that speakers fail to disambiguate prosodically when other linguistic information is available to express linguistic meaning. Typically, phonetic indicators of prosodic phrasing (including phrase-final segmental lengthening, tonal movement, or/and pausing) are reduced when syntactically ambiguous sentences are presented after short disambiguating paragraph contexts. In the presence of such contexts, speakers’ attention is not drawn to the ambiguity. In contrast, when speakers read decontextualised ambiguous sentences, and if they are made aware of the meanings involved and instructed to indicate meaning differences through differences in pronunciation, clear prosodic disambiguation is produced (c.f. Lehiste, 1973; Albrinott, McKoon & Ratcliff, 1996). Such evidence suggests that the use of expert and aware readers in production experiments will produce results that may be quite different from normal conversational speech, thus reducing the generalizability of the results.

Strategies to address the problem of speaker awareness have included recording speakers who have not been instructed to disambiguate and using sentences that contain temporary syntactic ambiguities (resolved within the utterance) rather than the global ambiguities used in the studies mentioned above (e.g., Warren, Grabe, & Nolan, 1995). However, the utterances obtained in this manner are still not representative of the spontaneous speech found in everyday dialogue, because it involves reading rather than talking. Readers and talkers have different pragmatic goals and different processing demands – unsurprisingly a number of researchers have documented the resulting differences between the prosodies of read and spontaneous speech (Ayers, 1994; Blaauw, 1994; Howell & Kadi-Hanifi, 1991).

1.2 Methodological strategies

While the study of spontaneous speech is desirable for its accurate representation of the productions that are most natural and typical of spoken communication, it presents a host of methodological difficulties. First, unscripted speech is largely unpredictable, making it difficult for experimenters to elicit pairs of utterances reliably that can be compared for a specific syntactic or pragmatic contrast. Second, spontaneous speech, even in a particular context, varies widely in length and content within and across speakers, frustrating attempts to collect multiple renditions of the same sequence from each of a set of comparable speakers. Such data are necessary
for reliable phonetic analysis of intonational features as well as for evaluating
the reliability with which speakers use prosodic structure to mark other types
of linguistic meaning.

Some control over lexical items and discourse content in spontaneous
speech has been achieved by creating conversational tasks that limit the topic
of conversation and provide a restricted number of objects to discuss. For
example, ‘map tasks’ (Anderson et al., 1991), route descriptions through
networks of nodes (Levitt & Cutler, 1983), or ‘tangram tasks’, where speakers
identify figures composed of geometric shapes (Clark & Shober; 1992, Clark
& Wilkes-Gibbs, 1986) have been used to examine such factors as the prosody
of utterance planning and repair, prosodic indicators of repeated reference, and
successfully controlled the set of words used in natural utterances in English
and Dutch by using a simple ‘time-telling task.’ The task was developed to
study the relation between the type of time expression (relative ‘ten past two’
vs. absolute ‘two ten’) and the preceding eye-movement on two types of clock
display (analogue vs. digital). Time-telling is a speech act ubiquitous across
cultures, and by manipulating the clock display, experimenters can easily
specify target words without explicitly instructing subjects. Therefore,
although the authors did not investigate prosodic patterns, this task holds great
potential applicability for the cross-linguistic investigation of prosodic
phenomena.

1.3 Eliciting specific contrasts in spontaneous speech

Researchers interested in specific prosodic contrasts have also used object-
manipulation tasks. To test whether probability of mention can induce
accentuation or de-accentuation, Nootenboom and Terken (1982) asked Dutch
speakers to describe changes in the configuration of movable and immovable
letters on a display. Displays had two to four letter symbols (P, T, K, B).
Immovable letters were marked with an underline (e.g., P). On each display,
there were either one or two movable letters, with a probability of mention of .1
for the former and .5 for the latter. Each display also had either one or two
underlined immovable letters, with the same probabilities of mention. The
speaker described the change in the location of a letter to a listener, who was
given a separate monitor to verify the speaker’s description. Most produced
utterances were of the form “De k komt links van de p (the k comes left of the
p),” and utterances that deviated from this pattern were excluded from
analysis. While the majority of utterances referred to both moving and fixed
letters, some ellipsis of the moving letter occurred when the probability of
mention for the letter was .1. As predicted, accentuation was more prominent
when the probability was .5 than when it was .1, for both the moving letter in
the subject position and the fixed letter in the predicate (in general, however, predicates were more likely to be accented than subjects regardless of the probability). Accents tended to be omitted or reduced when the letter was successively mentioned in the same grammatical position across trials. Nooteboom and Terken argue that their results reflect the speaker’s judgment about how accessible the referent is for the listener.

Terken (1984) claimed that the ‘moving letter task’ led to rather flat discourse structures with short sentences, limiting generalization to natural discourse. To elicit a more hierarchical structure in which the utterances are grouped into intermediate units with a stronger internal coherence, he developed another task in which a speaker gave instructions for constructing the front view of a house by assembling cardboard pieces. Since speakers did not get feedback from listeners, the data took the form of monologues. The speaker planned the order of construction and gave instructions using labels on the back of the pieces (e.g., roof, window, front door). Results showed that monologue utterances were formed units grouped around successive new topics (i.e. new part of the house). The probability of accentuation for the entities repeated within these units of instruction was low, which contrasted with a high probability of accentuation for the entities introduced right after the unit boundaries. Accentuation was also more likely for the sentence non-initial (as opposed to initial) positions, and for non-subject (as opposed to subject) position. Again, these results were interpreted to reflect the availability of the listener-oriented information that the speaker needed to assess in order to give instructions.¹

This ‘house-construction task’ was also used by Venditti and Swerts (1996) to investigate the relation between discourse structure and pitch range in Japanese. Their analysis is limited to data from a single speaker, but the study reports a tendency for a higher pitch range at the beginning of a discourse segment, suggesting that pitch range is a reliable indicator of discourse structure in pitch accent languages such as Japanese.

The challenge of the house-construction task stems from the freedom it gives to the speaker. Because the order of the instruction is determined freely by the speaker and because there is no intervening feedback from the listener, the length of instruction may vary to a large degree across speakers.² In addition, it is hard to control the information status of particular words used in

¹Terken and Nooteboom (1987) tested how listeners use accentuation and deaccentuation in a verification task using a display setting similar to Nooteboom and Terken (1982). The results showed that new information was verified faster when accented, whereas given information was verified faster when de-accented. While their findings confirm the function of accentuation/deaccentuation in utterance comprehension, whether speakers accent/deaccent with listeners in mind remains an empirical question.

²Terken (1984) reports that the total number of words per monologue varied from 219 to 799.
the monologue, which makes it difficult to obtain multiple renditions of target words under certain information structures across speakers.

The primary goals of a methodology for examining the relationship between produced intonation and the linguistic information conveyed in a discourse, then, are to elicit a sufficient amount of speech of a specific type of interest and to do so without changing the nature of the speech people typically produce in conversation. Structuring the speaker’s task can reduce the unpredictability of spontaneous speech, and some task designs may increase the number of productions that include the linguistic alternations of interest. Recent research has moved away from reading tasks and has begun to examine both the impact of real world situations and objects on production, as well as the discourse effects that arise in conversation rather than in monologues. In the following section, we present a comparison of the differing production results obtained for the same syntactic alternation in three object-based tasks.

