The Parsing of Prosody

Mary E. Beckman

Department of Linguistics, Ohio State University, Columbus, Ohio, USA

Prosody has been demonstrated to play some role in parsing the grammatical structure of utterances of texts that might otherwise display surface (syntactic) ambiguity. In the context of this research, it is important to be able to state with reasonable certainty what the prosody of an utterance is. This paper considers the fact that the prosody is itself a grammatical (phonological) structure that must be parsed. First, prosodic categories that determine or are marked by the intonational pattern are described for English and Japanese, concentrating on the categories “pitch accent” and tonally marked “phrases” at levels of the prosodic hierarchy above the (prosodic) word. Then various potential ambiguities in parsing these categories are discussed, with attention to both the phonological and the phonetic contexts which are most conducive to ambiguity between alternative prosodic parses, and to the implications of these ambiguities for claims about the relationship between prosody and syntax.

THEME

The stated theme of this special issue of Language and Cognitive Processes is ambiguous. Originally and probably primarily, “prosody and parsing” was intended to invoke discussion of the role of the prosody of an utterance in parsing its syntactic structure and semantics—knotty questions ranging from whether there are direct and consistent intonational and durational correlates of syntactic boundaries to how accent placement works in...
interpreting the semantics of complex noun phrases or in determining the
truth conditions of sentences containing words such as *only*. But the phrase
could also refer to the parsing of the utterance's prosody itself—and this
alternative meaning is the topic of this paper. That is, rather than elaborate
on yet another aspect of the ways in which prosody enhances or enables the
parsing of other sorts of linguistic structures, I will discuss the ways in which
the physical medium and the grammar of sound constrain the parsing of
prosodic structure. In order to make the discussion cohere as strongly as
possible with the other theme, I will concentrate on several specific examples
of how phonetic ambiguities in the prosodic parse complicate our
understanding of its role in these other parsings. Moreover, in order to be
able to describe the examples with the necessary detail, I will limit them to
just two domains: ambiguities of intonational phrasing and ambiguities of
accent placement.

To adequately describe these ambiguities in the prosodic parse, it will be
necessary to introduce some background technical information, beginning
with some explanation of the implicit claim that prosody is indeed something
that needs to be parsed in perceiving linguistic structure in the stream of
speech. We can tell that prosodic structure is parsed because it is
language-specific and prone to cross-language interference and misparsing.
For example, uninitiated native speakers of English have been known to
misinterpret the long closure interval of Japanese geminate stops as a pause
and thus to mishear a necessarily word-internal syllable sequence as being
divided across a strong syntactic boundary. Native speakers of Japanese, on
the other hand, frequently equate the nuclear fall of a common English
declarative intonation with the lexical pitch accent of Japanese, and so tend
to produce what sounds like far too many nuclear stresses in speaking
English. To understand either of these examples of prosodic interference,
we need to have a precise understanding of the phonological structures
involved. For example, to understand why there are too many nuclear
stresses in Japanese-accented English utterances, we need to understand the
relevant phonological differences between the roles of pitch accent in the
Japanese and English intonational systems.

Similarly, to understand how phonetic ambiguities within a language can
constrain the contribution of prosodic phrasing and accentuation to the
parsing of other grammatical structures for an utterance, we must
understand the relevant language-specific prosodic categories and their
phonological markings. In order to minimise the amount of such technical
background information that must be introduced, I will concentrate on
examples having to do with aspects of the intonational pattern of the
utterance (specifically in the domains of intonational phrasing and
accentuation), and will limit the source languages to just English and
Japanese (the minimum number necessary to show that the structures are
indeed language-specific). The next two sections of the paper, therefore, elaborate on the notion that prosody is a structure in its own right and, in the context of this elaboration, present the background information on the English and Japanese intonation systems. Then I deal with the relevant ambiguities of phrasing, and with the ambiguities involving accent.

A STRUCTURAL VIEW OF PROSODY

In describing the phonological grammar of intonational patterns for a language, there are two things a linguist must provide: a way to specify all of the relevant tonal categories in the tune (the pitch pattern) and a way to specify how this tune aligns with the text (the segmental string) of the utterance. For both sorts of specification, I assume the account of English and Japanese intonation originally developed by Pierrehumbert and her colleagues (see, e.g. Liberman & Pierrehumbert, 1984; Pierrehumbert, 1980, 1993; Pierrehumbert & Beckman, 1988; Pierrehumbert & Hirschberg, 1990) and recently incorporated in the ToBI transcription systems for English (Pitrelli, Beckman, & Hirschberg, 1994) and Japanese (Venditti, 1994). This account is couched in the framework of Autosegmental and Metrical Theory (see Goldsmith, 1990, for a review). This is a generative phonological framework in which the tune is specified using an independent string of tonal segments, and the prosody of an utterance is viewed as a hierarchically organised structure of phonologically defined constituents and heads. This view contrasts with an older generative framework (e.g. Chomsky & Halle, 1968), in which the only organising constituent in the phonology was the (alphabetic) segment, the minimal constituent for grouping together synchronously specified distinctive features of phonetic content, such as \{+nasal\} and \{+coronal\} for the segment \{n\}. In this older view, “prosody” was simply another pretheoretical cover term for “suprasegmentals”—features such as \{+H(igh pitch)\} or \{1stress\}, which were awkward to fit into the strictly alphabetic framework. In metrical theories, on the other hand, “prosody” has a precise technical meaning. It does not designate any set of distinctive features, but instead refers to the “meter”—that is, to the structural organisation itself.

This notion of prosody as raw organisational structure is not original to current metrical theories, especially as the notion is applied to the phenomena of pitch accent and stress. For example, more than 50 years ago, Arisaka (1941) proposed that pitch patterns in Japanese function primarily to organise words into a succession of “syntagms”, a proposal that was echoed in Trubetzkoï’s (1958) independent formulation of the “delimitative” and “culminative” functions of pitch accent and stress in languages such as Japanese and English (as opposed to the “contrastive” function of tone in languages such as Cantonese and Yoruba). However, it
was not until the late 1970s that the current metrical theories began to extend the notion to a larger range of prosodic phenomena by formalising it in an explicit representation of a hierarchy of organisational constructs that are realised in the arrangement of segmental and tonal properties at many more levels than Arisaka’s “syntagm”. It probably took generative phonologists so long to adopt this understanding of prosody and to formalise it as an all-pervasive organisational hierarchy because the idea is itself a difficult one. Arisaka and Trubetzkoy limited its application to a single level of organisation above the segment, so that their contemporaries could easily misinterpret their statements to be another way of describing the merely “suprasegmental” nature of pitch accent and stress, and even today, nearly two decades after Liberman and Prince’s (1977) first formal metrical phonology treatment of English stress, it is not uncommon to see prosody used as another term for “suprasegmentals”, as in “Prosodic features such as pitch, temporal structure, loudness” (Vroomen, Collier, & Mozziconacci, 1993, p. 577), or “the importance of pitch, and its physical correlate, the $F_0$ contour, as a prosodic feature of German” (Mixdorff & Fujisaki, 1994, p. 2231). Since much of what I say below about prosodic phrasing and stress crucially depends on an appreciation of this notion of prosody as organisational structure, let me first amplify with a metaphor, by building a “phonological” grammar for the sample of “words” and “utterance” in Fig. 1.

Consider first the distinctive letter shapes that contrast the isolated “word” strings in the upper panel of the figure. We can specify the inventory of segments by a generative grammar in which the unmarked “o” shape can be modified by adding an extender line or by stretching the shape vertically to make a larger ovoid capital letter. Segments with extenders can be further sub-classified by whether the line is specified to extend upward (“b” and “d”) or downward (“q”) from the “o” and, if it extends upward, whether it is to the right or the left. Enlarged ovoid segments can be further sub-classified by whether the oval is continuous (“O”) or open to the right (“C” and “G”), and, if open, by whether the opening has a cross stroke across the lower lip (“G”). The inventory of paradigmatic contrasts between {+ extender} versus {-extender}, or {+ stretch} versus {-stretch}, and so on, constitutes the distinctive features that define the letter segments of the language. In the word strings in the upper panel, there is some hint of syntagmatic organisational properties in addition to these paradigmatically contrasting distinctive features. For example, there are no strings where “q” occurs singly. Where there is one “q”, there are always at least 10 other identical segments around it. The lower panel of the figure shows the full “utterance” from which these isolated strings have been extracted, and reveals the prosodic structure defined by these syntagmatic constraints. The positions where {+ extender} and {+ stretch} segments can occur are defined by the
cube structures. Or, to put it in terms of the needs of the reader rather than of the typist, these segmental features define the cubes; they allow the viewer to parse the cube structures in the matrix of segments on the page. The syntagmatic constraints can now be understood in terms of how the letters in different strings must be positioned with respect to each other in order to form three visible faces of the cube. We could represent these positional properties by specifying “suprasegmental” distinctive features \{+cube\} and \{+top face\} for the letters in addition to the “segmental” features \{+extender\} and so on, but a better notation would be one that is more directly related to the physical properties of the cube and of how the number and the shapes of the faces depicted on the page relate to the angle of view. The structural constraints on where the segmental features occur can then be stated in terms of their functions in defining the visible faces of the cubes so specified. The more direct representation has an added advantage of being agnostic about any ordering of priority between the segmental content of the structure and the structure proper. The grammar need not commit to a serial processing architecture which first specifies the segmental features and then derives the suprasegmental features from them, or the other way around. Rather, the two aspects of the utterance can be specified simply as constraints on the output. Thus the specification that a particular set of letters are all \{+extender\}, and the fact that these letters contrast with the surrounding background field of \{-extender\} letters to define the faces of the right-facing cube, are two aspects of the same view, which will be produced in parallel by the typist and perceived in parallel by the reader.

