Chapter 2
The original ToBI system and the evolution of the ToBI framework

Mary E. Beckman, Julia Hirschberg, and Stefanie Shattuck-Hufnagel

2.1 Introduction

The term **ToBI** has come to be used in two different ways. Originally, it was the name of an annotation system, developed in the period 1991 to 1994, which was designed for use in labelling intonation and prosody in databases of spoken Mainstream American English (Beckman and Hirschberg 1994). Very quickly, however, it also came to refer to a general framework for the development of prosodic annotation systems in other varieties of English (e.g,, Mayo, Aylett and Ladd 1997 [Glasgow]) and in other languages (e.g., Grice, Reyelt, Benzmüller and Batliner 1996 [German]; Venditti 1997 [Japanese]). In this chapter, we will try to identify the essential properties of a ToBI framework annotation system by describing the development and design of the original ToBI conventions. In this description, we will overview the general phonological theory and the specific theory of Mainstream American English intonation and prosody that we decided to incorporate in the original ToBI tags. We will also state the practical principles that led us to make the decisions that we did.

Before we begin, however, we should explain a practical terminological decision. Although the original ToBI for Mainstream American English (MAE) is the most completely developed of the ToBI-framework systems, and also the system most completely tested by use, the development of ToBI-framework systems for other languages makes this dual usage increasingly awkward. Therefore, we will adopt the following convention for distinguishing the two uses in this chapter. We will reserve the unmodified term 'ToBI' for the developmental framework, and use the prefixed term 'MAE_ToBI' for the original system.

The chapter is organised as follows. Section 2.2 briefly chronicles how the MAE_ToBI system came into being. The purpose of this chronicle is to bring out the practical principles that guided the system's development. Section 2.3 briefly describes the consensus account of English intonation and prosody on which the MAE_ToBI system is based. In such a necessarily abbreviated account, we will not be able to begin to do justice to the nearly 80 years of observation and instrumental research that have made the intonation systems of MAE and its close relative, standard Southern British English (including RP), among the best-understood in the world. Section 2.4 catalogues the different components of a MAE_ToBI transcription and lists the salient rules which constrain the relationships between different components. This section also expands upon the theoretical foundations and practical consequences of adopting the general structure of multiple labelling tiers, and particularly the separation of the labels for tones from the labels for indexing prosodic boundary strength. Section 2.5 then describes some of the extensions of the basic ToBI tiers that have been adopted by some sites. This section also compares our decisions about the number of tiers and about inter-tier constraints with the analogous decisions for some of the other ToBI systems described in this book. Section 2.6 discusses the status of the symbolic labels relative to the continuous phonetic records that are also an obligatory component of the MAE_ToBI transcription. In particular, we describe the status of the Tones tier relative to the fundamental frequency record, and contrast this status to the epistemological claims implicit or explicit in some other transcription frameworks. Section

2.7 then closes by listing several open research questions that we would like to see addressed by MAE_ToBI users and the larger ToBI community.

2.2. The beginnings of MAE_ToBI

MAE_ToBI was developed in a series of four meetings, with delegates from a large number of sites, representing several disciplines. Participants in the workshop included engineers who wanted to train automatic speech recognition systems and build better text-to-speech systems, psychologists who wanted to investigate the relationship between prosody and human language processing, computer scientists who wanted to build better dialogue models and speech generation systems, phoneticians who wanted to test theories about tone association and alignment, and so on. This diversity of aims was also reflected in the diversity of sites for the meetings. The initial Prosodic Transcription Workshop (1-2 August, 1991) was organised by Victor Zue at the MIT Laboratory for Computer Science. The Second Prosodic Transcription Workshop (5-6 April 1992) was organised by Kim Silverman at NYNEX Science and Technology, Inc. The Third Prosodic Transcription Workshop (17-21 June 1993) was organised by Mary Beckman at the Department of Linguistics, Ohio State University. And the Fourth Prosodic Transcription Workshop (5-6 August 1994) was organised by Mari Ostendorf at the Department of Electrical, Computing, and Systems Engineering, Boston University.

Each of these four meetings was termed a 'workshop' and the term was appropriate. At each of them, we arrived prepared to work hard, and we did work hard to agree on what phenomena we wanted the conventions to cover and to decide on how to cover these phenomena. Also, before each workshop, participants prepared a set of exercises, transcribing a common set of utterances, to focus discussion at the meeting. The transcriptions prepared for the first meeting gave us an immediate basis for comparing existing transcription conventions across sites, a comparison which showed us that we had considerable grounds for consensus already in the group. That is, delegates from four of the eleven sites represented at the first workshop happened to have transcribed tone contours using the same autosegmental model of intonation (Pierrehumbert 1980; Beckman and Pierrehumbert 1986), and informal comparison of these transcriptions at the workshop suggested a high degree of agreement among them. Also, delegates from another three sites had transcribed prosodic grouping using the same system of numerical indices to boundary strength (Price, Ostendorf, Shattuck-Hufnagel and Fong 1991), and had compared their transcriptions prior to the meeting in order to be able to present actual numbers on inter-transcriber reliability. The transcription exercises prepared for the second meeting similarly served as the basis for our first group-wide inter-transcriber consistency check (Silverman et al. 1992). Moreover, thanks to Patti Price, we benefited from a meticulous record of the first meeting, and continued to make detailed minutes at subsequent meetings. These records allowed us to circulate a summary of the decisions made at each meeting immediately afterwards. They also served as the basis for successive drafts of the ToBI conventions, which were revised after each of the first three meetings before their final codification in Beckman and Hirschberg (1994). Between the second and the fourth meetings, we also developed a set of on-line training materials to accompany the draft annotation conventions and for use by new transcribers (Beckman and Ayers 1994). These training materials allowed us to recruit naïve transcribers, who had not attended any of the meetings, for the second inter-transcriber consistency check (Pitrelli, Beckman and Hirschberg 1994).

To understand some of our decisions at these workshops, it is useful also to know what prompted us to convene them in the first place. The immediate impetus was the example of the Penn TreeBank project (Marcus, Santorini and Marcinkiewicz 1993). By agreeing on a common

core of syntactic labels, the Natural Language Processing community had been able to develop a large online corpus of syntactically annotated English text. This corpus has been used for exploring aspects of syntactic theory (e.g., Srinivas and Joshi 1999), for testing models of sentence processing (e.g., MacDonald, Pearlmutter and Seidenberg 1994), and for improving the performance of syntactic parsers (e.g., Collins 1999; Charniak 2000). We were inspired by the TreeBank example to try to find an analogous set of consensus tags for intonation and prosody. In the short term, we wanted a similar tool that would allow researchers at different sites to share in the work of developing a large pool of prosodically transcribed online speech databases for a broad range of uses in speech science and technology. In the longer term, we wanted to provide a common vocabulary so that researchers at different sites could interpret each other's data and contribute complementary analyses and extensions of a common core of methods and datasets. There are practical principles to extrapolate from four aspects of this original purpose and of the work that we did to achieve that purpose.