2. Elicited speech in object-based tasks

This section discusses three recent studies that used tasks involving the manipulation of real-world objects to elicit the prosodic disambiguation of syntactic structure by naïve speakers. All three studies examined high/low prepositional phrase (PP) attachment in English, a syntactic ambiguity much-studied in the psycholinguistic literature. Phonetic analyses of sentences pronounced by trained speakers have shown that high attachment ([V NP PP]_{vp}) can be indicated to the listener by placing a prosodic boundary (segmental lengthening on the final word plus a following pause) after the NP. Low attachment ([V [NP PP]_{vp}]_{vp}) pronunciations tend to group the PP-modified NP as a single prosodic phrasal unit (Cooper & Paccia-Cooper, 1980; Price, Ostendorf, Shattuck-Hufnagel, & Fong, 1991). Example target sentences from each study are included below. All three studies also addressed the effect of situational context that disambiguates the syntactic ambiguity – does the presence of such context change speakers’ intonation such that prosody no longer resolves syntactic ambiguity? The studies show conflicting results. Although all three studies looked at both comprehension and production, we focus on the production data here, with an interest in understanding the impact of methodological decisions on the prosody produced in the experimental setting.

2.1. A partially scripted game-based task

Schafer, Speer & Warren (2000, 2001, 2003, in progress) constructed a game-based task to elicit utterances with specific syntactic contrasts. They compared
productions from a group of Midwestern American English (MAE) and New Zealand English (NZE), and collected multiple renditions from the same speaker, as well as across speakers. In their task, participant pairs worked together to negotiate the movement of objects around a game board. Speakers were required to use a fixed set of sentence frames and game piece names to construct instructions, requests and acknowledgements. Sentence frames were constructed to contain syntactically ambiguous and unambiguous sentences. The target PP attachment ambiguity sentence was ‘I want to change the position of the square with the triangle;’ which referred to two possible game moves: the triangle could be an instrument used to move the square or it could be part of a description of a house-like piece, the ‘square with the triangle.’ Each participant pair, one ‘Driver’ and one ‘Slider,’ had slightly different boards, and assumed different roles in the game. The “Driver” knew the ultimate goals for the pieces and issued instructions to the “Slider,” who followed the instructions and knew the locations of bonuses and dangers, but not the goals. The Slider asked questions about which piece to move and provided confirmation that game moves were complete. Both participants were aware of basic game rules governing the types of movement for game board objects. Game play lasted approximately two hours. Participants quickly became familiar with the set of utterances available for use under the rules of the game and learned to produce them fluently and without recourse to printed sentence lists. The researchers were careful not to pronounce game piece names to participants, so as not to model pronunciation of the complex NP ‘the square with the triangle.’ At the beginning of each game, participants were shown a board with labeled game pieces and asked to name each one (e.g., ‘triangle,’ ‘square,’ and ‘square with the triangle.’) The researchers termed the elicited data ‘quasi-spontaneous speech.’ By careful planning of gameboard layouts, game situations were constructed in which potentially ambiguous utterances remained ambiguous or were contextually disambiguated or were pronounced in situations with a contextual bias towards one of the meanings.

Fluent productions were analyzed both acoustically (duration, pausing, F0 patterns) and auditorily to locate the position of the strongest prosodic break in the utterances (using Pierrehumbert's Autosegmental Metrical framework encapsulated in the ToBI annotation system (Beckman & Ayers, 1997)). In accordance with the categories of the ToBI transcription, intonational phrase breaks are stronger than intermediate phrase breaks, which are in turn stronger than word-level breaks. The analysis produced three main groups of utterances, those with (1) the strongest break just before the preposition, (2) the strongest break at another location in the utterance, and (3) those with breaks at both of these locations. For both MAE and NZE speakers, phonetic measures and ToBI annotations showed that prosodic phrasing was used to indicate syntactic constituency. When speakers used ambiguous PP attachment sentences to indicate high attachment meanings, the strongest break followed
the word *square*. Both NZE and MAE speakers in both Driver and Slider roles produced a greater combined duration of *square* and following silence for the high-attached form, indicating instrumental use of the triangle. Sentences were annotated by a team of trained ToBI transcribers who had no information about the speaker's intended syntax. Annotation results revealed a strong influence of intended syntactic structure on relative boundary strength, so that the strongest boundary followed *square* in the majority of high attached tokens, but only 4% of low attached tokens. The strongest boundary followed a word other than *square* more often for low than high attachment sentences. Sentences that were produced with more than one strongest boundary, with one of these following *square*, (e.g. *...change the position*_[VP] *of the square*_[VP] *[with...]*) were more likely for high than for low attachment sentences.

Although speakers in the game task consistently used prosody to disambiguate these utterances, they varied the contours that they used to supply the disambiguation. Within and across speakers, many different pitch accent, phrase accent, and boundary tone combinations were used for the same morphosyntactic structure. For example, for the MAE speakers' sequence *change the position of the square with the triangle*, the authors found 63 distinct patterns on 79 high-attached utterances, and 87 patterns on 101 low-attached utterances.

To address the questions of whether context can reduce the need for prosodic disambiguation, and whether speakers must be instructed to disambiguate in order to do so, the researchers varied the discourse situations created by the game. Interestingly, they found that the degree of prosodic disambiguation of PP ambiguities did *not* vary according to the level of ambiguity in the situation. PP sentences were uttered in three types of game context. In ambiguous contexts, it was possible either to move a combined square-and-triangle piece or to use a triangle to push a square (the game rules did not allow the square to move on its own). In unambiguous contexts, one of the relevant pieces could not be moved, or it made no sense to move it since it was in its final goal location. In biased contexts, one interpretation was more likely because of the recent history of the game – for instance, a triangle had just been shifted to a space next to a square, setting up a subsequent move in which the triangle would be used to move the square. Although the three types of context varied the likelihood of the two syntactic interpretations, annotation and durational analyses from matched subjects showed no difference in the size of the contrast between high and low attached PPs. The authors interpreted the data to indicate that speakers produce largely the same prosodic contrasts regardless of the level of situational ambiguity.
2.2. A scripted toy-moving task

Contrasting results were found in a subsequent study of the same syntactic attachment ambiguity type, conducted by Snedeker and Trueswell (2003). They created a task in which the speaker uttered a series of commands involving the manipulation of a set of toys to a listener separated by a screen. Participants did not interact in this task, apart from the speaker asking if the listener was ready. Speakers received novel text with ambiguous PP attachment sentences such as ‘Tap the frog with the flower’ before each utterance and performed minimal manipulations of the toys. In the first experiment, speakers were asked to produce both high and low PP attachments and were shown toy displays consistent with either interpretation. In an example display, speakers saw a flower instrument, a plain frog, a frog holding a little flower, a distracter animal, and a distracter instrument. On each trial, the experimenter laid out a new set of stimuli, pronouncing the name of each toy or toy part, but carefully avoiding pronunciation of the complex NP (e.g., ‘the frog,’ ‘the flower,’ but not ‘the frog with the flower’). Next, the experimenter acted out the desired toy manipulation, either using the flower instrument to touch the plain frog, or using her hand to touch the frog with the small flower. The speaker then received the sentence as printed text on a card, and the experimenter acted out the desired toy manipulation again. Finally, the textual stimulus was removed and the command produced by the speaker from memory. Speakers were told “to say each sentence in such a way as to get the Listener to perform the same action on the other side of the screen” (p. 106). The structure of this task drew the speakers’ attention to the syntactic alternation between the instrumental and modifier readings of the sentence and instructed speakers to disambiguate prosodically. A post-experiment questionnaire tested for ambiguity awareness and showed that 97% of speakers reported having been aware of the possible syntactic alternation. Results from analyses of duration and ToBI annotation showed clear use of prosodic phrasing to indicate syntactic attachment with boundary patterns substantially like those described above for the Schafer et al. data. That is, speakers’ intended high-attachment utterances were more likely to be pronounced with a boundary after the direct object (e.g., ‘frog’) than those intended as low-attachment.