Returning now to the analogous notion of prosody as the organisational structure of speech, the representational devices of autosegmental and metrical phonology are like the more direct grammar that we would like to have for specifying cubes in Fig. 1. They have allowed us to go beyond the unsatisfactory definition of prosody as “suprasegmental” by redefining and eliminating the traditionally “suprasegmental” features in two ways. First, by allowing several parallel autonomous streams of specification for different “autosegments”, a tonal category such as \{+H\} can be treated as merely another segmental content feature, no different from \{+coronal\} and \{+nasal\} except in that it is typically specified for an entire prosodic constituent such as a syllable rather than just for an onset or coda consonant at the edge of the syllable. This is analogous to specifying \{+extender\} over the entirety of the right-facing cube, as opposed to specifying \{+down\} only over the top face of that cube. Second, by providing a direct representation of the syllable and of phonological categories below and above the syllable, the historically problematic features of \{1stress\} (“primary stress”), \{2stress\} (“secondary stress”), and so on, can be replaced by a more direct iconic representation of the multi-layered structures that give rise to these different levels of prominence contrast for syllables. This is analogous to using a more
direct representation of the cubes and their visible faces to replace the segment-by-segment specification of *ad hoc* “suprasegmental” features such as \{+cube\} and \{+top face\}. As with the cube representation in Fig. 1, this more direct representation of stress as structure *per se* can be agnostic about any ordering of priority between the occurrence of distinctive features that mark various positions in the organisation of speech and the organisational structures that provide the positional constraints on the occurrence of the distinctive features. (This is an important point to understand for the description in the next section of how stress figures in the prosodic hierarchy of English.)

Different metrical theorists propose different iconic representations for the various phonological structures that organise the speech stream. Broadly speaking, the main differences arise from paying closer attention to
constituent edges or to heads (two different kinds of position in the structure). For example, many languages give special status to the edge of the intonational phrase—a phonological constituent that can be defined for English and Japanese and many other languages in terms of the distribution of a tonal contrast between H% and L%. These are high- versus low-pitched boundary tones that in standard Southern British English, for example, distinguish a fall–rise tune commonly used in yes–no questions from the simple fall common in citation form declaratives. There are other quasi-universal defining features for the intonational phrase as well, such as the distribution of the most felicitous locations for pause. Often, too, there are language-specific features involving the blocking of some segmental sandhi phenomena. Thus in Korean, all word-initial obstruents are nasalised when the preceding word ends in a nasal, unless an intonational phrase boundary intervenes (Jun, 1992). Metrical phonologists who investigate such sandhi phenomena (e.g. Nespor & Vogel, 1986) tend to adopt strictly layered prosodic tree representations, in which segmental material for each constituent at any given level is parsed totally into one or more constituents at the next lower level, and is included in exactly one constituent at the next higher level. Such a tree representation is ideal for capturing the locations of the boundaries of the intonational phrase, and the relationship of necessary congruence between boundaries of larger higher-level constituents and those of lower-level constituents. That is, the final segments of the last syllable in an intonational phrase are also final to that intonational phrase. Thus, saying that a tone marks the right edge of an intonational phrase in effect specifies that the speaker will produce the pitch pattern that expresses the tones to be coincident with the spectral patterns that express the rhyme segments of the last syllable, and that the hearer can interpret the intended prosodic structure by perceiving the prosodically conditioned alignment between tune and text.

On the other hand, there are also salient prosodically conditioned phenomena that seem to bear little or no relationship to constituent edges. For example, in English and other languages with similar intonation systems, certain tones are realised concurrently with the segments for the head syllables of the most prominent words in the intonational phrase. Narrow focus of information can be signalled without explicit syntactic marking simply by the location of these pitch accents. Therefore, these tones can occur in any position with respect to the phrase edge. But the phonotactics of the English intonation system require that they occur only on prosodically “strong” head syllables. If narrow focus of information singles out an inherently “weak” syllable, as in Bolinger’s (1961) example in (1), the requisite prosodic status of the syllable as a head syllable induces a nonce pronunciation as a “strong” syllable:
1. This whisky wasn't EXported from Ireland; it was DEported.

Thus in this example, the vowel in the first syllable of deport will be a long tense \(i:\), rather than the lexically specified reduced lax front vowel or schwa. Metrical phonologists who study such stress-conditioned phenomena tend to adopt the metrical grid or grid-compatible representations of prosodic structure (e.g. Hammond, 1986; Prince, 1983), instead of the head-neutral nested constituents of the standard strictly layered prosodic tree. However, since the examples in this paper are of both types, the following descriptions of the English and Japanese prosodic hierarchies will be agnostic about such fine points, and simply describe each of the relevant levels in terms of whichever representational device is more amenable to the phenomena defining the level. The levels of interest in English and Japanese, then, are the following.

**PROSODIC CONSTRUCTS RELEVANT FOR INTONATION**

First, we will be concerned with the *intonational phrase*. As suggested above, this constituent is common to both languages and is defined by the distribution of \(H\%) and \(L\%\) boundary tones. As the name suggests, a “boundary tone” is a distinctive tonal specification which is *aligned* with (i.e. realised in the pitch pattern in such a way as to be phonetically coincident with) the one or two segments nearest to the phrase boundary. (The wording here should not be taken to mean that the segments are somehow primary. One could say equivalently that “the segmental specifications are aligned to the last tone specified for the phrase”. There is no logical priority in the phonological representation of the prosodic constituency mediating the alignment, and indeed, the phonetic patterns argue against such a serial processing view. In many languages, including English and Japanese, there are usually not enough tones specified in a phrase’s tune to realise the spectra of each consonant and vowel concurrently with some targeted pitch value, and yet the last vowel, and any following coda consonant if it is voiced, will lengthen to accommodate the pitch pattern when there are several tones in quick succession at the boundary.)

In both English and Japanese, the relevant boundary for the tonal mark of the intonational phrase is the final one. That is, in both languages, each intonational phrase has an obligatory specification for a \(+H\) \((H\%)\) or \(-H\) \((L\%)\) tone to be aligned to the segment(s) at its right edge. Figure 2 shows a pair of English utterances, each broken into two intonational phrases, but contrasting in the types of boundary tone marking the end of the
first and second phrases. Figure 3 shows an example of a Japanese utterance that is clearly marked as three intonational phrases by the rises to the H% boundary tones at the end of the negative *Nnnn*, and at the end of the declarative particle *-yo*.

(In these and other figures, I use the fundamental frequency contour as the phonetic representation of the tune, and plot the F0 values on a linear rather than a log scale because the psychoacoustics literature suggests that a log scale is not correct for representing linguistically relevant aspects of pitch level relationships. See Pierrehumbert and Beckman (1988, pp. 22–23) for further discussion, and Hermes and van Gestel (1991) for more recent evidence against a log scale for speech fundamental frequency.)

The intonational phrase is the highest-level prosodic constituent that can be defined in either language in terms of such qualitative phonological markings. At the other end of the prosodic hierarchy is another constituent common to both languages, and indeed to all languages—namely, the syllable—which is defined by language-specific phonotactic conventions that can be summarised cross-linguistically by the principle of alternation between less sonorous edge elements (consonants) and more sonorous head elements (vowels). While the definition of the syllable is not directly relevant to the examples of ambiguity at the intonationally defined levels that I will discuss in the next three sections of this paper, it is useful for illustrating the general characteristics of ambiguities at all levels of the prosodic hierarchy. Structure is least ambiguous when it is clearly marked by a salient phonetic contrast between adjacent alternating elements, as in the perfectly alternating CVCVCV sequence that syntagmatically contrasts stops and open vowels in the American place name *Catawba* or the Japanese surname *Takada*. It becomes ambiguous when the alternation is not a perfect binary one or when the phonetic contrast is less salient. For example, the lack of a clear alternation between consonant and vowel after the first CV in *towel*, *mayor* and *higher* makes these words highly ambiguous between a disyllabic and a monosyllabic parse, and there are American English dialects where they have become homophones with the monosyllables *tile*, *mare* and *hire*. Similarly, the very small sonority contrast between the medial reduced vowel and the following sonorant consonant in words such as *family* and *chocolate* make these prone to reparsings without the medial syllable. And, indeed, in some dialects, these have become lexicalised as disyllabic words. {Japanese also has historical reparsings involving the phenomenon of high-vowel devoicing. See Beckman (1994) for a discussion of such inherent ambiguities in English and Japanese and in several other languages that are typologically similar to each.}

In discussing English intonational categories, it will be useful to talk also about a constituent just above the syllable, the foot, defined as a “heavy”
Anna came with Manny.
FIG. 2. Fundamental frequency contours for utterances of the sentence *Anna came with Manny*, each produced as two intonational phrases, and illustrating L% versus H% boundary tones at the utterance–medial and utterance–final phrase boundaries. The rendition in (a) puts the greater “answer” focus on *Manny* and backgrounds *Anna*, whereas the rendition in (b) puts greater focus on *Anna* and backgrounds *Manny*. 
FIG. 3. Fundamental frequency contour of an utterance of the sentence *Nnnn, Kazumi-ga yonda yo, Ma'ri-o* "No, Kazumi did call [her]. Mari, [that is]", produced as three intonational phrases with medial H% tones marking the phrasal boundaries before the main clause and between the verb and the postponed object NP. The apostrophe after the vowel of the first syllable in the transliteration of the name Ma'ri is the standard way (after McCawley, 1968) of indicating that there is a pitch accent on that syllable (see the discussion of accent in Japanese below). The utterance was provided by Jennifer Venditti, from recordings for experiments in progress on phrasing and pitch range in Japanese.
syllable plus any following “light” syllables. Here “heavy” means a syllable with some minimum structural specification for stress, as marked by the distinctive features of its head vowel. A syllable is “heavy” if its head is a tense vowel or diphthong or is an unreduced lax vowel (in which case there is an obligatory following coda consonant), as in the first and third syllables of the word *automatic*, as opposed to its second and fourth syllables. Thus *automatic* comprises two feet. It also exemplifies the principled optimum, a perfect alternation between stressed and unstressed syllables. Example (2) shows a tree and a grid representation for the word, representations which emphasise the foot-level constituent structure and the rhythmic alternation between strong head syllable and weak trailing syllable, respectively:

2. Tree and grid representations of stress pattern in *automatic*

\[
\begin{array}{c}
F \\
\sigma_1 \sigma_2 \\
\text{au to ma tic} \\
\text{automatic}
\end{array}
\]