First, the MAE_ToBI system did not spring out of thin air. Rather, it is based on long history of studies of English intonation, stress, and phrasing. Moreover, a sizeable group of researchers in the original MAE_ToBI community were versed enough in various aspects of this history to have provided some basis for consensus even at the first meeting. (Section 2.3 reviews the antecedents of this consensus.) The lesson for the larger speech community is that any new set of ToBI conventions for another language needs to reflect a fairly broad and well-grounded understanding of the intonational and prosodic grammar of the language. Ideally, the conventions will be based on a large and long-established body of research in intonational phonology, dialectology, pragmatics and discourse analysis for the targeted language variety, but at the least, they should be based on rigorous analyses of the intonational phonology. Where established analyses are available for only a subset of the phenomena that users want to label, the development of a ToBI framework system can help formulate the relevant questions for further research, but system development should not run too far ahead of knowledge.

Second, the initial MAE_ToBI group was large enough and diverse enough — approximately 25-30 people attended each workshop — to allow us to pool expertise of various types. The system that eventually emerged was designed to cover only a subset of the prosodic features which we wanted to identify — namely, those that required hand labelling and which we collectively felt we knew enough about to label consistently. Some other features, such as the segmental makeup of each word or the location of its primary lexical stress, could be generated automatically using resources such as online dictionaries; thus, these were not explicitly included as part of the MAE_ToBI system. Still other features, such as phrasal stress patterns or subtle variation in the extent of pitch rises or falls that might be related to discourse-prominence relationships within and across intonational phrases, appeared to require more basic research before they could be labeled consistently; it was hoped that the creation of large labeled corpora could facilitate this study. The lesson we derive from our discussions of what to include in the MAE_ToBI conventions is that ToBI conventions should be efficient. One should not waste transcriber time by asking the transcriber to symbolically mark phenomena that can be extracted from the signal or derived from online resources automatically. (We review several examples, one involving location of main lexical stress, in Section 2.5.)

Third, the MAE_ToBI system was intended for use by an even larger community of end users, with a wide variety of interests and theoretical convictions. It is based on a broad consensus, which involved some amount of compromise on the part of every delegate to the workshops. The lesson to extrapolate for the larger ToBI framework is that a viable ToBI system needs a suitably large and diverse group of users who have agreed to act communally to

develop and adopt the system as a community-wide standard. One corollary principle is that the conventions should be easy enough to teach that their use is not limited to a few experts to do the transcription and to themselves train apprentice labellers. Therefore, there must be a freely available manual for teaching the system to new transcribers, with many recorded examples of transcribed utterances graded from easy to difficult. Another corollary principle is that the conventions need to be used and maintained consistently across transcription sites. Therefore, in the course of developing a ToBI framework system, there must be rigorous tests of inter-transcriber consistency, and there should be agreed-upon centres for maintaining the standard with periodic rechecks and evaluation of any proposed revisions. A third corollary principle is that mechanisms must be provided for customizing transcriptions to particular needs without compromising the common core. (We elaborate on this point in Section 2.5.)

Fourth, the building of transcription conventions was a strikingly iterative process. It involved much discussion and often impassioned argument, interspersed with actual transcribing by all delegates of actual recorded utterances provided by all of the sites. In each iteration, we had a chance to see what kinds of prosodic phenomena were important to others, and how a proposed change in the transcription conventions would affect other transcribers' ability to capture what they heard in the signal and what they wanted to capture for their research. Moreover, the discussion of problems and proposed changes to the conventions was always grounded in the examination of actual speech signals. For example, to argue for making a distinction between two pitch accent categories that had been tagged the same way in the previous draft of the conventions, one at least had to articulate clearly what aspects of the signal supported the proposed distinction and show that the difference was based on more than one pair of examples. Ideally, one also could mimic the tunes on other texts and invoke a recognizable difference in felicitous contexts of use. The lesson for the larger community is that a good ToBI system is not simply a transcription system. It is also a tool for observing the signal and communicating one's observations to the larger community in a common language. A ToBI transcription never replaces a permanent record of the speech signal with a symbolic record; rather, it seeks to integrate a symbolic commentary with the data upon which it is based. A related lesson is that some phenomena of interest, such as phrasal pitch range variation, can be represented by continuous measures that are derived directly from the signal in conjunction with the symbolic labels. We elaborate on these points in Sections 2.4 and 2.6. First, however, we describe the consensus model of MAE intonation and prosody on which the original ToBI system was built.

2.3. The antecedents of MAE_ToBI

The MAE_ToBI system is based on a consensus model that makes five salient claims about intonation and prosodic structure in the language. First, the prosodic pattern for an utterance can be projected onto separate tiers representing conceptually independent structural types. In particular, the intonation contour can be represented linearly by an autosegmental string of **tones**, whereas the metrical hierarchy of intonational phrases and lower-level prosodic groupings should be represented hierarchically, for example by a numerical **break index** value for the perceived degree of disjuncture between any two words. Second, the intonation contour is decomposed into relatively high and relatively low pitch levels: **H** versus **L** tones. These pitch levels are static targets in paradigmatic contrast with each other; 'relatively low' means low relative to the local phrasal **pitch range**, rather than low relative to the nearest pitch peak or plateau. This means that, for example, there can be simple L* and H* pitch accents in contrast

with each other, as well as rising and falling accents, as in a dynamic tone model. Third, the local pitch range is determined by a variety of effects, such as phrasal prominence relationships or the occurrence of **downstep** (a compression of phrasal pitch range that reduces a 'downstepped' non-initial H target and all following H tones within the phrase) and **upstep** (a raising of the phrasal pitch range beginning at a H- phrase accent). These effects are specified independently of the tone level, so that a H tone in one part of the intonation contour for an utterance can be lower than a L tone elsewhere in the same utterance. Fourth, the tones for any phrase are distinguished functionally either as being **edge tones** or as being affiliated with **pitch accents**. The absolute pitch value of a tone depends on its function as well as on its position; for example, a L tone that defines the beginning of a L+H* rising pitch accent can be higher than a L% tone at the following intonational phrase boundary. The function of a tone also determines its timing relative to the autosegmental projection of consonants and vowels; a pitch accent is aligned to the segments of the relevant stressed syllable whereas an edge tone is aligned to the segments at the relevant phrase boundary. Fifth, there are contrastively H versus L edge tones at two levels of intonational phrasing, associated with two different degrees of juncture or boundary strength (i.e., the **intermediate phrase** versus the **intonational phrase**). Moreover, the lower-level edge tone — the **phrase accent** — is aligned to affect the pitch contour beginning immediately after the last tone target of the accent that is aligned to the syllable with **nuclear stress** (the most prominent accented item in the intermediate phrase). Thus, the phrase accent defines the beginning of the post-nuclear **tail**[1]. This decomposition of the contour beginning at the syllable with nuclear stress means that the intonation contour over the pre-nuclear **head** can be described in terms of the same inventory of pitch accent types available for the nuclear accent.