In a second experiment, speakers saw the same PP attachment ambiguity texts, but toy scenarios and acted-out instructions were consistent with just one of the possible syntactic interpretations (that is, arrays contained either a flower and a plain frog, or a frog holding a little flower, but not both). These collections of objects could not draw speakers’ attention to the ambiguity of the text sentence, and the post-experiment questionnaire showed only 31% of subjects reported an awareness of ambiguity. Results showed that although speakers who received the high attachment visual material still produced more
boundaries after ‘frog’ than those who received the low attachment materials, phonetic and phonological marking of the PP attachment difference was less consistent than in the first experiment. Crucial differences were significant in items analyses but not subject analyses, and significant differences did not appear in as many of the phonetic category comparisons. The authors concluded that situational context can influence speakers’ production of prosodic resolution of syntactic ambiguity.

There are a number of possible explanations for the difference in the results found by Schafer et al. (2000, 2001, 2003) vs. Snedeker and Trueswell (2003) (detailed discussion may be found in the original articles and in Schafer et al., in progress). In section 2.4, we discuss those that stem from the methodology used, particularly those due to the use of text, the construction of their tasks and materials, and the instructions to speakers.

2.3 A toy-moving task with pictorial instructions

The third study we present used schematic line drawings to elicit the production of ambiguous PP attachment sentences in ambiguous and unambiguous situations (Kraljic & Brennan, 2005). Participant pairs, a ‘director’ and a ‘matcher,’ were seated together before an array of toys, with the director behind and to the side of the matcher to prevent eye-contact. Directors instructed matchers to move toys, and the members of each pair were allowed to converse with one another freely during the task. Critical PP attachment ambiguities took the form of the sentence, “Put the dog in the basket on the star,” where the referent action could be either moving a dog in a basker onto a star location or moving a dog into a basket already on a star. Speakers pronounced the target utterances in situations that were either ambiguous (e.g., objects included a dog in a basket, a star, another basket on a another star, and a distracter animal) or unambiguous (e.g., objects included a dog in a basket, a dog in a box, a star, and another basket in the object modifier condition; or a dog, a basket on a star, a basket, and another animal in the goal modifier condition). Speakers were given example sentence frames during a practice session, one frame with the target syntactic form and another with an alternative syntax containing function words. On each trial, the experimenter first pronounced the names of the toys in the array, including the complex NPs (e.g., ‘the dog,’ ‘the star,’ ‘the dog in the basket’). After this, the speaker received for private viewing a card with line drawings of the objects in the array, an arrow indicating the instruction to give (e.g., from the dog in the basket to the star), and a dot next to objects to mention in the instruction (e.g., the basket). Speakers were told to use the verbs ‘put’ and ‘place,’ and to give the instructions in a brief manner, but no further instructions were given about what to say. Kraljic and Brennan noted that they needed to add additional card
instructions to “prevent speakers from falling into a repetitive pattern and producing all utterances with the same mechanical prosody.”

The open-ended nature of the instructions produced various sentence forms, but roughly half the utterances produced by the participants had the targeted syntactic structure. These were submitted to durational analyses of the object and object of the preposition nouns, which showed that speakers reliably produced a prosodic distinction between modifier and goal interpretations of *Put the dog in the basket on the star*. These results held regardless of the level of situational ambiguity, that is, prosodic boundaries were produced to coincide with the major syntactic boundary intended by the speaker whether the situation was ambiguous or unambiguous.

Kraljic and Brennan conducted a follow-up experiment to determine speakers’ awareness of ambiguity. Anecdotal reports from the first experiment indicated that speakers perceived ambiguity in some of the unambiguous trials. Speakers in a separate timed ambiguity detection task showed a range of individual differences in performance, with 90% accuracy for some, and below chance performance for others. Mean time to judge ambiguity was over four seconds, with times ranging from 2.528 to 7.864 seconds. They interpreted these results to indicate that speakers were often unaware of ambiguity at the time they planned and produced their instructions. An additional follow-up experiment compared prosodic disambiguation performance with and without a matcher present. The method was similar to that in the first experiment with the exception that half of the speakers were told to speak into a microphone in order to create instructions that would be played for a matcher at a later time. Results from duration analyses showed prosodic disambiguation in both conditions, with no differences in the level of prosodic marking between these two conditions.

### 2.4 Methodological differences and differences in results

Differences between the tasks used in these three experimental settings may shed light on reasons for the differences that appear among their results. Here, we focus on differences in instructions, differences in the use of text materials, and differences in task and materials construction that might lead to differences in the process by which speakers plan and execute their utterances.

*Instructions to subjects.* Looking across these three studies, it appears that speakers whose attention is drawn to a prosodically resolvable syntactic alternation, and who are instructed to use prosody to disambiguate, produce stronger prosodic cues to the resolution of syntactic ambiguity. In Snedeker and Trueswell’s first experiment, speakers were aware of syntactic ambiguity and were implicitly instructed to use prosody to distinguish between two meanings of a syntactic ambiguity. Speakers saw two possible toy
combinations that were possible referents for the sentence to be pronounced, watched the experimenter’s demonstration, and then read the text they would memorize as an instruction from the card. In contrast, participants in the Schafer et al. study were instructed to use only the given sentence frames and game piece names, and those in the Kraljic and Brennan studies were instructed to use particular verbs and to be brief, but were not told to disambiguate. Speakers in Snedeker and Trueswell’s second experiment were much less aware of syntactic ambiguity than those in the first experiment. Although they produced some disambiguation, it was less reliable when compared to the strong phrasing cues produced with instruction and awareness of ambiguity. Schafer et al. examined ambiguity awareness by testing the strength of prosodic boundaries used across the two hours of productions in their experiment, and found no difference in disambiguation strength between initial and final PP productions. Kraljic and Brennan also tested ambiguity awareness, and concluded that it was unlikely that it had influenced productions in their experiments.

Use of text. Speakers’ production of prosodic phrasal cues to syntactic boundary locations seems reduced when text prompts are is given for each utterance, perhaps because pronunciation in this situation becomes more typical of read speech. Prosodic disambiguation was least evident in Snedeker and Trueswell’s second experiment, where novel text was provided before each trial. Schafer et al. and Kraljic and Brennan also used text, but in the form of frames at the beginning of their experiments which were then dispensed with during the rest of the tasks. Schafer et al. also used text labels for game pieces at the start of the experiment, contrasting with the other two studies which provided spoken versions of object names. Speakers did not seem to be systematically affected by the presence of possible models for the pronunciation of the complex NP – although both Schafer et al. and Snedeker and Trueswell were careful not to provide such cues, the pattern of results they found diverged. Kraljic and Brennan provided complex NP pronunciations, but a comparison of their results to those of Schafer et al. shows no evidence that speakers produced increased disambiguation in the presence of such models.