As with the syllable, less perfect alternations can be conducive to ambiguity. The heavy first syllable in the Southern British pronunciation of the word *automata*, for example, is stressed at this level of the hierarchy, and thus contrasts with the light first syllable in the word *atomic*. However, the results of an experiment in which Fear, Cutler and Butterfield (1995) cross-spliced the first syllables of *automobile*, *automatic*, *automata* and *atomic*, and then asked native speakers of Southern British English to make naturalness judgements of the original and cross-spliced versions, suggest that the stressed status of the first syllable of *automata* is less salient than that of the first syllables in *automobile* and *automatic*, where the heaviness from the vowel specification is reinforced by the contrast to the following light syllable. Further evidence for this ambiguity inherent to non-alternating stress patterns comes from historical reanalyses (analogous to the reanalyses of *family* and *chocolate* as two syllables). For example, in many American English dialects, *automata* has undergone sound change to have a reduced first vowel identical to that in *atomic*. The British English pronunciations of the words *raccoon*, *baboon*, etc., with reduced first syllable where the American pronunciations have preserved the heavy {æ}, are similar examples of diachronic reparsing. As suggested above, the metrical grid is a good device for depicting the contribution of rhythm to the structural ambiguity here. Example (3) gives metrical grid representations of the contrasting stress patterns in the British and American pronunciations of *automata*:
3. Grid representations of British (left) and American (right) pronunciations:

\[
\begin{array}{ccc}
 & x & x \\
\text{automata} & x & x \\
\end{array}
\begin{array}{ccc}
 & x & x \\
\text{automata} & x & x \\
\end{array}
\]

As with the structural representation of the cubes in Fig. 1, the structural representations of stress in (2) and (3) circumvent the need for a priori decisions about derivational primacy or processing priority, in contrast with the older strictly linear Generative Phonology representations in which the stress specification was a surface suprasegmental feature derived by rule for syllables from prior underlying vowel specifications. That is, the tree representation and the grid equivalently allow a phonological account which merely specifies that this set of segmental features is the qualitative phonological mark of this structural position in well-formed surface representations. Thus the question of which contrastive feature specifications must occur for this level of prominence specification is separated from the questions of which specification is prior, or even whether either is prior—that is, the representations in (2) might be understood simply as a parsing filter for a model in which the two specifications are computed in parallel in production and perception. [Pierrehumbert (1993) suggests that such a constraints-based account is more compatible with speakers' behaviour when accenting lexically reduced syllables, as in example (1). I will return to this below.] The latter set of issues, then, can be addressed properly in their own right, in experiments designed to uncover human processing strategies.

Although the syllable and the foot themselves are not defined in terms of the intonational pattern, they are crucial prosodic constituents for talking about intonational patterns in English, because any syllable which is the head of its foot can be a potential site for the other main type of intonational event besides the edge tones, namely the *pitch accent*. Cross-linguistically, we can define a pitch accent as a tone (or a closely linked group of tones) that is contrastively localised to a syllable. That is, a pitch accent is phonologically specified to occur around some designated syllable, and utterances can minimally differ in which syllables are so designated. In the terms of Autosegmental Phonology, the tone (or tones) of the pitch accent is “associated” with the syllable. A syllable that has a pitch accent associated with it is also called “accented”.

The pitch accents of English are intonational morphemes that encode various pragmatic relationships between the discourse context and the particular semantic value that is composed for the text by choosing the focused word or phrase. For example, a {+H} accent signals commitment to
that particular value and its relevance to the discourse, while a \{−H\} accent signals lack of commitment. A pitch accent can be a single tone specification, such as \{+H\}, or it can be a sequence of two tones, such as \{−H\} \{+H\}. Bitonal pitch accents contrast as well in whether it is the first or the second tone specification that is phonologically associated to the accented syllable, with the preceding or trailing unassociated tone being realised at some fixed temporal distance regardless of the alignment to the segments of the accented syllable. We adopt Pierrehumbert's convention of writing pitch accents with plus signs linking the two tones of bitonal accents and an asterisk marking the associated tone; for example, H*, L*, L + H*, L* + H, with the last two bitonal accents in this list contrasting in whether the high- or low-pitched tone target aligns to the accented syllable. \{See Beckman and Ayers (1994) for a tutorial introduction to the full set of pitch accents types differentiated in the ToBI transcription system, and Pierrehumbert and Hirschberg (1990) for a description of their pragmatic meanings.\}

As suggested above, pitch accent association in English refers to the discourse-related notion of focus of information. Accents are placed to indicate which words or phrases are most salient to the purpose of the discourse, and there is a vast and old literature on how syntactic structure interacts with the phonology of “stress” to constrain the scope of the focus indicated by any particular accent placement. \{See Ladd (1980) for a still relevant review of the literature up to the early 1980s, and Selkirk (1984) and Rooth (1992) for several important developments in the literature since then.\} What is relevant from the phonology are three constraints on what constitutes a well-formed text–tune association.

First, a syllable must have a minimal segmentally defined stress prominence in order to be accented. That is, a pitch accent can associate only to a syllable which is the head of its foot. Therefore, when a speaker wants to focus on a particular syllable (as opposed to the word containing the syllable) and that syllable is lexically unstressed, then the syllable must be promoted to being the head of a foot. This is the source of the nonce pronunciation for deported in example (1) above.

Second, it is difficult to produce a long stretch of text before the first tonal specification in a (prosodic) phrase, particularly if the phrase is the first in the utterance. Therefore, when the encoding of focus specifies only that there should be a single pitch accent late in a phrase, speakers tend to place at least one more pitch accent on an earlier heavy syllable. This constraint is the source of the perceived reorganisation of stress relationships in bipedal words such as automatic or Chinese when they occur early in a phrase such as automatic bill-paying or Chinese lantern \{Beckman & Edwards, 1994; Shattuck-Hufnagel, Ostendorf, & Ross, 1994\}.

Third, there must be at least one pitch accent somewhere in every (prosodic) phrase, and the placement of the accent or accents structurally defines two qualitative levels of stress prominence above the minimum level
defined for the head of a foot. That is, if there is only the one obligatory accent, it must associate to the “most stressed” syllable in the phrase, and if there are two or more accents, the last pitch accent typically associates to the “most stressed” syllable in the phrase. (The “most stressed” syllable is determined by the syntactic constraints on the interpretation of focus mentioned above and also by lexical specification in words with two or more feet, such as mailman, automatic or antihistamine. That is, in such multi-pedal words, one of the heavy syllables is lexically specified as the “primary stress”, to be the one that associates to the obligatory pitch accent when the word is produced in isolation or is focused in the discourse.)

This third constraint seems to be the source of patterns which older linear generative phonology accounted for by the “nuclear stress rule” (Chomsky & Halle, 1968), and so we adopt the name “nuclear accent” to designate the last pitch accent and the term “nuclear accented” to describe the last accented syllable in the phrase. That is, a nuclear accented syllable is a syllable that would have been assigned “primary stress” in these older accounts. Of course, this equation is not an exact one, since Chomsky and Halle (1968) understood the patterns in terms of the serial processing architecture of derivational rules, a viewpoint which has taken a long time for phonologists to overcome. For example, it still very much coloured Selkirk’s (1984) understanding of the relationship between accent and higher grid levels, so that she felt forced to decide between a “pitch-accent-first” account and a “stress-first” account. However, if accents “come first” (in order to account for how accents are assigned to words under focus), and only then are associated with the accented syllables by docking on the sites lexically specified for “primary stress”, it is difficult to see how to generate two or more accents on a word, as in very deliberate pronunciations of words such as automatic. Figure 4 gives an example of two such cases in a single utterance. (See Pierrehumbert (1993) for a recent recasting of the basic facts as constraints, in an account that allows for pronunciations such as these.)

In introducing the second and third of these three phonological constraints on accent placement, I have deliberately avoided specifying what is meant by “(prosodic) phrase”, because to understand precisely this constraint, we must introduce another prosodic constituent. In English there is a second level of intonational phrasing below the intonational phrase proper, a level that is very similarly defined by the distribution of tones marking its right edge. Each full intonational phrase consists of at least one “intermediate phrase”, marked again by a contrast between [+H] and [−H] edge tones, so that a full intonational phrase actually ends with a sequence of two such specifications, one for its own right edge and one for the right edge of the last intermediate phrase within it. The recognition of two levels of intonational phrasing for English is quite old (cf. Trim, 1959), but Pierrehumbert’s account of the English intonation system (see Beckman & Pierrehumbert, 1986) is the first to formalise the difference by recognising
FIG. 4. Fundamental frequency contour for a spontaneous utterance of the sentence fragment *from Indianapolis Indiana to Kansas City, round trip (non-stop)*, with two pitch accents on each of the words *Indianapolis* and *Indiana*. Utterance from the Boston University ATIS data, used with the permission of Mari Ostendorf.
that the two levels of phrasing are distinguished by the behaviour of the edge-related tones. Unlike the H% versus L% specifications for the intonational phrase proper, the \{+H\} or \{-H\} tone for the intermediate phrase is not localised strictly to the phrase boundary. Instead, the tone fills up the space after the last phrase-internal tonal specification, which is the obligatory or last pitch accent (i.e. the nuclear accent). In several of the most common English tunes, the edge-related tone for the intermediate phrase differs in pitch level from the nuclear pitch accent, giving a fairly sharp and salient prominence-lending pitch fall or pitch rise at the stressed syllable, and in phrases where the nuclear accent is quite early, there can then be a fairly long high- or low-pitched “tail” realising the edge tone. Pierrehumbert (1980) recognised these peculiarities of the tone defining the intermediate phrase by calling it the “phrase accent” and notating it with H- or L- rather than H% or L%. The ToBI transcription system for English also adopts these conventions and reserves the term “boundary tone” for the edge tone defining the higher-level intonational phrase.

Because of the constraints that every intermediate phrase must have at least one (nuclear) pitch accent, and that the last accented syllable is typically the “strongest”, this level of the constituent hierarchy is crucial in the structural definition of stress levels for English. Incorporating accent, three levels of stress prominence can be qualitatively differentiated (see Beckman & Edwards, 1994; Vanderslice & Ladefoged, 1972). First, a heavy syllable (the head of a foot) is more prominent than (i.e. stressed relative to) any light syllable. Second, although a syllable must be the head of a foot to be accented, not every heavy syllable need bear a pitch accent. Thus, in an intermediate phrase containing more than one foot, any accented syllable is stressed relative to any heavy syllable that is not accented. Finally, although an intermediate phrase must contain at least one accented syllable, it can contain more than one, in which case typically the last accented syllable is more prominent than any preceding accented syllable. Thus, when the word automatic is produced as a full intonational phrase (and hence a full intermediate phrase) all by itself, as in a citation form pronunciation, it must have a nuclear accent on its second heavy syllable, and may also have a pre-nuclear pitch accent on its first heavy syllable, as depicted in (4):

4. Grid representations of the stress pattern for citation pronunciations of the word automatic, produced with just the obligatory nuclear pitch accent (left) or with two pitch accents (right):

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>nuclear accented</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>x</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>accented</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>x</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>heavy</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>syllable</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

| automatic | automatic |
This convention of typical nuclear accent placement means that speakers can manipulate main and subordinate prominences to highlight all content words or to focus narrowly on one. For example, a phrase with a pre-nuclear accent on an early content word and nuclear accent on a late content word is likely to be interpreted as broad focus over the phrase as a whole, whereas narrow focus can be indicated on a word early in an utterance by placing the nuclear accent on that word, so that any subsequent content words are “deaccented” (i.e. produced in the post-nuclear tail where the phrase accent precludes the realisation of any pitch accents). This also means that to produce an utterance with double foci explicitly on two different words (i.e. to make two equally stressed nuclear-accented syllables), it is necessary to make two intermediate phrases, with an intervening intermediate phrase break marked by a phrase accent, as illustrated by the utterances in Fig. 5. The phrasings and stress patterns are represented in example (5). Note that in (5), as in (3) and (4), I follow Selkirk (1984) in choosing to represent the stress pattern with grids rather than with trees, even though this obscures the relationship between the prominence of nuclear accent and the intermediate phrase. Although the nuclear accent beats and the heavy syllable beats in the grid could be represented as the heads of intermediate phrases and stress feet in a tree structure, as in (2), there are no qualitatively marked prosodic constituent edges between the foot and the intermediate phrase to allow such a representation for the beats of pre-nuclear accents (see Beckman & Pierrehumbert, 1986).