The immediate antecedents of the MAE_ToBI model are Pierrehumbert (1980) and Beckman and Pierrehumbert (1986) for the decomposition of the intonation contour into functionally distinct groups of H versus L tone; Ladd (1983) for the treatment of downstep; and Price et al. (1991) and Wightman, Shattuck-Hufnagel, Ostendorf and Price (1992) for the treatment of juncture. However, all aspects of the model are also grounded in a long history of research on intonation and prosody in the language. For example, both the claimed relationship between the intonation contour and the sense of disjuncture or phrasing, and the notion of the pitch accent, predate Pierrehumbert's (1980) grammar of MAE tunes by several decades. The MAE intonation system is closely related that of standard Southern British English (SBE), which is one of the most studied in the world. Our modern understanding of the inventory of SBE intonation patterns traces its roots to astute observations by teachers of English as a foreign language beginning in the first decades of the last century (Palmer 1922; Armstrong and Ward 1926) and studies of large corpora in the 1960s (Halliday 1967; Crystal 1969) and later (e.g., Gussenhoven 1984). Ladd (1980) summarised important points of consensus among the then current competing models of SBE intonation, and pointed out the relationship between these models of SBE and Bolinger's (1958, 1964) observations of intonation patterns in MAE. These consensus properties include (1) the use of pitch changes (dynamic tones) rather than pitch levels (tone targets) to identify an abstract representation of the tune, independent of the associated text; (2) the connection drawn between tune and phrasing for at least one, and often for two levels of grouping (cf. Trim 1959); (3) the notion of pitch accent as encoded, for example, in Kingdon's (1939) **tonetic stress marks**; and (4) the specification of a distinct inventory of tone patterns for the nuclear versus pre-nuclear accents (cf. the review of treatments of the head in

Ladd 1980). The second and third of these properties are the earlier antecedents for the claimed link between edge tones and phrasing, and for the notion of pitch accent.

Note, however, that the first and fourth of the properties that Ladd listed are not shared by the model of American English intonation that is encoded in the MAE_ToBI conventions. Instead, MAE_ToBI follows Pierrehumbert (1980) in adopting a tone target model rather than a dynamic tone model, and also a decomposition of the intonation contour starting at the nucleus that unifies the inventory of nuclear and pre-nuclear accent types.

The unified inventory for nuclear and pre-nuclear pitch accents builds on Bruce's (1977, 1982) seminal work on Swedish accent and intonation. In Bruce's model the pitch contour for an utterance is decomposed into a sequence of independent tone targets associated with different parts of the text. Every content word has an associated word accent and every intonational phrase has a phrase accent. Word accents are anchored at the stressed syllable in each successive word, the phrase accent is inserted into the tone string immediately after the lexical pitch accent of the word bearing the sentence stress (which is not necessarily the last content word), and boundary tones are anchored at the phrase boundaries. Variation in word tone shapes can then be neatly explained (and handily captured in speech synthesis) by modeling the effects of neighboring tones (e.g., **undershoot** when two tones are crowded together) and of phrasal position (e.g., the word accent tones are successively reduced by downstep after the phrase accent) independent of the word tone specification. Pierrehumbert (1980) proposed a similar decomposition of American English intonation contours into tones contributed by the pitch accents, and two types of edge tones: boundary tones proper (which are anchored only at the intonational phrase edge) and phrase accents (which are anchored both to the end of the intermediate phrase and after the nuclear pitch accent in the phrase). With this decomposition of the pitch contour around and after the nuclear accented syllable, Pierrehumbert was able to unify the treatment of accent in the head and in the nucleus, and to account for the shape of the tail. The model also resolves a fuzziness inherent in dynamic tone accounts of alignment when a contour tone is "stretched out" over different numbers of syllables under different conditions of stress and phrasing. The fall-rise nucleus, for example, does not simply flatten out to yield a shallower fall and shallower rise when the post-nuclear tail is longer than a few syllables. Rather, the low inflection point stretches out into a plateau, so that the fall remains steep and anchored at the stressed syllable and the rise remains steep and anchored at the phrase end. A level tone model specifies this alignment pattern by decomposing the dynamic shape into a sequence of simple tones that are aligned to specific prosodic landmarks. The fall and the rise of a fall-rise nucleus, for example, can be described as a sequence of H* pitch accent, L- phrase accent, and H% boundary tone, which are anchored invariantly to the appropriate prosodic domain however many syllables there are after the sentence stress. (See Pierrehumbert 1980, 2000, Liberman and Pierrehumbert 1984, and Pierrehumbert and Beckman 1988, for a discussion of this and other empirical advantages of the level-tone model as compared to a dynamic-tone model.)

The tone target analysis adopted in the MAE_ToBI model also resolves problems with earlier models of MAE intonation that posited more than two tone levels. For example, Bolinger (1951) criticised Trager and Smith's (1951) four-level analysis for its failure to differentiate continuous pitch range variation from categorical intonational contrast. (See Pierrehumbert 1980 or Pierrehumbert 2000 for a summary of this and other problems.) Pierrehumbert's solution was to posit tone targets, but only at two rather than four or five levels, and then to develop a more elaborated model of the relationship between tone target and backdrop pitch range in order to account for the observations of systematic phonetic differences among more than just the one

relatively low ('L') and the one relatively high ('H') tone level. The elaborations of the model of the relationship between tone target and pitch range are of two types: positionally-conditioned local variation in the realisation of the same tone (e.g., a 'L%' boundary tone target anchored at a phrase boundary typically is lower than the 'L' target in a 'L+H*' pitch accent) and more global variation in the backdrop pitch range (e.g., downstep reduces the pitch of a 'H' and all following 'H' targets until the end of the phrase).

Pierrehumbert's two-tone model of English intonation owes much to insights about tone patterns in African tone languages (e.g., Leben 1973; Anderson 1978), and has been adopted widely in the subsequent literature on intonation in English (e.g., Gussenhoven 1984) and in other languages (cf. Ladd 1996). In common with the Africanists, Pierrehumbert (1980) understood downstep to be triggered by a H L H alternation of tone levels within a phrase. In English, such an alternation only occurs when at least the first of the H tones is part of a pitch accent. In standard Japanese, downstep occurs specifically when a HL sequence is a pitch accent (see, inter alia, McCawley 1968, Pierrehumbert and Beckman 1988). An alternative understanding of downstep in English, therefore, is that it is triggered by the alternation of tone levels specific to a rising or falling pitch accent (see Beckman and Pierrehumbert 1986). A third alternative, proposed by Ladd (1990), is that downstep is a direct, iconic signal of reduced prominence for a later accent relative to an earlier accent. None of these accounts, however, has been tested against large corpora of labelled speech. The MAE_ToBI conventions, therefore, adopt a more theory-neutral approach, following Ladd (1983). Specifically, downstep is marked explicitly on the first affected tone, using the downstep diacritic common in autosegmental treatments of African tone languages. Thus, "L+!H*" signifies a rising pitch accent that is downstepped relative to an immediately preceding H tone target. While the immediate effect is that the peak value is intermediate between that of immediately preceding H and L tones, this is not a mid (M) tone, but a H tone in a compressed pitch range. That is, the "!" in the tag also indicates the beginning of a stretch of speech in a compressed range, such that not only the downstepped !H target in the L+!H*, but all subsequent H tones will be lower relative to their 'expected' value had they occurred before the downstepped tone.