Task and materials construction. The three studies considered here differed greatly in the complexity of their manipulations of visual situational context, the variety of productions made by the speaker, and the use of a conversational partner. All of these factors may lead to differences in the process by which speakers plan and execute their utterances.

Snedeker and Trueswell’s experiment provided the simplest contexts and sentence structures with the least variety. In contrast, both the task and the stimuli in other studies were more complex visually as well as syntactically. In Snedeker and Trueswell’s task, the same very simple syntax was used repeatedly with successive arrays of objects. When arrays provided two referent actions, the syntactic simplicity may have made it easier for speakers
to recognize and remember the two alternate syntactic forms of the word strings. When arrays provided a single referent action, Kraljic and Brennan’s observation about ‘mechanical prosody’ may be relevant – speakers may have fallen into a single phrase pronunciation routine, producing uninformative prosody due to the repetition of homogeneous, short sentences.

Snedeker and Trueswell allowed minimal contact and no task-related feedback between speaker and listener. In their second experiment, the speakers needed to convey neither a syntactic distinction nor a sentence-level semantic contrast. In contrast, Kraljic and Brennan did not restrict interaction between director and matcher; they allowed discussion and feedback to correct misunderstanding of instructions, and they asked the speaker to observe the matcher’s behavior to see if the intended meaning was conveyed. Schafer et al. restricted the specific utterances available for interaction, but built mechanisms for feedback and correction into the scripted sentence frames they used. However, if the presence of a conversational partner who provides feedback is critical to the production of prosodic disambiguation, why did Kraljic and Brennan find comparable prosodic disambiguation in ambiguous and unambiguous situations not only when conversants sat side-by-side viewing a common display, but also when a lone speaker taped instructions for later use? These results suggest that the task of planning a message (at some level higher than mere lexical access and phonetic implementation) may be the key distinguishing factor between disambiguating productions and ambiguous ones. When speakers have a clear message in mind and must do some high-level planning to construct that message, as they do in spontaneous speech, they seem to disambiguate PP sentences.

3. Analyzing Scripted Speech

This section discusses two recent studies that used tightly scripted speech to elicit utterances for phonetic analysis of prosodic structure. Script reading with role-play, where speakers imagine a conversational setting and produce utterances accordingly, is the primary method used to elicit intonational contours that convey pragmatic contrasts. Here, we direct our attention to the pitfall of assuming a particular pragmatics for scripted dialogues and of training speakers to produce the pitch contours of interest, using as an example Arvaniti and Garding’s (to appear) study of English accent types used to convey emphasis. We then discuss the phenomenon of prosodic patterns that emerge in a script-reading task, but occur less often in spontaneous speech, using as an example Ito’s (2002) study of prosodic focus in Bermeo Basque and Tokyo Japanese. The tonal patterns that emerge more frequently in free speech than in read speech will be further discussed in the following section.
3.1 Pragmatics of scripted dialogues

Arvaniti and Garding (to appear) use scripted dialogues to examine traditionally-assumed categorical distinctions among H*, L+H*, and L*/H (since Pierrehumbert, 1980) in two dialects of American English. Participants played roles in two types of dialogues. The scripts were structured such that a target word (e.g., Raymona) would convey four different degrees of emphasis. In one type of dialogue, the target word was expected to be produced with H* at the lowest end of the emphasis continuum and with L+H* at the highest end. In the other type, the target word was expected to carry L*/H with all four degrees of emphasis (See Table 1, taken from Arvaniti & Garding, Table 1).

<table>
<thead>
<tr>
<th>Investigator:</th>
<th>H*/L+H* or phone dialogue</th>
<th>L*/H or date dialogue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaker:</td>
<td>Who was on the phone?</td>
<td>I need someone who</td>
</tr>
<tr>
<td></td>
<td>Raymona.</td>
<td>would hold her own.</td>
</tr>
<tr>
<td>Investigator:</td>
<td>Who?</td>
<td>But someone who would</td>
</tr>
<tr>
<td>Speaker:</td>
<td>Raymona.</td>
<td>understand politics.</td>
</tr>
<tr>
<td>Investigator:</td>
<td>It was who?</td>
<td>But someone who would</td>
</tr>
<tr>
<td>Speaker:</td>
<td>Raymona.</td>
<td>be really funny, too.</td>
</tr>
<tr>
<td>Investigator:</td>
<td>Raymond?</td>
<td>But someone who would</td>
</tr>
<tr>
<td>Speaker:</td>
<td>RAYMONA!</td>
<td>look totally hot!</td>
</tr>
</tbody>
</table>

Based on their measurements of local F0 minima and maxima and the alignments of those minima and maxima, Arvaniti and Garding argued for clear phonetic distinctions between L+H* and L*/+H. Their analysis further revealed a larger effect of emphasis on Hs than on Ls, and interesting tendencies in the alignment of non-starred tones (i.e., L in L+H* being mostly within the stressed syllable and H in L*/+H not being delayed under emphasis as much as a starred H). These detailed reports undoubtedly take us to a deeper level of understanding of the relation between the phonological status of tones and their phonetic realization. In addition, Arvaniti and Garding point out dialectal differences between Minnesotan (MN) and Southern Californian (SoCal) speakers; in general, SoCal showed later tonal alignment than MN, and whereas SoCal seemed to maintain the contrast between the flat and rise-fall tonal shapes, MN seemed to lack a flat H*. These findings clearly merit attention as such dialectal variations are expected to be ubiquitous and must
not be ignored during the investigation of natural discourse processing. However, we must not also overlook potential pitfalls of the above methodology. Although we are not arguing that the data are unrepresentative of the two dialects under analysis, we suggest that more convincing results might be obtained with a more natural, communicative task. There are two reasons behind this suggestion.