5. Grid representations of the three stress levels and tree representations of the intermediate and intonational phrasings for the sentence *Anna came with Manny*, produced (a) as one intermediate phrase, as in Fig. 5a, (b) as two intermediate phrases, as in Fig. 5b, and (c) as two full intonational phrases, as in Fig. 2a or 2b:

![Diagram](image-url)
Anna came with Manny

H° L− L°
FIG. 5. Fundamental frequency contours for utterances of the sentence Anca came with Manny produced as (a) one intermediate phrase versus (b) two intermediate phrases. The one-phrase rendition in (a) has a pre-nuclear accent on Anna which makes its (first) head syllable less stressed than the head syllable of Manny, whereas the two-phrase rendition in (b) has two nuclear accents and phonologically equal stresses on the head syllables of the accented words.
In summary, then, the prosodic hierarchy for English includes qualitatively marked stress contrasts at three structural levels. Moreover, while the lowest level of stress can be predicted from the specification of the vowel types in the speaker’s lexicon for the words in the text in the absence of any special narrow focus of the sort illustrated in example (1) for the first syllable of deported, the stress pattern at the higher two levels cannot be stated except by knowing where pitch accents are placed and whether an accent is the nuclear accent in some intermediate phrase. These intonational facts constitute the qualitative markers of stress at these levels. To be sure, there are concomitant quantitative differences in vowel quality and the like. For example, accented vowels tend to be longer and articulated closer to the periphery of the vowel space (see de Jong, 1995, for a review and some recent data). However, these are minor variations compared with the qualitative difference between inherently longer full vowels and inherently very short reduced vowels that categorically defines the stress contrast between heavy and light syllables at the lowest level of the stress hierarchy, and could be called ancillary to the tonal markers (Beckman & Edwards, 1994). Thus it is not possible to talk about stress at the two higher levels without explicitly or implicitly assuming an intonational pattern for an actual or an imagined utterance of the text.

This describes all of the prosodic categories that are relevant for discussing the English examples of phonetic ambiguity, and we can now turn to the relevant categories that make the Japanese prosodic system differ from English. It was noted above that Japanese, like English, has the syllable near the lower end of its prosodic hierarchy, and the intonational phrase, marked by an obligatory H% or L% final boundary tone, at the upper end of the hierarchy. Japanese also has a tonal event that fits the cross-linguistic definition of pitch accent given above, namely “a tone or closely linked group of tones that is contrastively localised to a syllable”. However, its role in the Japanese intonational system differs fundamentally from the role of pitch accents in English. First, Japanese does not have any contrast among different pitch accent shapes indicating different pragmatic meanings. Instead, there is only one shape—a simple fall on or after the accented syllable, notated in the Japanese ToBI transcription system as H* + L, using Pierrehumbert’s conventions for English. Moreover, Japanese has nothing like the English system of qualitative contrasts in syllable stress which relates accentuation to focus of information. Instead, accent placement is determined entirely in the lexicon and in the morphology of compound formation. That is, there are lexically accented words, contrasting among themselves in which is the accented syllable, and there are also contrasting lexically unaccented words. The lexically unaccented words are not a marginal type—unlike, say, the “unaccentable” monosyllabic function words of English, such as the, to and of, which might be considered lexically
specified as unstressed, but which must take an accent when they are produced in citation form, and then have nonce pronunciations with full vowels such as \{Di:\} for the, \{tu:\} for to, and so on. Rather, the unaccented words of Japanese are a substantial subset of the content words of the language, and they surface without accent even in citation form. Indeed, it is very easy to devise sentences containing nothing but such lexically unaccented words, and an utterance of such a sentence will contain no accented syllable. In other words, there is no constraint comparable to the “nuclear stress rule” of English.

On the other hand, Japanese is similar to English in that focus of information can be indicated prosodically without syntactic markings. That is, there are functional analogues that play the same discourse roles as do accent placement and nuclear accent position in English. One of the analogues is the patterns of phrasing at a level below the intonational phrase, the accentual phrase. The accentual phrase is a prosodic constituent defined intonationally by the distribution of two tones that together create a delimitative rise in pitch on the first or second syllable of the phrase. The end of the delimitative rise is specified by a phrasal H- tone, which is associated to the first or second syllable of the phrase, and the beginning of the rise is specified by a L% boundary tone marking the actual edge. The timing and extent of the rise is determined by the characteristics of the initial syllable. The phrasal H- associates to the first syllable if that syllable is accented or if it is “long” (i.e. if it is closed by a coda nasal or has a long vowel or diphthong as its head), in which case the L% is realised only at the phrase edge. Otherwise, the phrasal H- associates to the second syllable, and the L% associates to the first syllable and is realised over a more extended temporal domain. Pierrehumbert and Beckman (1988) called the associated and unassociated L% tones the “strong” and “weak” versions of this boundary tone, because the shorter duration and typically higher tone target of the weak L% makes the delimitative rise to the following phrasal H- less salient.

{There is some controversy over whether the L% originates as a “final” boundary tone marking the right edge of the preceding accentual phrase, or as an “initial” boundary tone marking the left edge of the current phrase, or whether there are actually two L% tones at every medial boundary, as suggested by the fact that there are L% tones at the end of a phrase before a pause and at the beginning of a new phrase after a pause. The phonetic consequences differentiating these three accounts are very subtle and hinge crucially on the model of phrases at higher levels. See the discussion in Pierrehumbert and Beckman (1988, pp. 85–87).}

Accentual phrasing is functionally analogous to pitch accent placement in English, because several words can be joined together within a single accentual phrase to indicate that they are not separate focus domains. When a content word is phrased together with a preceding word in this way, it is
said to be “dephrased” (in analogy to “deaccented” for English). This is illustrated by the contrasting phrasings of the same adjective–noun sequence in the first two utterances in Fig. 6. A phonological consequence of dephrasing is that only the first pitch accent is realised. That is, a “first accent wins” rule deletes the pitch accents of all subsequent lexically accented words if the phrase contains more than one. Thus the accentual phrase can be defined as the domain of at most one pitch accent, as well as by the distribution of the phrase-initial rise from L% to H-.

A second functional analogue in Japanese to accentuation in English comes from prominence relationships among the fundamental frequency peaks of neighbouring phrases. Phonologically these relationships reflect the specification of “pitch range”, which can be modelled roughly as the distance above an abstract reference line of the F0 target value of the H* of each phrase’s pitch accent (or of its phrasal H- if it is unaccented). If the pitch range on a phrase is large, there will be a large and salient rise from the preceding L% to the phrasal H- (followed by a large fall if it is accented). When a sequence of content words is phrased as a sequence of so many separate accentual phrases, typically the first phrase will have the largest pitch range, and subsequent phrases will be subordinated to the first by successively reducing their pitch ranges to produce a more or less gradual downtrend of the successive fundamental frequency peaks. A new focus domain can be indicated by a “reset” of pitch range and the beginning of a new downtrend, and narrow focus on a content word late in the utterance can be indicated by resetting to expand its pitch range relative to preceding phrases. Reset thus defines a level of prosodic grouping above the accentual phrase, a larger constituent which I will tentatively identify with the contrast between obligatory final L% and H% that defines the intonational phrase. [Because they assumed H% to be limited to utterance-final position, Pierrehumbert and Beckman (1988) took reset to be the qualitative mark of a level of phrasing intermediate between the accentual phrase and the intonational phrase. However, subsequent work by Venditti, currently in progress, shows that H% can occur without a following pause, as in Fig. 2 above. This suggests that there is no qualitatively marked intermediate phrase between the accentual phrase and the intonational phrase.]

There are phonological factors affecting both of these relationships among words or potential accentual phrases. First, the second of a sequence of two words is far more likely to be dephrased if one or the other of the words is lexically unaccented than if both are accented. Thus, the first production of /kimi’dori no kanarija/ in Fig. 6a is more typical than the second production in Fig. 6b. Second, if the first of two phrases is accented (i.e. contains a lexically accented word), there is a “downstep”—a steep and global reduction of pitch range that affects all subsequent phrases up to the next reset. Therefore, the downtrend over the sequence of accentual phrases
in an intonational phrase will be steeper the more accented phrases it contains. Panels (a) and (c) to (e) of Fig. 6 illustrate these typical patterns for all possible combinations of accented and unaccented items in utterances containing two content words.

In summary, then, focus of information in Japanese can be indicated qualitatively by making the focused word or phrase stand as the first in its higher-level intonational phrase, so that it is the first in a new downtrend sequence and any effect of downstep induced at a preceding accent is undone by the reset. That is, the focused item is not dephrased relative to preceding material, and instead has it its own independent specification of (expanded) pitch range. These phonologically governed manipulations of pitch range are thus the grammatical source of the phonetic prominence that marks informational focus, functionally equivalent to the intermediate phrasing and placement of the nuclear pitch accent in English. Other phonetic cues can also be produced concomitantly, such as lengthening all of the segments in the phrase and producing more peripheral vowel qualities for all of its vowels (see Maekawa, 1994a). These aspects of focus are reminiscent of the ancillary marks of nuclear-accented syllables in English, but they are not localised to any single most stressed syllable and seem rather to be more direct quantitative effects of the focus on the phrase as a whole. Accentual phrasing and pitch range reset have also figured prominently in discussions of the role that prosody plays in syntactic parsing in Japanese, and so we begin our account of phonetic ambiguities by considering where and how accentual phrasing can be ambiguous.