The treatment of juncture in Price et al. (1991) also has a long history. Essentially, it unifies earlier phonetic work on the relationship between juncture and syntax, and on discourse level effects on duration and pitch range (e.g., Lehiste 1960; 1975; O'Malley, Kloker and Dara-Abrams 1973; Cooper 1976; Nakatani, O'Conner and Aston 1981; Gee and Grosjean, 1983), with the treatment of segmental sandhi effects in metrical phonology (e.g., Selkirk 1978; Nespor and Vogel 1986). The further insight achieved by comparing transcriptions at the first Prosodic Transcription Workshop in 1991 was that this treatment of juncture could also be linked to the treatment of intonational phrasing in Beckman and Pierrehumbert (1986) by identifying the perception of juncture at break index levels 3 and 4 as being cued in part by intonation contour shapes at the edges of Beckman and Pierrehumbert's (1986) intermediate phrase and intonational phrase, respectively. We also agreed that any perceived variation in boundary strength above the intonational phrase (BI=4) could not be identified with the domains of prosodic effects such as the distribution of phrase tones, and therefore, that 'utterance' and 'paragraph' ends should not be marked (see Section 2.4). The transcription of tones following Beckman and Pierrehumbert (1986) and Ladd (1983) thus provides the **To**nes part of MAE_ToBI, and the transcription of perceived juncture at the ends of prosodic units in Price et al. (1991) provides the **B**reak **I**ndices part. In the next section we will review these and other obligatory parts of a MAE_ToBI record.

2.4. Overview of the MAE_ToBI conventions

A full MAE_ToBI record of an utterance has at least six parts, listed in Table 2.1 and illustrated in Figures 2.1 through 2.3. Of these six parts, two are continuous phonetic records and four are symbol strings. The primary continuous phonetic record is an audio recording of the utterance. In the case of the utterances in three figures, these are digital recordings, on the CD that accompanies this book. The waveform in the top panel of each figure is a graphic representation of this recording. The other continuous phonetic record is some representation of the fundamental frequency (F0) contour. This could be an analogue representation such as a narrow-band spectrogram, or a digital representation such as a string of numbers calculated by some F0 tracking algorithm. If the representation is of the latter form, it is useful to provide also some graphical representation of the numbers, as in the bottom panel of each figure, which shows the output of an autocorrelation-based F0 tracking program applied to the audio recording. The panel in between the waveform and the F0 contour in each figure shows the four obligatory symbol strings, ordered vertically in these displays starting with the tier of labels for **Tones** (i.e., a symbolic transcription of the intonation contour) at the top. A full list of the ten basic tonal morphemes and of the other Tones-tier labels is shown in Table 2.2. The two symbol strings just below the Tones tier are labels for all of the **Words** in the audio recording (i.e., an orthographic transcription) and labels of **Break Indices** (i.e. a number indicating the perceived degree of boundary strength) for each of the labels in the Words tier. The four[2] basic Break-Index values and the several diacritics and other labels for phenomena such as a marked prolongation that disrupts the intonation contour are given in Table 2.3. At the bottom of the panel showing the symbol strings is the **Misc** tier of labels for events such as coughs or disfluencies — anything marking speech events of interest or questions or unusual configurations of labels on the other three symbolic tiers. (The Misc tier may optionally contain other labels, depending upon the aims of the labeling project. See, for example, the Misc-tier labels in Figure 2.1 and the discussion in Section 2.6.)

Insert Table 2.1 about here: The six obligatory parts of a MAE_ToBI record.
Insert Table 2.2 about here: The inventory of MAE_ToBI Tones-tier labels.
Insert Table 2.3 about here: The inventory of MAE_ToBI Break-Indices-tier labels.
Insert Figure 2.1 about here: waveform, F0 contour, and ToBI xlabel windows for utterance <okay, there are a couple flights>
Insert Figure 2.2 about here: waveform, F0 contour, and ToBI xlabel windows for <iraqi>
Insert Figure 2.3 about here: waveform, F0, and ToBI xlabel windows for <quincy>

Each of the labels on the symbolic tiers is an index to events observed in one or both of the continuous phonetic records, and this indexing function is accomplished by the time stamp associated with the label. For example, the time stamp for a label on the Words tier indexes the end of that word in the audio signal, and each label on the Break-Indices tier should share the same time stamp as the label for the immediately preceding word. Thus, in Figure 2.1, there are six word labels (discounting the '<SIL>' — see below), and each of these labels is aligned with a 1 or a 4 on the Break-Indices tier beneath it. Some of the labels on the Tones tier also inherit these time stamps. For example, the 'H-H%' sequence at the end of *okay* in Figure 2.1 is aligned with the time stamp marked by Break-Index 4. Other Tones-tier labels (e.g., the 'L*' preceding the 'H-H%' sequence in Figure 2.1) must be provided with their own independent time stamps that refer to the fundamental frequency record or to both the F0 and the audio recording. That is, the tags on the Tones tier are of three types, with slightly different alignment conventions. The

three types of tone labels are (1) edge tone labels at each break index 3 and 4; (2) pitch accent labels for each accented syllable within the intermediate phrase; and (3) two 'phonetic' labels at points where it is useful to extract times and fundamental frequency values in investigation of peak alignment and phrasal pitch range. These last two labels are the '**<**' tag that in careful labelling can be used to mark accent peaks which are not aligned with the relevant accented syllable, and the 'HiF0' label that helps gauge the pitch range for an intermediate phrase that contains at least one accent-related 'H' tone. For example, the utterance in Figure 2.1 contains two intermediate phrases, but only the second has a HiF0 label, because the H tones in the first are both edge tones rather than accent tones. The placement of these two labels depends on the placement of accent labels, and will be described in more detail below (see also Section 2.6). By contrast, the edge tones depend on the Break-Indices tier (or the Words tier in some cases).