First, we share the intuition of an anonymous reviewer quoted in their paper, who points out that their speakers were “highly trained, practiced speakers, and they were coached until they managed to produce something close to the desired contours (in Notes 6)”. In their description of the procedure, Arvaniti and Garding state that the second author demonstrated to each speaker “how to consistently use the same contour within each dialogue while increasing the level of emphasis”. This is problematic especially in the L*+H-inducing dialogues, such as the above date dialogue, where the pragmatics of the target lines change as the dialogue proceeds. To test how likely it was for their script to induce the intended target tonal contour, we asked two native speakers of Midwestern American English to take the Speaker role in the date dialogue in Table 1 without training them with the target contour. Neither speaker produced L*+H on “Raymona”, and one produced a stronger accent on “would” as the dialogue proceeded. As we explained the intended tonal patterns, the other speaker felt that the L*+H accent would be appropriate for the Speaker’s first line, but would be infelicitous for the rest of the lines. This is because the Speaker’s intention should shift toward convincing the Investigator instead of staying with “the sly/cajoling attitude,” which was part of the context that was explained to the speakers before recording in Arvaniti & Garding’s experiment. Our first speaker agrees with our second speaker that “Raymona” must already be part of the common ground in the second line and thereafter. Our first speaker argues that in a natural discourse, “Raymona” would normally be replaced by the pronoun “she”. For the production of “She would….,” the speaker’s emphasis should be on the predication of the background/topical subject, and thus, having an accent on the auxiliary, i.e., “She WOULD…”, would be the most natural response in the Speaker’s role. In addition, Arvaniti and Garding’s participants listened to pre-recorded example sequences of their parts without the intervening Investigator’s lines before the recordings. This training makes it highly possible that the speakers who were analyzed were able to successfully calibrate themselves upon listening to the example recordings and then maintain the intended intonation patterns while paying less attention to the pragmatics of scripted dialogues. If this is the case, we cannot be certain whether the observed pitch accents would surface in the reported phonetic forms in the speech of the same speakers were they speaking freely outside of such supervision.
The second problem with this method concerns their unreported speakers. Arvaniti and Garding remark that speakers had more difficulties with controlling the degree of emphasis than with maintaining the target pitch contours. In the endnote, however, they state that there were eight SoCal speakers whose data were not reported (they report data from six speakers). Five were excluded due to improvisation, being bilingual, and history of speech therapy while the other three could not produce the intended contours. Thus, three out of nine SoCal monolinguals with typical speech development who initially provided data had difficulties producing the intended pitch contours for their scripted dialogues despite example recordings and explicit instructions to maintain the intonation patterns throughout each dialogue. While this does not invalidate the reported differences in tonal scale and alignment between L*+H and L+H*, their data may not necessarily demonstrate “distinct phonological categories with distinct pragmatics (p19 in ms)” that would emerge in natural discourse. One way to improve the use of scripts for prosodic analysis would be to pretest or norm the materials with multiple naive speakers to ensure that the target dialogues naturally induce the intended tonal patterns without training. Another way to achieve a better mapping between accent distribution and pragmatics is to separate the tonal annotation and the phonetic analysis from discourse analysis, and then to examine the correspondence between the prosodic structure and the discourse structure. We propose this approach with more naturalistic speech data later in this chapter.

3.2 Intonation in text-reading: Basque & Japanese

The above criticism of using highly controlled speech data should be directed to many other studies, including the first author’s past work, where the researcher had no other choice but to select for analysis those utterances that had been produced in an intended manner. In her dissertation, Ito (2002) conducted a comparative production study between Bermeo Basque and Tokyo Japanese, which happen to show crucial similarities in their prosodic systems as pitch accent languages. In order to investigate the effect of focus on the tonal scale and alignment of the two adjacent lexical pitch accents, Ito prepared a set of question-answer pairs to be read by participants for both languages (see (1) for an example dialogue set in Basque).

(1) Example dialogue set in three focus conditions: Bermeo Basque

a. Zer topa zu tallerran? “What did you find in the garage?”
   Ba. ÁNAN MÁLLUE topa dot. “Well, I found Anna’s hammer.”

b. Tallerran zure aman mállue topa zu? “Did you find your mom’s hammer in the garage?”
   Ez. ÁNAN mállue topa dot. “No, I found ANNA’S hammer.”
c. Tullerran Ànan zerrie topa zu? “Did you find Anna’s saw in the garage?”
Ez, Ànan MÁLLUE topa dot. “No, I found Anna’s HAMMER.”

This experiment was also designed to investigate the interaction between the focus effect on the pitch accents and the distance between the two adjacent accents, or the inter-accent-interval (IAI). The target words were chosen so that the number of morae/syllables varied from one to three in both languages: (1) Ànan málue, (2) Ànan janáirije, (3) Ànan arkomári (Anna’s hammer, Anna’s meal, Anna’s shirt, respectively in Basque); (1) hadé-na midori-no, (2) hadé-na kimidori-no, (3) hadé-na fukamidori-no (loud green, loud lime-green, loud dark-green, respectively in Japanese). In order to examine how the size of IAI affects the effect of focus using data that contained as little noise as possible from phonetic variation, Ito tried to minimize the phonotactic differences across conditions within each language. In addition, she recorded each dialogue 12 times with each participant across three to four sessions (trials were randomized within each session). Obtaining multiple productions of controlled linguistic targets was a crucial requirement for detailed phonetic analysis, where F0 height and peak alignment must be compared across three focus conditions, three IAI conditions, and four speakers in each language. Since the size and the timing of F0 changes are largely affected by the voicing quality and the inherent duration of segments, changing target phonetic sequences across experimental conditions would introduce noise that would make the evaluation of effects extremely difficult. In the field of laboratory phonology, controlling the linguistic materials for phonetic analysis has been justified mainly for this reason. There is no doubt that we gain indicative results from carefully controlled read/practiced speech, and we by no means intend to conclude that the results of phonetic analysis of scripted utterances cannot demonstrate the characteristics of the target prosodic systems. However, investigators must keep questioning to what extent findings with scripted laboratory speech account for phonetic phenomena during natural, unmonitored speech acts. We also need to remain cautious upon making claims about observed prosodic patterns and pragmatics, as the script-reading task may induce particular prosodic patterns that may not be observed frequently in natural conversations.

Ito (2002) observed an interesting pattern among some speakers of Bermeo Basque and Tokyo Japanese when they tried to perform the above script-reading task. In both languages, speakers were asked to read the answer line on each card (approximately 2.8” x 8” or 7 x 20.5 cm) where the target words under focus were printed in bold with underlines. During the Japanese experiment, Ito noticed that some speakers put an accent on –na in hadé-ná or –no in midori-nó, when the word was under narrow focus. Kori (1989) claims that Japanese speakers sometimes put an accent on the case-markers when they express narrow focus, and such accenting on the case-marker may place more
weight on the grammatical function than on the meaning of the word. Since the purpose of the recording was to analyze the tonal patterns of the lexical accent in the stem, Ito demonstrated her intended intonation pattern with a pitch range expansion over the accented mora in the stem word and asked the speakers not to put an accent on the case-marker. Most of her speakers dropped case-marker accenting after a couple of practice utterances and did not need a reminder in the successive recording sessions. Interestingly, Ito also noticed that some Basque speakers put an accent on the last syllable of the narrowly focused words (i.e., malluÉ), which sounded very similar to the Japanese case-marker accenting pattern. With these speakers, Ito attempted to make a similar correction through the interpreter who was taking the question part in the scripted dialogues. With one speaker, however, Ito had to give up collecting data because he could not switch from his final accenting pattern to a simple pitch range expansion over the lexically accented syllable.

In the following section, we introduce an interactive task from which we have collected multiple renditions of limited sets of target words in both English and Japanese. Interestingly, we have not observed the focus-related case-marker accenting in Japanese during this task as often as during Ito’s past experiment with scripted dialogues. On the contrary, a different type of final accenting, which never appeared during the reading task has been observed frequently during this unscripted speech.

4 Interactive speech task: Directing tree decoration

The task described here was developed to investigate the relation between tonal patterns and the information status of words in a natural discourse in Midwestern American English and Tokyo Japanese. By using a set of common color terms and object names for Christmas tree ornaments, we have successfully collected multiple spontaneous productions of those target words in a natural conversational setting in both languages. We controlled the sequence of decorations to collect multiple repetitions of target words. The same lexical items appeared in different conditions for the same speaker, and in addition, these items bore comparable information status across speakers. This enabled a reliable statistical analysis of the mapping between information structure and tonal patterns, and a reliable phonetic analysis of the intonation produced.