AMBIGUITIES OF ACCENTUAL PHRASING

Arisaka’s (1941) discussion of the syntagm (see pp. 19–20) might be understood also as a statement about the relationship between prosody and syntax. However, the modern understanding of the accentual phrase and its role in the relationship between prosody and syntax does not begin until Hattori explicitly equated the tonal patterns for the accentual phrase with the *bunsetu*, a syntactic grouping that corresponds roughly to a minimal NP, that is a noun plus any following postpositions (see various articles from the 1940s and 1950s reprinted in Hattori, 1960). McCawley (1968) adopted this quasi-syntactic definition and recast it in the phonological framework of the time by proposing generative rules to delete accents in postpositions after accented nouns and to predict accent placement within inflected forms of verbs, which he assumed would constitute accentual phrases of their own. Poser (1984) also related accentual phrasing to syntactic structure, stating again that a noun and any following postpositions will constitute a single accentual phrase, and describing various regular conditions for dephrasing in many larger syntactic structures. For example, he stated that a head noun will be dephrased after the modifying verb in a relative clause. Also, he
FIG. 6. The first two panels show fundamental frequency contours for the noun phrase /ki'mi'dori no kanarija/ “yellowish-green canaries”, produced (a) as one accented phrase and (b) as two accentual phrases. The sequence here is accented-unaccented. The remaining panels show fundamental frequency contours for utterances of (c) the accented–accented sequence /ki'mi'dori no hima'wari/ “yellowish-green sunflowers”, (d) the unaccented–accented sequence /ai'ro no hima'wari/ “indigo sunflowers”, and (e) the unaccented–unaccented sequence /ai'ro no kanarija/ “indigo canaries”. The contours illustrate downstep, which lowers the second accentual phrase peak in (b) and (c), and dephrasing of the second content word in the other utterance of the accented–unaccented sequence in (a) and in the two other sequences containing at least one unaccented word in (d) and (e). The recordings were provided by Jennifer Venditti, who explained them as “descriptions of curtain material” when eliciting them from the speaker.
claimed that constructions with auxiliary verbs such as *yonde-miru* “try reading” can be distinguished prosodically from VP sequences such as *yonde miru* “read and see”, in that the auxiliary verb will necessarily be dephrased but the VP sequence will be produced reliably as two accentual phrases. Selkirk and Tateishi (1988) offer another slightly different account, using Selkirk’s (1986) “end-based theory”. The end-based account predicts, for example, that an object and following head verb might constitute a single accentual phrase, but that other minimally contrasting sequences of noun and verb in which the noun is not dominated by the VP will never be dephrased. This is an important claim for the theme of “prosody and parsing” because the topic, subject and object case-marking postpositions -*wa*, -*ga* and -*o* are elided easily in casual speech. In short, a common assumption is that there is a fairly straightforward mapping between syntactic structure and prosodic structure, such that prosodic constituency at the level of the accentual phrase should play an important role in the on-line parsing of the syntax of a Japanese utterance.

However, other work suggests that the syntax–prosody mapping may not be so straightforward, or at least that it may not be straightforward to test such claims about the mapping, because of some fairly common conditions of ambiguity in the parsing of the accentual phrasing of an utterance. The condition most conducive to ambiguity is a sequence of two accented words in which the second phrase has initial accent, so that the boundary tone cannot be associated with the first syllable of the second phrase (i.e. it is the shorter, higher “weak” L%), and the first phrase is either short or has a particularly late accent, so that there is not much room for the L% to be fully realised on the segments just before the phrase boundary, as shown schematically in Fig. 7b as compared to Fig. 7a. If the prominence relationship between the two words is such that the second phrase has a very reduced pitch range relative to the first, there can be complete phonetic undershoot of the delimitative rise from the L% to the phrasal H-. The phrasal H- might then be realised as nothing more than a vague plateau in the pitch contour, as in Fig. 7c. In cases of extreme subordination of the second phrase to the first, it might not be possible to distinguish such a sequence of two accentual phrases from the case where the second word has been dephrased (see the discussion in Pierrehumbert & Beckman, 1988, pp. 189–190).

The auxiliary verb in the “try X-ing” construction discussed by Poser (1984) is etymologically identical to the verb /mi’ru/ “to see” and thus is underlyingly accented on its first syllable. The main verb that Poser used in illustrating his claim about the phrasing of the auxiliary verb construction is also accented on its first syllable—/yo’mu/ “to read”. That is, he based his claim about the phrasing of the construction on utterances of /yo’Nde-mi’ru/ “try-reading”, as contrasted with the verb–verb sequence /yo’Nde mi’ru/
FIG. 7. Schematic pitch ranges, tone targets and “pitch” contours in: (a) *soroetarisite yomu* “arrange and such, and read”, where the accent in the first phrase is six moras from the end of the phrase; (b) *yonde miru* “read and see”, where it is on the penultimate mora; and (c) *yonde miru* “read and see”, in a context where *miru* “see” is old information and the prominence on its phrase is reduced relative to that on *yonde* “read and …”.

“read and see”. Since the auxiliary verb is by definition subordinated to the main verb, this example provides exactly the condition most conducive to the ambiguity described above. Thus an equally plausible alternative account is that the two syntactic structures do not differ categorically in the accentual phrasing, but rather simply that the pitch range on the auxiliary verb is considerably reduced relative to that of the preceding main verb so that the delimitative rise marking the phrase boundary is phonetically obscured. Kubozono (1993, p. 134) observed many instances of two separate accentual falls in utterances of the similar sequence /no'Nde-mi'ru/ “try drinking”, a result that is incompatible with Poser’s dephrasing account given the “first accent wins” rule. Maekawa (1994b) made a more controlled test of the alternative account by examining the slope of the fundamental frequency fall over the boundary region in /no'Nde-mi'ru/ “try drinking” and comparing it to /no'Nde-iru/ “is drinking”, a syntactically similar construction in which the auxiliary verb is not accented. He found consistent differences in slope which suggested that the second accent in /no'Nde-mi'ru/ is not deleted, as it would be if there were real dephrasing here.

Kubozono’s (1993) observations show that even the basic *bunsetu* consisting of noun and following postposition need not be dephrased. His corpus of utterances also included readings of phrases containing accented postpositions, such as /ro'oma-yo'ri/ “from Rome” and /ro'NdoN-ma'de/ “as far as London”. Although the majority of these utterances showed single smooth pitch falls starting at the accent on the noun, he did note a small minority of cases in which there were two discernible accentual falls. Sagisaka and Sato (1983) also observed “secondary accents” in such phrases, in a corpus of utterances that was large enough that they could relate the occurrence of the second discernible pitch fall to the distance between the two accented syllables. They found that the proportion of observations of two falls increased with the separation between the postposition’s accent
and the noun's accent, a result that suggests that the postposition is not necessarily dephrased but merely usually so subordinated to the noun's pitch range that the delimitative rise cannot be realised. In other words, the standard claims about accentual phrasing as a more or less direct grammatically conditioned reflex of syntactic structure seem less plausible than an indirect account of pitch range subordination mediated by the relative degrees of disjuncture between the two words in the discourse structure. [Further evidence for such an indirect, discourse-mediated relationship between syntactic grouping and prosodic phrasing comes from observations of spontaneous speech, where even a monosyllabic unaccented postposition such as the subject or object case marker can be realised with a full delimitative rise, to put a kind of metalinguistic focus on the role of the NP in the discourse segment (see, e.g. Muranaka & Hara, 1994).]

**AMBIGUITIES OF ACCENT PLACEMENT**

Although English does not have a level of phrasing comparable to the accentual phrase in Japanese, there are comparable ambiguities to the one described above, involving pitch–range relationships and accent placement. As in Japanese, English has a phonologically governed phenomenon of downstep, which is triggered not by every accent, but by a particular subset of accent types. In particular, the bitonal accents tend to trigger downstep (as in the sequence of L+H* accents on *Indianapolis* in Fig. 4), and there is a type of accent, notated !H* in the ToBI transcription system, which contrasts with the plain H* in being downstepped relative to any preceding H tone within the same intermediate phrase. (The contrast between H* and !H* is illustrated in Fig. 8 in the context of preceding H*. The symbol “!” is a common notation for downstep in African tone languages, and the ToBI system uses it to explicitly mark all downsteps, whether the downstep is predictably triggered by a preceding bitonal accent, as for the L+!H* on the second foot in *Indianapolis* in Fig. 4, or is the contrastively downstepped !H* pitch accent, as in Fig. 8.) Downstepped accents are prone to ambiguities about deaccenting, analogous to the ambiguity over whether there has been dephrasing in Japanese, because of the multiplicative nature of downstep. That is, each successive downstep is a constant proportional reduction of the pitch range, so that the absolute reduction in pitch levels for the H tone targets becomes geometrically smaller the longer the sequence. Thus, when there is a sequence of several downsteps within an intermediate phrase, the absolute phonetic difference between two phonologically different tone levels can become extremely reduced, so that it can become very unclear in the pitch pattern whether there is a pitch accent on a word late in the phrase.

In theory, this type of ambiguity is possible for any type of downstepping sequence. For example, in a succession of many L+!H* accents, it should
become very difficult to parse the rise from the L target to the H tone in the last accent or two, particularly if the pitch range specification for the intermediate phrase is already very reduced to indicate some kind of discourse-related subordination to a neighbouring phrase. However, because of the meaning of the L+H* accent (see Pierrehumbert & Hirschberg, 1990; Ward & Hirschberg, 1985), it is unusual to find more than two or three such accents in a row, at least in American English. Also, the low target of the L+H* accent is considerably lower than the L% boundary tone in the delimitative rise for the Japanese accentual phrase, and so a succession of these accents typically presents a more or less salient alternation of L and H targets to be parsed as a succession of rising accents.