That is, the edge tones are the 'L-' and 'H-' phrase accents (marking the ends of all intermediate phrases), the 'L%' and 'H%' final boundary tones (marking the ends of all intonational phrases), and the '%H' initial boundary tone. The original MAE_ToBI decision that only the end of every word needs to be marked on the Words tier is related to the fact that every intermediate phrase must end with a phrase accent, and every intonational phrase must end with the phrase accent of the last intermediate phrase and an immediately following boundary tone, whereas the initial boundary tone is rather marginal in English. (Typically, the first well-defined tone target in an utterance is a pitch accent, on the first likely candidate syllable for accent, giving rise to the percept of stress shift if it is a syllable with lexically "secondary stress" — cf., Shattuck-Hufnagel, Ostendorf and Ross 1994.) Thus, a phrase accent or a final boundary tone can simply inherit the time stamp for the break index that marks the end of the relevant intermediate phrase or intonational phrase. The '%H' initial boundary tone, on the other hand, must be aligned with the beginning of the phrase-initial word, and this will not already be marked in the Words tier, unless the transcriber has inserted a '<SIL>' or '#' label after each pause (see below).

The pitch accents include two accents in which the F0 remains low or falls to a lower level on the accented syllable ('L*' and 'H+!H*'), two accents in which the F0 remains high or rises to a peak on the accented syllable ('H*' and 'L+H*'), and a scooped accent ('L*+H') which has an F0 minimum within the accented syllable followed by an F0 peak. In careful labelling, these three sets of accent labels are placed differently with respect to the accented syllable. That is, minimally the label for a pitch accent should be placed somewhere within the syllable to which it is associated, so that the time stamp can identify the accented syllable when there is more than one candidate in a word. In careful transcriptions, however, further constraints are followed — namely, this time stamp is placed at the F0 minimum for 'L*', 'H+!H*', and 'L*+H', and at the F0 maximum for 'H*' and 'L+H*' so long as the maximum occurs within the accented syllable. If the maximum is later than the end of the syllable, the accent label is aligned to the amplitude peak within the accented syllable, and a '<' label is placed at the actual F0 peak.

The '<' can also be used to mark the necessarily later peak for 'L*+H' in such careful labelling, which is particularly useful if the aim is to investigate the phonetics of alignment. For example, we have observed differences across datasets in how late a peak can be and still be perceived as a 'L+H*' rather than a 'L*+H', and it would be useful to know what stylistic or dialectal differences influence this category boundary. Analogously, the HiF0 label was motivated by its usefulness for investigating the phonetics of pitch range variation as a cue to discourse topic structure or intentional structure (e.g., Ayers 1994; Grosz and Hirschberg 1992). This use of the HiF0 label is related to a claim implicit in our decision not to mark break index

values above the intonational phrase. There was a strong feeling among a sizeable group at the first workshop that the grouping of intonational phrases into 'utterances' and 'paragraphs' is qualitatively different from the grouping of morphemes into 'prosodic words'[3] dominated by intermediate phrases, and into intermediate phrases dominated by intonational phrases. The consensus understanding of the metrical hierarchy in phonological theory at the time was that units such as prosodic word and intonational phrase constitute a non-recursive (and possibly even strictly-layered) tree (see, e.g., Nespor and Vogel, 1986, Pierrehumbert and Beckman, 1988). That is, a prosodic word cannot dominate other, smaller prosodic words at a lower level of grouping, and an intonational phrase cannot dominate other, smaller intonational phrases. Also, each such unit can be defined in terms the distribution of categorical phonological markers in a way that makes the bracketing amenable to a finite numerical index. The phonological domain for the distribution of the phrase accent is a '3' because that is where the intermediate phrase is in the hierarchy; segmental sandhi phenomena such as palatalisation of a coda consonant by a following palatal (e.g., in *meet you*) does not occur across a domain bounded by a phrase accent, and a boundary tone never occurs within this domain. Units such as 'utterance' and 'paragraph', by contrast, seem to be recursive, reflecting something like a hierarchy of larger discourse topic and embedded smaller subtopics (see, e.g., Grosz and Sidner 1986). Work such as Lehiste (1975) further suggested that this recursive discourse structure will be reflected iconically in such continuous phonetic measures as the durations of inter-utterance pauses or the backdrop pitch range over successive intermediate phrases, rather than by direct categorical markers such as the distribution of boundary tones. Therefore, we decided not to label degrees of perceived disjuncture above the intonational phrase, so that the break indices could be interpreted as a direct implementation of the non-recursive prosodic hierarchy.[4] The MAE_ToBI conventions thus differ from other marking systems such as INTSINT, which propose symbolic labels for 'paragraph' and the like, based on the implicit claim of a single category of extra 'pitch resetting' at paragraph beginnings or a single discrete level of 'extra-low pitch' at paragraph ends (Hirst and Di Cristo 1999: 17).

Note that there is an ambiguity inherent in the conventions for relating the words and break indices to the audio recording. If only the end of each word is marked, as specified in the original MAE_ToBI conventions, then the Words tier labels a string of events. That is, each Words-tier tag (and the corresponding Break-Indices label) marks the boundary between a pair of words or between the last word and the following silence. As a result, the onset of the first word and any mid-utterance pauses are not marked. However, there are many occasions, such as when training automatic speech recognition systems, when it is more useful to treat the 'words' as labelling intervals in the signal rather than single time points; for these purposes '<SIL>' or '#' is often used to mark any word beginnings which are not coterminous with the end of the preceding word — i.e., at the beginning of the first word in the recording and after every pause. This was the expedient we adopted in converting the labels for the example utterances that accompany the *Guidelines to ToBI Labelling* into an EMU database.[5] These '<SIL>' or '#' labels are thus the one exception to the rule that every word label must have a corresponding break index value. They are also a systematic source of inconsistency between 'dialects' of MAE_ToBI, although the inconsistency seems to be disappearing rapidly as more and more sites adopt this convention.[6]

This distinction between interval and event is encoded explicitly in the syntax for the Misc tier, which specifies two types of labels, differentiated by the interpretation of the time stamp. There are labels for localised events, such as 'disfl' (which marks the approximate time point where some disfluency is perceived), and there are paired labels for effects that occur over

identifiable longer intervals, such as 'disfl<' and 'disfl>' (for the beginning and end of an identifiable stretch of disfluent speech). The dual syntax was our solution to the question of how to label a motley set of phenomena that could be either intervals (similar to words) or events (similar to breaks or tone targets). Another solution would have been to allow only the second type of label, but to permit the pairs of labels for more localised events to be separated by only one time frame, as in the EMU version of the *Guidelines to ToBI Labelling*.

We have illustrated the syntax of the Misc tier with the disfluency labels, because false starts and repairs are a perennial source of concern in analyzing spontaneous speech. They are also a common source of 'ungrammaticality' or uncertainty about how to mark tones and break indices. One of the practical principles that we listed in Section 2.2 is that a ToBI annotation system needs to be reliable in order to be useful. This means that labels need to be applied consistently across sites and among transcribers at any given site, and that mechanisms must be provided for dealing with transcriber uncertainty and phonetic ambiguity. The MAE *ToBI Annotation Conventions* note that disfluencies "are not automatically detectable, and the absence of markings for them makes it difficult to parse the Tones and Break-Indices tiers. For these reasons, transcribers are urged to mark disfluencies on the miscellaneous tier using 'disfl<' and 'disfl>' (or 'disfl' if the disfluency is extremely localised)" (Beckman and Hirschberg 1994: 5). Related labels on the other tiers are the 'p' diacritic on the Break-Indices tier, to mark the "perception of audible hesitation (for example, an abrupt cutoff or a prolongation)", and the '%r' label on the Tones tier, "to indicate a 'contour restart' — i.e. the initiation of a new intonational contour after a disruption".