4.1 Task Description

Participants were told to work as a member of a pair to decorate trees. A confederate acted as the decorator, while a naive participant played the role of
the director. The director sat inside a soundproofed booth and gave instructions regarding which ornament to pick and where to place it on the tree according to a monitor display. The confederate/decorator followed these instructions, communicating with the director through sets of headphones and microphones. Through a window on the booth, the director could see the tree and the decorator’s hands, but not his/her face and the ornament tray. The ornament tray was hidden from the director so that the ornaments would not be included as discourse entities in the common ground between the director and the decorator before the director mentioned them. The experimenter sat beside the decorator to advance the monitor display. The relative locations of the director, the decorator, the experimenter and the equipment are depicted in Figure 1.

![Figure 1. Director, Decorator, and Experimenter during tree-decorating production experiment.](image1)

On each trial, a photo of a tree appeared first on the left side of the monitor display. After a moment, the target ornament appeared on the right side, and a text tag appeared simultaneously indicating the location at which it should hang on the tree. Each tag was a combination of a color adjective and an object noun (e.g., orange drum; see Figure 2). Directors were not explicitly instructed to use the tags to name ornaments, but most of them did. The decorator (a confederate) was instructed not to use the tagged ornament names, but to use anaphoric expressions such as ‘this/that (one)’ and ‘it’ to refer to the ornaments when confirming their locations on the tree (e.g., ‘Does it go right below this one?’). This was done to control the relative frequency and information status of target words during the experiment. In order to collect the recordings of the target words in comparable pragmatic contexts across

![Figure 2. Monitor display for the director in the tree-decorating task.](image2)
multiple participants, it was crucial to reduce the number of mentions outside the structured orders. For the target ornament tags, eight color adjectives and eight object nouns were used in both English and Japanese. These words were chosen to maximize the number of voiced segments for F0 tracking, and to balance accented words and unaccented words in Japanese (Table 2).

<table>
<thead>
<tr>
<th>targets</th>
<th>Japanese</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accented</td>
<td>buru’u-no ‘blue’</td>
<td>buru’u-no ‘blue’</td>
</tr>
<tr>
<td></td>
<td>guri’in-no ‘green’</td>
<td>guri’in-no ‘green’</td>
</tr>
<tr>
<td></td>
<td>ore’Nizi-no ‘orange’</td>
<td>ore’Nizi-no ‘orange’</td>
</tr>
<tr>
<td></td>
<td>guruma-no ‘gray’</td>
<td>guruma-no ‘gray’</td>
</tr>
<tr>
<td>Unaccented</td>
<td>beeju-no ‘beige’</td>
<td>beeju-no ‘beige’</td>
</tr>
<tr>
<td></td>
<td>miznari-no ‘sky color’</td>
<td>miznari-no ‘sky color’</td>
</tr>
<tr>
<td></td>
<td>chairo-no ‘brown color’</td>
<td>chairo-no ‘brown color’</td>
</tr>
<tr>
<td></td>
<td>toomei-no ‘clear’</td>
<td>toomei-no ‘clear’</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>noun</th>
<th></th>
<th></th>
<th>noun</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>‘ball’</td>
<td></td>
<td></td>
<td>blue</td>
</tr>
<tr>
<td>guri’in-no</td>
<td></td>
<td></td>
<td></td>
<td>egg</td>
</tr>
<tr>
<td>ore’Nizi-no</td>
<td></td>
<td></td>
<td></td>
<td>onion</td>
</tr>
<tr>
<td>guruma-no</td>
<td>ie’</td>
<td></td>
<td></td>
<td>house</td>
</tr>
<tr>
<td>beeju-no</td>
<td>kane</td>
<td></td>
<td></td>
<td>bell</td>
</tr>
<tr>
<td>miznari-no</td>
<td>doramu</td>
<td></td>
<td></td>
<td>drum</td>
</tr>
<tr>
<td>chairo-no</td>
<td>ningyoo</td>
<td></td>
<td></td>
<td>doll</td>
</tr>
<tr>
<td>toomei-no</td>
<td>ame</td>
<td></td>
<td></td>
<td>candy</td>
</tr>
</tbody>
</table>

Table 2. English and Japanese color adjectives and object nouns for the tree-decorating experiment.

4.2 Discourse Analysis

As previously mentioned, our participants were neither trained nor instructed to use particular carrier phrases or sentences to perform the task. The only direct linguistic input that the director received was the tag for each ornament on the screen, and thus directors were engaged in spontaneous conversations with the confederate decorators to accomplish the tree decoration task. Although we presented photo-slides in particular orders, we had no other control over the details of the discourse structure. Therefore, it was necessary to analyze the structure of each recorded discourse to evaluate the information status of target words. Our approach was fundamentally comparable to Terken’s (1984, see Introduction). However, while he used a hierarchical utterance-grouping strategy for his analysis of accent distribution, we adapted the framework of discourse analysis proposed by Grosz and Sidner (1986). We chose this method of identifying the discourse structure of the data because it provides well-defined criteria that are crucial for the description of accent choice in spontaneous dialogues. According to this model, utterances are grouped into subunits of conversation or discourse segments (DS) around local discourse purposes that encompass the meaning of successive utterances. This is an intention-based approach, where the coherence among utterances is assumed to be driven by the speaker’s intentions. The role of intonation in marking the discourse structure is widely discussed in Grosz and Hirschberg.
(1992), Hirschberg and Grosz (1994), Hirschberg and Nakatani (1996), and Nakatani, Hirschberg and Grosz (1995), etc.

We assumed that all the utterances exchanged during the decoration of one tree belong to one level of discourse, which in turn is composed of a set of smaller DSs. Table 3 shows an example piece of conversation in English, which is grouped into several DSs.

<table>
<thead>
<tr>
<th>Time Spkr Utterance</th>
<th>Time Spkr Utterance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>054.2</strong> Dir: At the very top there’s a white hat. Dec: [shows] That one? Dir: Yeah. Dec: All right. [places] Okay. Next?</td>
<td>bad, I forgot to tell you, this is ... this gonna be on the next down; there’ll be three in this row.</td>
</tr>
<tr>
<td><strong>069.5</strong> Dir: Uh, a blue(H*) house(H*). [Dec: shows through booth window] Dir: Yeah. Dec: And which way did you say ... right or left?</td>
<td>Dec: And this is on the right, yes? Dir: Yeah. Dec: [places on tree] Okay? Next?</td>
</tr>
<tr>
<td><strong>108.9</strong> Dir: Uh, this is an orange(L+H*) bell(H*). [Dec: shows through window] Dir: Okay. Next. Dec: [places on tree] Okay.</td>
<td><strong>188.3</strong> It’s a green(H*) onion(H*). doesn’t really look like an onion.</td>
</tr>
<tr>
<td><strong>108.9</strong> Dir: Uh, this is an orange(L+H*) bell(H*). Dec: [places on tree] Okay.</td>
<td><strong>188.3</strong> It’s a green(H*) onion(H*). doesn’t really look like an onion.</td>
</tr>
<tr>
<td><strong>108.9</strong> Dir: Uh, this is an orange(L+H*) bell(H*). Dec: [places on tree] Okay.</td>
<td><strong>188.3</strong> It’s a green(H*) onion(H*). doesn’t really look like an onion.</td>
</tr>
</tbody>
</table>

The orders of ornaments were arranged so that each color adjective and object noun appeared three times within each tree (combining with a different word to form a complex NP across those three tokens). The information status of each word was tagged as ‘first mention’ or ‘consecutive mention’. Note that the specification of ‘first’ vs. ‘consecutive’ mention can apply at both the higher discourse level (D-level) and the local level (DS-level). For example, a drum may appear for the second time in a given tree, but it can be the first mention within a local segment (e.g., within a row of ornaments on the tree). This dual tagging of first vs. consecutive mention allowed us to investigate the effect of discourse structure on the distribution of particular pitch accents, such

---

1 Readers may find the details of the project on the intonation and discourse led by B. J. Grosz at http://cslu.cse.ogi.edu/nsu/issg897/reports/grosz1.html
as L+H*, which is often used to express the contrastive status of a word (Pierrehumbert & Hirschberg, 1990).