Ambiguity in determining whether there is an accent at any given location in a downstepped pitch range, therefore, is far more common in a downstepping sequence involving the single-tone !H* pitch accent. A succession of !H* pitch accents presents no very salient alternation of tone type. Instead, the existence and location of the accents must be interpreted from the much smaller difference in pitch level between one !H* target and another !H* at the next level of downstep. Figure 9 gives an example of such an ambiguity, an utterance which was transcribed in two different ways in two independent ToBI transcriptions. The transcribers agreed that in the phrase *show me the round-trip flights*, there is a L+H* accent on round and at least one more accent, a !H*, somewhere in the following material to account for the only gradual fall from the L+H* peak. (If round were nuclear accented, the F0 peak for its L+H* would be followed by either a flat plateau for a H- phrase accent or a steep fall to a L- phrase.) However, they disagreed on the placement of the following accent. One transcription put the nuclear !H* on trip and the other put it on flights. Since ToBI transcriptions are made while listening as often as necessary to the digitised utterance as well as looking at the fundamental frequency contour, inter-transcriber disagreement of this sort can be taken as an indication of ambiguity in the perception of accentual prominence. There are two independent transcriptions of a large set of utterances from the Boston University FM Radio News corpus, made in an experiment to gauge inter-transcriber reliability before the transcriptions were used in a stochastically trained automatic extraction of parameters for machine synthesis of intonation patterns (Ross, 1995). Even for this corpus of carefully read speech, of 131 cases where at least one of the two transcriptions has !H*, there are 24 examples of this sort of disagreement between transcribing !H* and transcribing no accent on the word (as compared to 16 examples of disagreement between !H* and H*). There are even more examples of this sort of confusion between !H* and no accent in the two independent transcriptions that are being made now of the spontaneous speech corpus from which Fig. 9 was taken (Ostendorf, pers.
FIG. 8. Fundamental frequency contours for two renditions of the sentence *Her name was Maria*, with pre-nuclear accent on *name* and nuclear accent on *Maria*. In both cases, the pre-nuclear accent is H*, but the nuclear accent in rendition (a) is H*, whereas that in rendition (b) is the downstepped !H*.
FIG. 9. Fundamental frequency contour for a spontaneous utterance of the sentence *Show me the round-trip flights between [Newark and Cleveland]*. Utterance from the Boston University ATIS data, used with the permission of Mari Ostendorf, who also provided the information about the alternative ToBI transcriptions.
comm.), and it was one of the more common disagreements among the many more transcriber-pairs in Pitrelli and co-workers’ (1994) study.

This ambiguity in parsing the placement of the later accents in a H* (!H*...) !H* L- sequence makes for an important difference in the salience of the stress pattern between this tune and the archetypal American English citation-form intonation, which consists of a sequence of one or more H* accents followed by a L- phrase accent and L% boundary tone in a “flat-hat” pattern. In this flat-hat pattern, there may be a small rise from a neutral pitch at the onset of phonation to the high pitch target of the first H*, giving some phonetic salience to the first accent. However, if the initial accent is on the first syllable in the phrase, and particularly if the syllable begins with a pitch-raising voiceless obstruent, there may not be even the phonetic rise to the initial H* target. Moreover, for stressed syllables between this initial rise and the nuclear pitch accent, it will be difficult to differentiate between a H* target specification and a phonologically unspecified point in the phonetic interpolation between the two H* targets, so that accentuation in this part before the nuclear syllable is inherently ambiguous. Thus, the most consistent and by far the most prominent event in the contour will be the large pitch fall from the nuclear pitch accent to the L-phrase accent. In !H* downstepping sequences, by contrast, the fall from the last !H* level to a L- phrase accent will be realised in the part of the phrase that has the most reduced pitch range. Phonetically, the most consistent and salient event will instead be the rise in pitch to the initial H* accent. A phoneme-monitoring experiment by Ayers (1995) suggests that other ancillary phonetic characteristics reinforcing the stress level defined by the pitch accent are also less salient in exactly the places where the pitch pattern is prone to ambiguity. Reaction time to detect a target phoneme in a syllable early in an utterance is slower when that syllable is a pre-nuclear accent than when that syllable is nuclear accented for early focus. This is true of the pre-nuclear H* accent both in the H* H* L- L% flat-hat pattern and in the H* !H* L- L% downstepping pattern. Reaction time to detect a target phoneme in a syllable late in an utterance, on the other hand, is slower when the syllable has the !H* nuclear accent of the downstepping pattern than when it has the H* nuclear accent of the flat-hat tune (albeit not so slow as when that syllable is deaccented after an early focus.)

As noted on pp. 31–32, the constraint that defines the qualitative prominence contrast between nuclear accented syllables and pre-nuclear accented syllables is evident in two types of behaviour: the lexical specifications for contrasting accent placements in citation form productions of multi-pedal words such as legislature versus legislation, and the syntactic constraints that have been observed on the focus interpretations of different accent placements in multi-word utterances. At least since Ladd (1980), it has been an axiom of standard generative grammar accounts of focus
domains and their relationship to intonation, that focus involves accent placement, independent of the specification of accent type. For example, Selkirk (1984) borrowed Pierrehumbert's sets of utterances of the flat-hat pattern on *Legumes are a good source of vitamins* and of the American English yes–no question intonation on *Are legumes a good source of vitamins?*, which show that the patterns of broad and narrow focus domains are exactly parallel between the two tunes. (The American English yes–no question intonation is the mirror image of the hat pattern, consisting of one or more L* accents followed by H- H%.) Placing the nuclear accent on *vitamins* allows many focus interpretations. It is ambiguous among broad focus on the utterance as a whole, and among narrower focus on {good source of vitamins}$_F$, on {source of vitamins}$_F$, or on just {vitamins}$_F$. By contrast, placing the nuclear accent earlier in the utterance allows only the narrowest focus interpretation; for example, putting the nuclear pitch accent on *good* can only be interpreted as narrow focus on {good}$_F$. This pattern of possible focus interpretations holds in both sets of utterances, whether the nuclear accent is the H* of the flat-hat citation pattern or the L* of the yes–no question tune. Since the independence of focus interpretation and accent type is thus assumed, most accounts of focus interpretation ignore accent type entirely, and the examples used to support the discussion seem to involve actual readings or imagined utterances of only the flat-hat pattern with varying nuclear accent placement.

However, given the different patterns of ambiguity of accent placement between the flat-hat pattern and the downstepping narrative tune, which differ only in whether the accents are H* or !H*, it seems important to re-examine the assumption that focus domain is determined solely by accent placement without interaction from choice of accent type. For example, we might ask whether the ambiguous scope of focus for utterances with late nuclear accent (as in the utterance of *Legumes are a good source of vitamins* with nuclear accent on *vitamins*) is in some sense an artifact of the inherent ambiguities in parsing pre-nuclear accent placement in the flat-hat tune. One indication that this might be so is that the tune typically contrasted to the flat-hat as evidence for the irrelevance of accent type is the American yes–no question tune. In this tune, the most salient accent placement again will be the last, at the place in the contour where pitch begins to rise sharply from the L* nuclear accent to the H- phrase accent. The pre-nuclear L* tones involve no salient alternation of pitch levels. Thus, there is an exactly analogous inherent ambiguity in parsing pre-nuclear L* accents in the American yes–no question intonation.

Another indication that we should question the current generative understanding of focus and accent placement is the uncertain status of pre-nuclear accents in most accounts. For example, although Ladd’s (1980) first chapter presents British analyses of intonation contours in terms of a
pre-nuclear “head” and a “nuclear tone” (which are the analogues in these systems of analysis to what Pierrehumbert analyses as, respectively, the first pre-nuclear accent and the sequence of tones from the nuclear accent to the end of the phrase), if we look at Ladd's later examples in the chapter on broad versus narrow focus, it is difficult to know how the pre-nuclear accents of the “head” should be treated vis-à-vis the focus domain. In Selkirk’s rule-based “accent-first” account, all accent placement is said to be determined originally by focus, and not just the nuclear accent placement (see Selkirk, 1984, pp. 268 ff.). However, again, if we look at her example sentences, it is difficult to know how to treat pre-nuclear accents. For example, her discussion of how a broad VP-focus is derived in pairs such as She sent a BOOK to MARY versus She sent the book to MARY (ibid., pp. 210 ff.), includes the claim that these differ intonationally—that there is a pre-nuclear accent on book only in the first sentence, where the direct object is also a focus because it is new information. The production indicated for the second sentence would seriously violate the second constraint on accent placement discussed above, and in reviewing these examples with native speakers of English, I have never found a speaker who can produce such a long pre-nuclear stretch without an accent either on sent or on book. My guess is that the intonation patterns that Selkirk had in mind for these examples actually differ in phrasing (or possibly in accent type) and not in the presence or absence of an earlier accent. That is, the lack of capitalisation on book in She sent the book to MARY might indicate the lesser salience of a pre-nuclear accent in the flat-hat pattern, whereas the accent indicated by the capitalisation on book in She sent a BOOK to MARY could be the more salient nuclear accent of a H* L- when there is an intermediate phrase break between it and Mary (or perhaps the nuclear accent on Mary is L+H*, providing a salient fall in pitch from a pre-nuclear H* on book).

A question to ask, then, is whether the patterns of focus interpretation would be the same in the narrative tune involving !H* accents, where the placement of the last accent itself can be ambiguous. There is some suggestion in the literature on intonational meaning that focus interpretations do in fact differ between the two intonations. In discussing the meaning of the !H* pitch accent, Pierrehumbert and Hirschberg (1990) state that sometimes the accent is functionally equivalent to deaccenting. They give the following overheard example:

6. We know you have great credentials. We’re looking for someone with just such credentials.
   H* !H* !H* L- L%

Pierrehumbert and Hirschberg note that the second sentence would have been just as felicitous in this context if the speaker had instead placed the
nuclear accent on *just*, deaccenting all following words, including the second occurrence of *credentials*. Substituting the flat-hat pattern without changing the accent placements, by contrast, would be quite anomalous. The listener might interpret the focus pattern indicated by the nuclear H* accent on *credentials* as implying that these are a different set of credentials. This and related examples of utterances with !H* accents constitute an important class of exceptions to the phonological generalisation that the last accent is placed on the most stressed syllable in the intermediate phrase. The phoneme-monitoring experiment described above lends further support to the idea. [See Ayers (1996) for a review of more examples from English. Terken (1995) describes an experiment eliciting prominence judgements for early and late content words in phrases extracted from spontaneous monologues, the results of which suggest that the facts for Dutch are similar.]