The MAE_ToBI conventions also specify a number of methods by which transcribers can indicate uncertainty in the absence of disfluency. For example, uncertainty about the strength of a break index is indicated by adding a '-' to the right of the index value. Thus, '4-' indicates that the transcriber found the preceding and following word to be somewhat more closely conjoined than is usual for words separated by a level 4 break index, but less clearly conjoined than those at a level 3 break index. Uncertainty on the Tones tier is indicated by a set of special symbols rather than by a diacritic, although all of these symbols include '?' as a final element. Thus, '*?' indicates uncertainty about whether or not a syllable has a pitch accent; '-?' indicates similar uncertainty about whether a phrase accent has occurred; and '%?' indicates uncertainty about whether a boundary tone has occurred. (Note that the latter two symbols should be accompanied by '3-' and '4-' on the Break-Indices tier.) Where tonal uncertainty concerns the type of tone, on the other hand, we employ 'X*?', 'X-?', and 'X%?' instead. So, while '*?' means 'I don't know whether this syllable is accented or not,' 'X*?' means 'I believe that this syllable is accented but I don't know which pitch accent type to assign to it.' The first sort of uncertainty is exemplified in the first intonational phrase in Figure 2.1, where the labellers could not decide whether the first syllable of *okay* has a pre-nuclear L*, and in the last intonational phrase in Figure 2.3, where the labellers could not decide whether the apparent rise in F0 onto *Shore* represents a pre-nuclear H* or was just a point in the interpolation between the preceding L% boundary tone and the following H* on *Cab*. The latter sort of uncertainty is exemplified in the second intonational phrase in Figure 2.2, where the succession of downstepped accents on *southern* and *Iraqi* has compressed the pitch range to such an extent that there is no objective way to decide between a third !H* and a L* for the nuclear accent on *cities*.

Note that in many cases, labeler uncertainty can be attributed directly to aspects of the signal — i.e. to real phonological ambiguity that cannot be resolved just by training the transcriber to label more carefully. The MAE *ToBI Annotation Conventions* describe two such cases of phonological ambiguity which can be tagged with the Tones-tier uncertainty symbols:

"A typical case where '*?' might be used is for a very strong syllable in a part of an utterance between a prenuclear H* and a nuclear H*, where the F0 contour is flat and high because of the preceding and following tones, making it difficult to detect intervening H* accents. A typical case where 'X*?' might be used is a part of an utterance where the labeller cannot tell whether an accent is a L* accent or a H* accent in a compressed pitch range." The nuclear accent in the second intonational phrase in Figure 2.2 is an example of the latter sort of ambiguity. Thus, typical uses of '*?' and 'X*?' are cases of ambiguity involving mismatch between the perceived prominence of a syllable in the audio recording and the tonal markings of accentuation in the F0 record.

The '2' symbol on the Break-Indices tier was intended similarly to mark cases of ambiguity involving two types of mismatch between the perceived sense of disjuncture and the tonal markings of prosodic grouping. The symbol can mean that the perceived disjuncture is at break index level 3 or 4 (e.g., final lengthening and pausing that is appropriate for an intermediate phrase or intonational phrase), but that there is no clear indication of a phrase accent in the tone pattern. (The sense of pause between each pair of words in the sequence of downstepped accents in the second intonational phrase in Figure 2.2 is a good example.). Alternatively, the symbol 2 can mean that the perceived disjuncture is at break index level 1 (an ordinary phrase-internal juncture) despite the clear occurrence of a phrase accent or even a phrase accent and boundary tone sequence. (The steep fall from H* to L- after *Quincy* in Figure 2.3 is a good example.) Thus, break index '2' is not part of the metrical hierarchy per se. Rather, it provides a way to tag such cases of mismatch without jettisoning the definition of break indices 3 and 4 in terms of the otherwise well-governed coupling between tune and prosodic grouping. This provision distinguishes MAE_ToBI from earlier tagging systems for English which have no projection of metrical structure separate from the 'tone unit' (e.g., Crystal 1969) or which disclaim any correspondence between higher-level units in the metrical hierarchy and the domain(s) for associating tune to text (e.g., Gussenhoven 1990).

An aspect of this comparison to earlier tagging systems that was very salient in the discussion at the second workshop is the issue of redundancy and its effect on efficiency. That is, the explicit projection of the metrical structure onto a separate Break-Indices tier even though the higher level breaks are defined in terms of categorical tonal marks introduces redundancy that is somewhat at odds with the principle that ToBI conventions should be efficient. For example, every time a boundary tone is marked on the Tones tier, a 4 should be marked on the Break-Indices tier, and vice versa. The MAE *ToBI Annotation Conventions* acknowledge this redundancy and recommend that transcribers 'avail themselves of routines for automatically inserting redundant labels on either tier'. The function of the symbol 2 brings out the value of separating the function of the phrase accent in marking prosodic grouping from its function of autosegmental contrast. For example, the L- after *Quincy* in Figure 2.3 contrasts both with H- and with !H-, but there is no clear pause to separate *Quincy* (the caller's response to the operator's opening *What city please?*) from the following sentence (the caller's request for the telephone number). The redundancy allows the transcriber to modularise the tagging of the two separate structures and makes the tagging system theory-neutral by comparison to the implicit claim of strict correspondence in Crystal (1968) or the explicit claim of complete independence in Gussenhoven (1990). At the same time, the MAE_ToBI system does not exploit this modularity as gracefully as do several of the other ToBI-framework systems described in this book. The dual usage of the symbol '2' is one of the most awkward aspects of the MAE_ToBI conventions, and a common source of confusion for new transcribers. By contrast, the ToBI framework systems for Japanese, Korean, and Greek have introduced an explicit mismatch

diacritic ('m') on the Break-Indices tier, so that break index 2 can be used for a well-defined level of the prosodic hierarchy for the language, such as the accentual phrase for Korean and Japanese, and the intermediate phrase for Greek (see, Venditti 1997, this volume; Jun this volume, ch. 9; Arvaniti and Baltazani this volume).[7] In the next section, we discuss some extensions of this modular design to other languages and to other phenomena.