4.3 Tone Annotation

Although speakers of the same dialects are generally expected to produce similar tonal patterns for words conveying certain pragmatics, we cannot assume a strict one-to-one mapping between choice of pitch accent type and pragmatics either within or across speakers. The degree of consistency in the use of pitch accents remains an open research question. In order to investigate the range of pitch accent choices during natural conversation, it is important to annotate the tonal patterns without knowledge of surrounding discourse context. For this reason, we ToBI-annotated the recorded decorator utterances individually and in random order.

Two transcribers were instructed to annotate the target words (i.e., color adjectives and object nouns), the pitch accents for the target words, and the boundary tones around the target NP (if any) (See Figure 3 for an example annotation). Each transcriber annotated a subset of utterances for the purpose of testing the inter-transcriber reliability (for a description of the calculation of reliability scores, see procedures in Pitrelli, Beckman, & Hirschberg, 1994.)

![Figure 3: Example ToBI annotation](image)

<table>
<thead>
<tr>
<th>WORD POSITION</th>
<th>FIRST/CONSECUTIVE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>D-level</td>
</tr>
<tr>
<td>Adj/Noun</td>
<td>D-f</td>
</tr>
<tr>
<td>Adj/Noun</td>
<td>D-c</td>
</tr>
</tbody>
</table>

Table 4. D-level to DS-level tagging correspondence

4.4 Preliminary Observations

The preliminary analyses presented here represent a subset of the data collected. For English data, analyses so far have focused on phonological form and information structure. We present ToBI annotation and discourse transcription data from two English speakers who produced similar discourse information structures. Results demonstrate that naive speakers of English predictably generate pitch accent patterns reflecting information structure. Those pitch accent patterns also serve as an indicator of the discourse
situations that give rise to intonational marking of information status and the frequency of this marking. For Japanese data, analyses so far have focused on the phonetic form of words in contrastive contexts, specifying focus effects for accented vs. unaccented words. We present analyses of fundamental frequency from twelve Japanese speakers and a brief analysis of the productions by three speakers of a boundary-related pitch movement characteristic of spontaneous (but not read) speech in Japanese.

As mentioned above, we did not instruct our speakers (directors) about what to say when they gave directions to the decorators. Thus, it was necessary for us to transcribe the discourse structures they produced in order to determine the pragmatic status of the target words from each speaker’s productions. The two English speakers discussed here consistently started a new subunit of conversation or discourse segment (DS), each time they directed the decorator to move on to the next row on the tree (see Table 3 for an example). Thus each DS was formed around the purpose of completing a row. To examine accent presence/absence and type by discourse status, target words were tagged as first mention or consecutive mention at both the discourse (D) and discourse segment (DS) levels (Tagging correspondence is shown in Table 4).

A word was tagged as first mention at the D-level when it appeared for the first time in a tree, and first mention at the DS-level when it appeared for the first time in a row. Second and the third mentions were considered consecutive at both D and DS-levels. Note that a word could be globally consecutive (D-c) but locally first (DS-f). In addition to consecutive/first mention, a word was tagged as contrastive, when the adjacent adjective or noun was repeated from the immediately preceding trial (e.g., ball in green candy → green ball; or blue in orange bell → blue bell).

<table>
<thead>
<tr>
<th>Domain</th>
<th>First/Consecutive</th>
<th># of trials</th>
<th>ADJ</th>
<th>NOUN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discourse (D)</td>
<td>FF 31</td>
<td>.94</td>
<td>.84</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FC 32</td>
<td>.84</td>
<td>.56</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CF 32</td>
<td>.91</td>
<td>.81</td>
<td></td>
</tr>
<tr>
<td>Discourse Segment (DS)</td>
<td>FF 67</td>
<td>.91</td>
<td>.88</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FC 62</td>
<td>.89</td>
<td>.39</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CF 56</td>
<td>.88</td>
<td>.79</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CU 6</td>
<td>.83</td>
<td>.83</td>
<td></td>
</tr>
</tbody>
</table>

Table 5. Accent placement by word position in discourse structure.

The data were examined to investigate the frequency of accentuation in DS-level first mention vs. D-level first mention. Table 5 summarizes the number of observations and the proportion of accentuation for target adjectives and nouns in each first/consecutive condition. The current data indicate a strong relation between word position and the presence of accent, with adjectives more likely to be accented than nouns. Target adjectives were produced with an accent more than 80% of the time regardless of their first/consecutive status at both D- and DS- levels. Even when a word was repeated within a row (i.e.,
DS-CF/CC conditions), an adjective bore an accent about 85% of the time. We also examined the relation between pragmatic contexts and the choice of pitch accent. Table 6 summarizes the distribution of L+H* accents in Contrastive/Non-contrastive contexts defined at the D- and DS-levels. At both levels, the L+H* accents appeared more frequently in contrastive contexts than in non-contrastive contexts. On the other hand, nouns tended to be produced without an accent more frequently when they were in contrastive contexts. Word position interacted with contrastiveness, such that contrastive adjectives were more likely to bear a L+H* than contrastive nouns, while given nouns were more likely than adjectives to be deaccented in the presence of adjacent contrastive accent (Ito, Speer, & Beckman, 2003).

Table 6. L+H* accents in contrastive and non-contrastive contexts in discourse

<table>
<thead>
<tr>
<th>Domain</th>
<th>F/C status</th>
<th># of trials</th>
<th>L+H* # of trials</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discourse (D)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contrastive</td>
<td>F</td>
<td>26</td>
<td>.46</td>
</tr>
<tr>
<td>Non-Contrastive</td>
<td></td>
<td>37</td>
<td>.03</td>
</tr>
<tr>
<td>Contrastive</td>
<td>C</td>
<td>37</td>
<td>.51</td>
</tr>
<tr>
<td>Non-Contrastive</td>
<td></td>
<td>91</td>
<td>.00</td>
</tr>
<tr>
<td>Discourse</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Segment (DS)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contrastive</td>
<td>F</td>
<td>60</td>
<td>.47</td>
</tr>
<tr>
<td>Non-Contrastive</td>
<td></td>
<td>10</td>
<td>.04</td>
</tr>
<tr>
<td>Contrastive</td>
<td>C</td>
<td>3</td>
<td>.67</td>
</tr>
<tr>
<td>Non-Contrastive</td>
<td></td>
<td>59</td>
<td>.03</td>
</tr>
</tbody>
</table>

To examine how first/consecutive mention affects prosodic phrasing and interacts with the lexically determined tonal patterns of Japanese, target phrases collected from 12 speakers were analyzed. Consistent with Ito (2002), results showed that intonational marking of first/consecutive mention was constrained by lexical accentual properties. When the target phrase (color + object) was initiated by a lexically accented adjective (e.g., buru’u-no ‘loan word blue’ or ore’Nzi-no ‘loan word orange’: Fig 4 A&B), the pitch range for the following noun was remarkably compressed, and the noun tended to be set off in a separate AP with a distinctive rise (shown with in Fig 4, A&B).