**AMBIGUITIES OF INTERMEDIATE PHRASING**

In English, the choice of particular tone sequences can create ambiguities in the phrasal parse as well as in the stress parse. For example, when the accented syllables are close together, it can be difficult to differentiate an intermediate phrase with two accents where the nuclear accent is L+H* from a sequence of two intermediate phrases with an intervening L- phrase accent. Figure 10 shows schematic “pitch” contours for a pair of such tunes, and a list of three other tune pairs that fit this description. Figure 11 gives an actual example from the Boston University FM Radio News corpus, which was transcribed in these two different ways in two different ToBI transcriptions. That is, one transcription attributed the low portion between the accent peaks to the L tone of the nuclear L+H*, whereas the other transcribed an intermediate phrase break marked with a L- phrase accent.

If nuclear accented syllables other than !H* are structurally more prominent than pre-nuclear ones, as described on pp. 31–35, then the two different transcriptions of the utterance in Fig. 11 correspond to two different stress patterns, and two different focus interpretations. The single nuclear accent (as in the single-phrase sequences in the tunes in the right-hand column in Fig. 10) is one focus domain, whereas the two nuclear accents (as in the two-phrase sequences in the tunes listed in the left-hand column) would indicate double foci. Thus, as hinted above in discussing Selkirk’s examples, this potential ambiguity in phrasing for these tunes complicates our understanding of the relationship between intonation and focus.

It could be that in ordinary circumstances, this ambiguity does not matter very much. For example, when the progress of a conversation hinges on the interpretation, perhaps speakers produce more deliberate pronunciations
which differentiate the two interpretations by other phonetic cues, such as a more pronounced final lengthening for the first intermediate phrase in the double-foci case. Beckman and Pierrehumbert (1986) discuss an analogous ambiguity of phrasing involving the L\(^*\) accent, which figured in our analysis of vocative tags and tag questions. The relevant examples were sentence pairs such as Mary will Win Manny, produced with a H\(^*\) L-H\(^%\) tune and a narrow-focus nuclear accent on win, versus Mary will win, Manny, which we analysed as a two-phrase sequence with the pattern H\(^*\) L- L\(^*\) L- H\(^%\). In the deliberately disambiguated lab-speech examples of this pair that we used, there was a considerably longer duration for the phrase-final nuclear accented syllable in the main clause before the tag, so that the fall from H\(^*\) to L- could occur before the following phrase boundary. (This is an example of the kind of accommodation from text to tune mentioned above in the definition of alignment.) However, this phrase-final lengthening was less pronounced in Mary will marry, Manny, where the fall for H\(^*\) to L- could be spread out over the following unstressed syllable. Beckman and Edwards (1990) similarly found that final lengthening is considerably less reliable when the target final syllable is not also nuclear accented. It seems likely that such durational cues will be even less reliable in speech produced outside the laboratory, in contexts where redundant textual cues might disambiguate the speaker's intentions, so that lengthening and pause can be used orthogonally to the intonational phrasing, for other rhetorical purposes. Woodbury (1993) describes such orthogonal rhetorical uses of pause in Yupik Eskimo narratives. As far as I know, there has been no systematic study of the phenomenon in English, but examples such as the utterance in Fig. 12 show that pauses and pause-like lengthenings can occur in the middle
FIG. 11. Fundamental frequency contour for the sequence...says Ballaga is sincerely committed... extracted from the sentence Massachusetts Water Resources Authority director, Paul Levy, says Ballaga is sincerely committed to pro-environment goals. There are L+H* accents on the underlined syllables, and a L- phrase accent marking an intermediate phrase boundary after Ballaga and possibly also after sincerely. The utterance is from the Boston University FM Radio News corpus, and used with the permission of Mari Ostendorf.
of tonally demarcated phrases. Here the lengthening of the [z] in the preceding *does* and the pause before the [s] of the following *stand* set off the phrase *YN* for special attention (perhaps in order to give the listener more processing time) without destroying the intonational coherence and prominence relationships among *What* (with its clearly pre-nuclear *L+H*\* accent), *Y* (with its quasi-nuclear *L+!H*\* accent) and *stand* (with its inherently less prominent downstepped *!H*\* accent). Such cases of deliberate pausing in the middle of a well-formed tune occurred often enough in the communally transcribed utterances used in developing the ToBI consensus transcription system that a separate “break index” value was prescribed for them (see Beckman & Hirschberg, 1994).

The examples cited so far have been ambiguities between no phrase boundary and an intermediate phrase boundary. There is also a type of ambiguity between an intermediate phrase and an intonational phrase boundary which stems from the contrastive features of the *L%* boundary tone. This tone is differentiated qualitatively from *H%* primarily by an absence of specification for pitch raising. For example, *H*\* *L- L%* (the flat-hat citation form pattern) differs from *H*\* *L- H%* (the “fall–rise” at the end of a flat-hat pattern with continuation rise) in that the nuclear pitch fall is not reversed at the end. Similarly, the *H*\* *H- L%* tune of the “stylised plateau” pattern discussed by Ladd (1980) differs from the *H*\* *H- H%* “high–rise” pattern discussed by Cruttenden (1994) and Hirschberg and Ward (1995) in that the pitch remains level rather than rising again at the end of the intonational phrase. At an utterance–medial boundary, then, any tune without a rise from the pitch level of the phrase accent is tonally ambiguous. It could be a full intonational phrase break with a *L%* boundary tone, or it could be a mere intermediate phrase break with just the *L-* or *H-* phrase accent.

There is some indirect experimental evidence that listeners interpret the intended degree of phrasal disjuncture in carefully read speech by relying on other (non-categorical) cues, such as the presence and degree of final lowering (Liberman & Pierrehumbert, 1984) in the case of *L-* versus *L- L%*, and the amount of final lengthening in both cases. For example, Wightman and Ostendorf (1994) showed that stochastic recognition models can be trained to recognise hand-labelled break index values for these two levels of phrasing in carefully read speech on the basis of normalised segment durations and pauses. On the other hand, examples such as the utterance in Fig. 12 make it unclear whether such results can generalise to the interpretation of phrase breaks in more complex discourses, where other discourse purposes can induce the speaker to place pauses and lengthenings orthogonally to the groupings indicated by the intonational phrasing. Until we know more about the dynamics of these temporal effects, we cannot guess how much listeners will be able to rely on final lengthening to
FIG. 12. Fundamental frequency contour for an utterance of the question *What does Y N stand for?*, from the Boston University ATIS data. The lengthening of the [z] before the Y and the pause between the N and the *stand*, set off the Y N for the listener's attention, at the same time that the downstep from the accent on *What* to that on *Y* act to make the tune cohere as a single intonational phrase.
differentiate an intended intermediate phrase break with L- from an intended full intonational phrase break with L-L% in speech produced outside of the laboratory.

To relate this ambiguity to the other sense of the “prosody and parsing” theme, we first must ask how important it is to distinguish these two levels of tonally marked phrases in parsing other grammatical structures. The answer depends to a large extent on the analysis of the meanings of the phrase accents and boundary tones. Pierrehumbert and Hirschberg (1990) characterise the meanings as follows. The H- phrase accent is “forward-looking”. It conveys that the interpretation of the current intermediate phrase should be explicitly related to the one that follows it, and thus contrasts with the neutral disjuncture of the L- phrase accent. The H% boundary tone, similarly, contrasts with the L% boundary tone in conveying that the current intonational phrase should be interpreted with respect to something in the surroundings, either to the preceding intonational phrase, or to the following phrase, or to the anticipated response by the other conversational participant(s).

The meaning of the H- phrase accent can be understood more concretely by analysing its specific interpretation in some common intonation patterns for the constructions in (7):

7a. Do you want coffee, or tea?
(H*) H* H- H* L- L%

b. Give him an inch, he'll take a mile.
(H*) H* H- H+!H* L- L%

In (7a), the medial H- indicates that the interpretation of the first phrase should wait for the list to be closed with the L- L% sequence on the following phrase. In (7b), it indicates that the first phrase should be interpreted as a conditional to the following phrase. The interpretation of the H- phrase as a conditional is reinforced by the meaning of the H+!H* in the second phrase, an accent which Pierrehumbert and Hirschberg (1990) identify as meaning that the listener should infer an explicit connection between the proposition and some mutually believed background knowledge. The specific interpretation of the H- in this tune as a conditional is a syntactic one. As (7a) shows, however, H- can be used felicitously also in contexts where this specifically syntactic interpretation is not appropriate. In other words, in Pierrehumbert and Hirschberg's compositional semantics (as in Ladd’s, 1980, approach, or in Gussenhoven’s, 1983), the intonation pattern has a pragmatic meaning of signalling the speaker’s intention for the interpretation of the text vis-à-vis the discourse context. Any syntactic interpretation follows only indirectly, as a conversational implicature.
derived from inferring the consequences of the pragmatic meaning for a particular text in its larger intonational and discourse context.

Such an account of an indirect relationship between intonational meaning and syntax is supported also by the many different specific interpretations of the H% boundary tone, which occurs in the continuation rise, in the yes–no question contour and in the high-rise contour (see the discussion of these in Hirschberg & Ward, 1995), as well as in the utterances shown in Fig. 2. The utterances in Fig. 2 exemplify a marked double-focus pattern discussed by Jackendoff (1972) and Ladd (1980). Each of the two utterances is broken into two phrases, so that there can be nuclear accent and focus on both names. However, the relationship between the two foci differs. In the utterance with the medial H% in Fig. 2a, the first phrase is interpreted as a “background” focus (or “theme”) to the “answer” focus (or “rheme”) in the second phrase. This might be an appropriate response to What about Anna? Who did she come with? In the utterance in Fig. 2b, by contrast, it is the second phrase that has the H% triggering the theme interpretation, so the first phrase (the one with the steep H* L- L% fall) is interpreted as the rheme. This would be an appropriate response to What about Manny? Who came with him? The direction of the interpretation of the H% differs in the two cases, but the specific interpretation with respect to the other phrase is the same. Whether the boundary tone makes the phrase forward-looking or backward-looking, the connection is one of subordinated “background” to the “answer” focus. Liberman and Pierrehumbert (1984) show that there are also other phonetic differences between the two orderings. In particular, the relationships between the peak fundamental frequencies for the two accents differ. The first pitch peak is higher relative to the second in the “answer–background” sequence, reinforcing the backward-looking interpretation of the H% as a relationship of subordination to the preceding phrase. Since Pierrehumbert’s intonational model ascribes the peak height relationship to a separate quantitative specification of backdrop pitch range for each phrase, this reinforcement of the backward-looking interpretation of the final H% is independent of the pitch accent types. It is available whether the nuclear accents in the two phrases are both H* or both L+H* or even both !H* (as in a later experiment by Liberman and Pierrehumbert reported in Pierrehumbert & Beckman, 1988, p. 181). The direction of the H% of the theme is independent also of the boundary type for the rheme, so long as there is a boundary; that is, the utterance-final H% can indicate subordinated backward-looking “background” information even when there is only an intermediate phrase break between it and the preceding “answer” focus, as in the production in Fig. 5b. It is only when there is no phrase break, making the accent on Anna a pre-nuclear (non-focal) accent, as in Fig. 5a, that the answer–background double-focus interpretation becomes problematic.
In Pierrehumbert and Hirschberg's compositional semantics for intonational morphemes, then, the H- phrase accent and H% boundary tone differ in scope (the H- determines the interpretation of the current intermediate phrase relative to the following intermediate phrase, whereas the H% determines the interpretation of the current intonational phrase), and also in the possible directions of interpretation (the high phrase accent seems to be only forward-looking, whereas the high boundary tone relates the interpretation of the current intonational phrase either to following or to preceding material). The meanings of the L- phrase accent and L% boundary tone also contrast with each other in scope, but otherwise they are more similar. They contrast with the meanings of their H counterparts by being essentially neutral (e.g. the L- is “not explicitly forward-looking”). The phonetically subtle tonal difference between L- and L- L% thus is matched by a correspondingly subtle difference in interpretation.