2.5. Extensions of ToBI

One of the first extensions of our work in developing MAE_ToBI was the application of the basic ToBI framework design to other languages, such as northern German, Tokyo Japanese, and several other languages with less well-studied intonation systems. Although we made a firm decision at the first workshop that any tagging system we developed would have to be language-specific, we suspected that aspects of the MAE_ToBI design, particularly the explicit separation of autosegemental tonal content from hierarchical metrical structure, could be extended to other languages. Indeed, all of the systems described in this volume have implemented Tones and Break-Indices tiers, as well as some form of orthographic tier to provide the initial set of time stamps for relating tones and prosodic unit boundaries to the audio and F0 signals. As Jun (this volume, ch. 17) points out, the applicability of the **To**nes versus **Break-Indices tier** structure to languages as typologically different as German, Japanese, and Cantonese suggests a prosodic universal: many languages seem to structure utterances into a hierarchy of prosodic units, at least some of which are categorically marked by the tone pattern.

It is important to emphasise, however, that the four labelling tiers described above, together with the audio and F0 records, are only the *obligatory* parts of the MAE_ToBI record. We had no expectation that these four tiers and two types of continuous record would suffice for tagging systems for all languages or even for all users of MAE databases. Rather, the originators of MAE_ToBI fully expected individual sites to add other ToBI tiers or other completely independent annotations as needed, in order to customise shared databases to their own purposes. For example, syntactic bracketing and part of speech can be projected in a hierarchical labelling system that is separate from the metrical hierarchy of break indices and the tonal projection in MAE_ToBI, in order to train algorithms for predicting intonational phrasing and accentuation in text-to-speech systems (see, e.g., Hirschberg 1993; Hirschberg and Prieto 1994; Ostendorf and Veilleux 1994; Koehn, Abney, Hirschberg and Collins 2000; Hirschberg and Rambow 2001). Discourse structure also can be marked separately, and there are several standard discourse tagging systems available, based on different models of discourse structure. For example, the tagging system described in Nakatani, Grosz, Ahn and Hirschberg (1995) implements Grosz and Sidner's (1986) model of the speaker's hierarchy of discourse purposes as a basis for segmentation and global coherence, whereas Allen and Core's (1997) DAMSL system, Carletta, Isard, Isard, Kowtko and Doherty-Sneddon's (1996) Map Task annotation scheme, and the annotation schema developed at the Discourse Resource Initiative workshops[8] implement more locally defined 'dialog act' models of discourse.

Like the MAE_ToBI system, a number of the discourse tagging schemes mentioned above have been tested in several inter-transcriber reliability exercises, and there has been much research in the ToBI community relating discourse structure tagged in these models to such continuous phonetic measures as syllable and pause durations, amplitude variation within and across discourse segment boundaries, pitch range relationships across successive intonational phrases, and so on. Studies now exist both for English (e.g., Grosz and Hirschberg 1992; Swerts and Ostendorf 1997; Hirschberg and Nakatani 1996) and for several other languages for which ToBI framework models are available (see, e.g., Venditti 2000 for Japanese). This research

tends to support our prediction that discourse structure does not have a one-to-one correspondence to categorically marked prosodic domains above the intonational phrase. That is, there do not seem to be special tones or other categorical events that distinguish, say, the ends of paragraphs from the ends of sentences internal to a paragraph. Instead, the discourse hierarchy seems to be marked only by a fine, continuous control of variation in phonetic measures that can iconically reflect such (non-phonological) relationships as coordination versus embedding of discourse segment purposes.

Another common extension to the original MAE_ToBI tiers is to tag consonants and vowels on a separate autosegmental tier from the tonal projection. Any site which uses ToBI labelled data to train an automatic speech recognition or speech synthesis system does this, and the emerging convention is to call such a projection a **Phones** tier. At many such sites, a first-pass Phones-tier labelling is done automatically using an alignment program, such as Aligner (Wightman and Talkin 1994) or some other similar HMM-based automatic transcription alignment system. For such sites, the Words-tier labels are then also derived automatically from the Phones alignment. A **Syllable** tier can also be added from the Phones tier, using a simple syllabification script, if desired. Stress labels can also be assigned, using an online dictionary in conjunction with the pitch accents on the Tones tier. Syrdal, Hirschberg, McGory and Beckman (2001) describe such a syllable-tagging scheme, designed for training a variable-unit concatenative text-to-speech system. Ostendorf and Ross (1997) used similar tags in training an automatic intonation recogniser.

At the Fourth Prosodic Transcription Workshop, we discussed whether the Phones and Syllable tiers should be obligatory, but decided that it was not practical to make them so until alignment software becomes commonly and freely available. Some ToBI framework systems for other languages, however, have made other decisions. For example, the Pan-Mandarin ToBI system system (M_ToBI, see Peng, Chang, Tseng, Huang, Lee and Beckman this volume) specifies that a syllable-by-syllable segmental transcription of the utterance must be provided on a **Syllable** tier. Moreover, each interval marked on this M_ToBI tier must be labelled for its perceived degree of stress on an independent **Stress** tier. The contrast between stress levels 1 and 2 is defined by the categorical absence versus presence of an associated (lexical) tone, reminiscent of the definition of the levels 'unaccented' versus 'accented' in the English stress hierarchy. It is important to note that, while they are like the break indices in being a numerical index of perceived 'strength', these Stress tier labels differ from the break indices in marking intervals rather than events; they index the syllable's own strength rather than the following boundary strength. Thus, they constitute another metrical hierarchy that is independent of the metrical hierarchy of prosodic groups on the Break-Indices tier. That is, where the break indices correspond to a bracketing hierarchy (a metrical tree), the M_ToBI stress levels correspond to a rhythmic hierarchy (a metrical grid).

It would be easy to project the stress labels from the Syllable tier in Ostendorf and Ross (1997) and Syrdal et al. (2001) onto a similar independent Stress tier for MAE, although there are questions that need to be addressed before MAE_ToBI could make such a Stress tier obligatory. In particular, should pre-nuclear accent and nuclear accent project different levels of stress? And how many levels should there be below the accented/unaccented decision?

The labels for the (obligatory) Syllables tier in the Cantonese ToBI system (C_ToBI, Wong, Chan and Beckman this volume), by contrast, do not mark stress levels. Rather, this C_ToBI tier provides a transliteration in the roman alphabet of the standard Hong Kong reading of the Chinese and 'serves the function of the Words tier for sites that do not have a way to input

and/or read Chinese characters' (Wong et al. this volume: #). Moreover, it is difficult to see how one could provide categorical definitions for consistently marking stress levels or syllables in Cantonese. Cantonese does have syllable-level lenition effects. However, these effects do not selectively target the segments in the less stressed syllable in a sequence of two, as in the superficially similar Mandarin lenition processes. Rather, the Cantonese fusion effects merely 'erase' the inter-syllable boundary, along a continuum from weakening or deleting the intervening consonant(s) to merging the two vowel qualities into an intermediate value. Moreover, it is typical for both syllables to maintain their status as tone-bearing units in the tonal projection even in cases where the consonant and vowel effects are so extreme that the syllable count is no longer clear in the segmental projection. Thus, Cantonese syllable lenition seems to be more a matter of prosodic grouping than of prominence-based rhythmic structure, and it is transcribed in C_ToBI by projecting a phonetic transcription of the affected syllable sequence onto a single prosodic unit on the **Foot** tier, while marking a value of 0 at the 'erased' boundary on the Break-Indices tier.