Figure 4A-D. F0 contours for First First (ff) vs. First Consecutive (fc) conditions for 4 accent pairs.
When the phrase was initiated by an unaccented adjective (e.g., tyairo-no ‘brown’ or toomei-no ‘clear’: Fig 4, C&D), the following nouns tended to merge with the adjective, forming a flat tonal unit as a whole phrase. F0 values were extracted from the midpoint of the first and the second mora of the target noun. If the noun is phrased into a separate AP from the preceding adjective, it should be shown as a phrase-demarcating %LH-pitch rise from the first to the second mora. In fact, this rise tended to be larger when the noun was new. In a regression model, the F0 value at the midpoint of the first mora and the information status of the noun accounted for 80% of the variability in the second mora value, with new nouns having on average an 8Hz larger rise size than given nouns ($t = 2.292, p < .05$) (Ito, Speer, & Beckman, 2003).

In sum, the data from these speakers indicated that while the overall intonation contour of a phrase was predictable from the lexical accentual specification of the words forming the phrase, the information status of words did affect the size of the AP-initial rise.

In addition to the above patterns, the Japanese spontaneous speech data exhibited a very interesting characteristic. Many female speakers produced a salient boundary pitch movement (BPM) with a clear rise and fall, typically over the AP-final case-markers. An example pitch trace is shown in Figure 5. The tone annotation shown for the BPMs in Figure 5 is given by Igarashi, an expert X-JToBI transcriber at Riken Brain Science Institute. According to Venditti, Maeda & van Santen (to appear), the first type of BPM (L%HL%) is categorized as a “rise-fall” boundary movement, whereas the second type (L%H%) may fall into either “prom” (for prominence-lending) or “insist” type.4

![Figure 5. Annotated pitch trace for boundary pitch movement (BPM) in spontaneous Japanese.](image)

---

4 Venditti judges this particular example BPM for *tamane’gi-NO* as a prominence-lending (prom) type. Igarashi does not agree with the categorical distinction between ‘prom’ and ‘insist’ proposed by Venditti et al. Igarashi considers the possibility of the two BPMs (H% and HL%) being at the opposing endpoints of a continuum of HL BPM. According to Igarashi’s view, H% is a HalfRise, of which the falling contour is truncated due to lack of segmental duration [-longsyllable], whereas HL% is fully realized RiseFall with [+longsyllable] durational structure.
Venditti et al. present the results of a perception study where participants were asked to rate the interpretation of each utterance on a set of semantic scales. The rise-fall pattern led to a strong sense of continuation and explanatory intention. Therefore, in their production study, Venditti et al. recorded speakers reading scripted dialogues to investigate the phonetic properties of five different types of BPM. To induce the rise-fall pattern, speakers were given an explanatory context. According to the data from two speakers, the average peak height of the rise-fall BPM was slightly lower than the peak of the accent preceding the BPM. As shown in Figure 5, this was not the pattern in our spontaneous speech data. The size of the tonal hump for a BPM is often much larger than the lexical accentual tonal movement. Often times, the tonal movement for a lexical accent was much less visible compared to that of BPM, as shown for tamane’gi-NO. The data from three speakers also show that the occurrence of BPM varies across individuals. The total numbers of BPMs (including prom, insist, and rise-fall types) were 23, 121, and 17, respectively. Interestingly, all speakers rarely produced BPM when they mentioned the ornament to pick (e.g., buru u-no tama ‘blue ball’), but frequently produced BPM when they mentioned the location of the ornament in subsequent utterances. The speaker who produced the utterance in Figure 5 had a tendency to start the utterance with sore-o ‘that-acc’ to mention the location of the ornament, and the overwhelming majority of her sore-o were produced with rise-fall L%HL% BPM. These results suggest that speakers were sensitive to the presence of the text label in their BPM production, so that utterances mentioning the text tags did not show this pattern, but those in subsequently planned speech did.

As mentioned earlier in section 3, Ito (2002) never observed rise-fall BPM in the read speech of female speakers of comparable age. The higher occurrence of rise-fall during our tree decoration task may have to do not only with the reference to text/labels, but also with the speaker’s explanatory role as a Director, consistent with Venditti et al.’s observation about the explanatory nature of rise-fall BPM. However, rise-fall BPM (HL%) is also known as a characteristic of casual speech by young females (Venditti et al., Venditti, Igarashi, Maekawa, in personal communication5). Therefore firm conclusions about the function of BPM will depend on future comparisons across different discourse purposes and speakers of different genders.

5 We are grateful for all the insightful comments from Kikuo Maekawa, Jennifer Venditti, Yosuke Igarashi, and Mary Beckman.
5. Conclusion

Investigating the prosodic characteristics of spontaneous speech is, without a doubt, a laborious and time consuming process. However, we have provided evidence of distinct benefits stemming directly from the use of naturally produced speech data. Most notably, we have found that spontaneous productions can include prosodic patterns unobserved in scripted recordings. The examination of spontaneous utterances has been made more tractable with the development of real-world object manipulation tasks, where investigators constrain the set of target words and phrases to be used by speakers, and control the degree of speaker’s ambiguity/contrast-awareness. Designing such tasks requires consideration of the conversational interactions that lead to natural discourse structure, effective elicitation of target phrases/words, and the means to gauge the speaker’s attention and awareness of pragmatic/syntactic contrasts in question. The analysis of the natural speech data should include identification of discourse structure; independent annotation of prosodic structures; replicable, objective acoustic measurements; and attention to dialectal and individual variations. Our on-going production study with the tree decoration task attempts to address each of these issues. Although the analyses are still to be completed, a parallel cross-linguistic eye-tracking study that we have conducted has benefited from the previous findings presented in this chapter. More specifically, the auditory stimuli created for the eye-tracking experiment reflect the prosodic patterns observed in natural speech. By using naturalistic prosodic patterns in the auditory instructions, we may assume that the listener’s reactions to those instructions reveal aspects of the comprehension mechanism that are also active during spontaneous oral communication. We believe that advanced investigation of prosodic effects on speech comprehension will be more accurately modeled as we have a better grasp of spontaneous prosodic patterns.

References


Beckman, M. E., and J. Hirschberg (1994): The ToBI Annotation Conventions. Manuscript, The Ohio State University, Columbus, OH.


boundary pitch movements for speech synthesis. In N. Campbell, et al. (eds)
Selected Papers from the Third ESCA Workshop on Speech Synthesis (tentative
Venditti, J. J., and M. Swerts (1996): Intonational cues to discourse structure in
Japanese. Proceedings of the International Conference on Spoken Language
7-486.