Steedman (1991a, 1991b) takes a rather different approach to the syntactic interpretation of intonation patterns, an approach which he works out in some detail for Jackendoff’s double-focus constructions. His analysis involves ascribing a rather different set of meanings for the phrase accents and boundary tones, in a grammar that derives the answer–background and background–answer interpretations for the double-focus construction without reference to such extra-tonal facts, such as F0 peak relationships or the larger intonational and discourse context. However, the analysis hinges on there being only an intermediate phrase boundary in the middle of the answer–background sequence, as in Fig. 5b. [He points this out in a footnote in Steedman (1991b), and the following discussion will refer to that paper rather than to the earlier article in *Language*, where there was no difference between the two.]

Steedman's analysis of the phrase accent and boundary tone meanings is couched in the formalisms of Combinatory Categorial Grammar (Steedman, 1987). In this theory, each element in the text is associated with a syntactic category type, defined as a function which takes a specific type of argument from the preceding or following context and produces a specific type of element as its result. For example, a transitive verb is associated with the category VP/NP, meaning that it combines with the type noun phrase on its right to produce something of type verb phrase. The types of derived constituents are exactly the same. For example, the VP constituent produced by combining a transitive verb with the following NP in applying the VP/NP function, is itself a function, of type SNP. That is, it combines with a noun phrase on its left to produce a sentence. Thus the syntactic type of a constituent is identical to its semantic function in the sentence, and the syntactic parsing of an utterance can thus be viewed as simultaneously providing a compositional semantic interpretation.
Besides the basic rule of functional application, there are only a few more rule types. A coordination rule combines constituents of like type to make a combined constituent of the same type, and a type-raising rule changes an argument into a function over a function with that type of argument. For example, “subject type raising” changes a noun phrase into type S/VP (i.e. a function looking for a verb phrase on its right to make a sentence). Steedman (1991a, 1991b) describes how the rules of combinatory categorial grammars—particularly the type-raising rules—can produce many alternative derivational paths for some sentences, all having the same semantic interpretations. He proposes that the choice of derivational path will be constrained by being identified with different intonational phrasings. Moreover, he proposes to treat intonational elements in the same way as textual elements, identifying syntactic types and combining them by functional application and other syntactic rules, to account for the different discourse functions of a text when it has different tunes. By incorporating intonational categories directly into the syntactic/semantic parse in this way, Steedman proposes to collapse the “informational structure” indicated by intonational boundaries together with the surface syntactic structure, eliminating a large source of indeterminacy for the listener (and presumably also for the language learner; see Steedman, 1996).

Steedman's analysis of the double-focus construction illustrates this approach. In most other syntactic formalisms, the syntactic structure of the double-foci answer–background renditions in Figs 2b and 5b is the same as that of the background–answer rendition in Fig. 2a and of the single-focus rendition in Fig. 5a. Steedman, by contrast, builds a fragment of a combinatory categorial grammar for pitch accents and phrasal tones that gives the answer–background double-focus interpretation a different syntactic parse from the background–answer interpretation. In order to understand this grammar, it is crucial to know that Steedman identifies the background–answer tune as specifically L+H* L- H% H* L- L%, and the answer–background tune as specifically H* L- L+H* L- H%. That is, unlike Pierrehumbert and Hirschberg’s (1990) analysis, Steedman’s analysis does depend on the pitch accent types in the two phrases, and on the level of intonational disjuncture in the answer–background case. His fragment of a combinatory categorial grammar for intonation, then, identifies the following syntactic types:

8. L+H* : = Theme/L- H%
   H* : = (Utterance/Theme)/L-
   : = (Utterance/\Theme)/L- L%

That is, the L+H* accent on the background focus is an element looking to combine with a L- H% sequence at the intonational phrase boundary to its
right to produce an intonational constituent of type theme. The H\* accent on
the answer focus can be either one of two types, which produce two different
types of rheme, depending on the argument type. If the argument is the L- at
a following intermediate phrase boundary, applying the function combines
the H\* with the L- to produce a “forward” category rheme—that is, a rheme
that combines with a following theme to produce an utterance. This is the
“answer–background” case. If the argument to the right of the H\* is the
L- L% at a full intonational phrase boundary, on the other hand, the
functional application produces a “backward” category rheme—that is, a
rheme that combines with a preceding theme to produce an utterance. Thus,
in Steedman’s analysis, the directionality of the interpretation of the
background focus is hardwired in the grammar, and depends on whether the
answer focus is in an intermediate phrase marked by L- or in a full
intonational phrase marked by L- L%. This means that Steedman’s analysis
has the problem that it will not generate the answer–background rendition
with the full intonational phrase boundary medially, as in Fig. 2b.

Ladd (pers. comm.) has suggested that the problem here is not with
Steedman’s grammar of intonational meaning, but with the prosodic theory.
Since there is no clear qualitative pitch contrast between H\* L- and
H\* L- L%, we cannot know for sure that the utterance–medial boundary in
Fig. 2b is not the same kind of boundary as the utterance–medial boundary
in Fig. 5b, and this indeterminacy should be incorporated into the structural
representation. Specifically, Ladd proposes to replace the strictly layered
tree structure of intermediate phrases underneath intonational phrases by a
recursive phrasal parse based on phrasal pitch range relationships, as in
Ladd (1986). The motivation for this proposal is similar to the reasons for
choosing the grid rather than the tree in (2). That is, choosing the grid rather
than the tree to represent the prominence of the head of the foot allows a
uniform representation of the relative structural prominences of syllables at
the higher levels shown in (4) and (5), despite the indeterminacy of any
prosodic edges between the foot and the intermediate phrase. However, the
two solutions to the problem of indeterminate edges differ radically in the
claims about the paradigmatic correlates of prominence. Like the prosodic
trees proposed by Nespor and Vogel (1986), the grids in (4) and (5) are
level-ordered and bounded in height. The levels of beats in the grid are
limited to ones that can be clearly labelled by some qualitative feature of
phonetic prominence, and the implicational relationships among the labels
define a strict order from lowest to highest grid levels. Because it allows
recursive embedding of phrases, Ladd’s proposal precludes such a
qualitative labelling, and thus is instead like the unbounded, unlabelled grid
proposed by Liberman and Prince (1977). This means that adopting Ladd’s
representation would preclude the generalisation that prosodic boundary
strength for these higher-level units is effectively neutralised just when the
presence of the boundary tone (the qualitative label for the intonational phrase) is unclear because there is no salient phonetic alternation with the phrase accent, so that the intonational phrasing in the double-focus construction is not ambiguous for the theme phrase where the boundary tone is H%.

Moreover, recursive phrasing will only compound the problem of indeterminacy for any grammar, as in (8), that uses the boundary strength after the background focus to set the type for the rHEME phrase as a “backward” or “forward” category. For the grammar in (8) to work, it must have reasonably salient boundary strengths. The distinction between L- and L- L% for the different rHEME types in (8), then, must be redefined in terms of other cues besides the qualitative tonal markers, cues such as the distribution of turn-medial pauses and final lengthening. Moreover, these other cues must be salient and reliable enough to support acquisition of the syntactic contrast. Thus, where Pierrehumbert and Hirschberg’s analysis can accommodate both the renditions in Fig. 2b and in Fig. 5b, and could stand whether or not there is evidence that other phonetic cues reliably support the subtle tonal contrast between them, the arguments for Steedman’s analysis are vitiated if such evidence is not forthcoming.

SUMMARY

This paper has reviewed an account of prosodic categories related to the intonation pattern in English and Japanese, and has described several common prosodic ambiguities. In describing these categories and in presenting these examples, I have tried to illustrate three main themes. First, prosody is not another word for “suprasegmentals”; rather, it is a complex grammatical structure that must be parsed in its own right. Second, certain contextual factors—ranging from the choice of a particular tonal sequence to the separation of accented syllables and specification of pitch range relationships between adjacent phrases—are conducive to common misparsings of the intended prosodic structure, or make for common inherent ambiguities in the parse. Finally, these ambiguities have profound implications for certain claims and assumptions in the literature on the relationship between prosody and syntax. In short, in examining the issue of prosody and parsing, it is imperative to be precise about the prosodic structures assumed, and it may be necessary to explore issues in the phonological structures in tandem with syntactic or semantic ones.

REFERENCES

Ayers, G.M. (1996). *Nuclear accent types and prominence*. Doctoral dissertation, Ohio State University, Columbus, OH.


Beckman, M.E., & Ayers, G.M. (1994). *Guidelines for ToBI labeling guide, ver. 2.0*. Manuscript, Ohio State University, Columbus, OH. (Tutorial and accompanying example utterances are available by contacting tobi@ling.ohio-state.edu.)


Beckman, M.E., & Hirschberg, J. (1994). *The ToBI annotation conventions*. Manuscript, Ohio State University, Columbus, OH. (Provided with the *Guidelines for ToBI labeling guide*, or available separately by contacting tobi@ling.ohio-state.edu.)


Trim, J.L.M. (1959). 'meidZer |@n 'main@ | "toungru:ps | in "ÎgliS Ù. Le maître phonétique.112, 26–29.


Venditti, J.J. (1994). Japanese ToBI labeling guidelines. Manuscript, Ohio State University, Columbus, OH. (On-line notes available from venditti@ling.ohio-state.edu.)