Comparing the (obligatory or imaginable) projections from the Syllables tier across the M_ToBI system, an extended MAE_ToBI system, and the C_ToBI system, then, we can make the following generalisation about the usefulness of the sort of modularity that the ToBI framework promotes. Projecting numerical representations of the two different metrical hierarchies separately from each other as well as separately from the categorical tonal or segmental marks of any level of grouping or prominence brings out the fundamental similarity between (some varieties of) Mandarin and (some varieties of) English, while emphasizing the rather different role of the syllable in Cantonese. Cantonese lacks anything comparable to Mandarin or MAE syllable stress, and there is no basis for projecting a tier that encodes each syllable's metrical grid level. A similar conclusion also accords with research on Tokyo Japanese (e.g., Beckman 1986), on Mayali and many other Australian languages (e.g., Fletcher and Bishop this volume), on many dialects of Basque (Hualde, Elordieta, Inaki and Smiljanić in press), on Quebec French (e.g., Cedergren and Perrault 1994), and so on, and it is difficult to imagine a Stress tier for any of these language varieties. Thus, the rhythmic structuring of utterances by a hierarchy of syllable prominences seems to be not nearly so wide-spread as the structuring of utterances by a hierarchy of categorically marked prosodic units, and it is quite appropriate that the framework as a whole is called ToBI, and not, say, 'ToBISL' (see Peng et al. this volume).

A third very common extension of MAE_ToBI is the recording of alternate analyses. For example, Syrdal et al. (2001) describe the use of a **Comments** tier to keep track of differences in proposed transcriptions when several transcribers are labelling a database together. (Figure 2.1 illustrates the use of the Misc tier for the same purpose.) The Syrdal et al. team found the Comments tier particularly useful in the initial stages of labelling a new voice or a new speaker style. Periodic discussion of recurring patterns in the sets of alternative transcriptions allowed the transcribers to compare across utterances to calibrate their criteria for distinguishing between superficially similar contours such as two rising shapes, and to articulate cues such as subtle differences in timing, slope, or transition extent. In this way, annotating many utterances together brought out important questions for future research, such as the possibility of inter-speaker or perhaps inter-dialectal differences in the location of the boundary between the L+H* and L*+H categories along a peak timing continuum. This question could be investigated in several ways. For example, one might elicit imitative productions using a synthetic continuum, as in Pierrehumbert and Steele (1989). Alternatively, one might gather and label a suitably large

multi-speaker corpus recorded in dialogue tasks that are designed to elicit tokens of these two accent types.

The **Phonetic** tier in the expansion of the Korean ToBI system proposed by Jun (this volume, ch. 9) is similarly motivated. That is, the use of a separate tier to tag 'non-canonical' tone targets at the edges of accentual phrases seems difficult to justify if the term 'phonetic' is taken too literally — i.e., if the labels on this tier are viewed as a substitute for a more direct phonetic representation such as the F0 contour. However, Jun's Phonetic tier seems intended simply as an interim device to keep track of different kinds of apparent mismatch between the Tones and Break-Indices tiers, in a system that was perhaps codified too early, on the basis of an overly restricted set of speech styles. If larger and larger numbers of completely fluent L-ending accentual phrases are discovered as the system is applied to a richer variety of spontaneous speech styles produced by Seoul speakers who do not command any other variety of the language, then the Phonetic tier labels could become the basis for a rigorous discussion of how to revise the K_ToBI annotation conventions to provide better coverage. Alternatively, given the recent history of rural in-migration in South Korea (more than 30% of the population now lives in the Seoul area), it may turn out that L-ending accentual phrases signal solidarity with some other major regional variety, such as Chonnam or Kyongsung Korean. In that case, the 'non-canonical' tone targets might be better accommodated, not by expanding the inventory of accentual-phrase tones posited for the standard Korean model, but by setting up the appropriate alternative inventories for the other dialects involved, and adding a **Code** tier to indicate points of code-switching between the standard and regional dialect inventories, as proposed for different varieties of Mandarin by Peng et al. (this volume).

In the same way, we can imagine something like Syrdal and colleagues' Comments tier or Jun's Phonetic tier becoming an extremely useful tool in the initial stages of investigating whether the MAE_ToBI system can be applied to other varieties of English, as in Fletcher and Harrington's (1996) study of rise times in standard Anglo-Australian English and Fletcher and Warren's (2000) investigation of the different nuclear rises in both standard Australasian varieties (see also Section 15.2 in Fletcher, Grabe and Warren, this volume). That is, for example, we currently do not know whether the greater prevalence of rising boundary configurations in Australian English compared to MAE and SBE is due to a different pragmatics for such sequences as H+!H* !H H% or to a slightly different inventory of pitch accent types. Until we have enough data to suggest a definitive phonological analysis, a more cautious procedure might be to note examples of the potentially categorically different nuclear rise types on a separate Phonetic tier. However, we would caution against thinking of such a tier as a 'truly phonetic' representation, since this invites a misinterpretation of the status of symbolic tags in the ToBI framework that the originators of the MAE_ToBI system hoped to preclude. Since the original ToBI has been criticised as 'somewhat vague' (Nolan and Grabe 1997: 259) regarding the status of its symbolic tags, we would like to clarify our conception of their function here.

2.6. The phonetic representations in MAE_ToBI

As Pierrehumbert puts it in a recent paper, the original MAE_ToBI system 'is at the level of abstraction of a broad phonemic transcription, or rationalised spelling system, such as those of Korean and Finnish. Just as a broad phonemic transcription for any language must be guided by the phoneme inventory of that language (as revealed by the lexical contrasts), a ToBI-style transcription of the prosody and intonation of any language must be guided by an inventory of its prosodic and intonation patterns' (Pierrehumbert 2000: 26). This point cannot be emphasised too strongly. Symbols imply symbolic categories, and their use implies that the labeller has

recourse to extensive prior research of the sort that would allow positing an inventory of categories relevant for the language variety being transcribed. A theory-neutral or 'non-linguistic' symbolic transcription is impossible, even in theory. This is true even for transcribing consonants and vowels, whether in nonsense words in the labeller's own language or in utterances of a language that the labeller does not speak. Using the IPA or any similar alphabetic device to label speech data brings with it two strong theoretical claims: that utterances in the language being transcribed can be segmented into consonant and vowel categories, and that the language has consonant and vowel inventories that will not be too different in design from the inventories of already analyzed languages on which the IPA is based. The fact that the IPA has been used successfully by field researchers to build phonemic analyses of hundreds of languages justifies the assumption of these claims as a plausible first hypothesis, but it does not change their status as theoretical claims. Using the IPA does not prevent arguments about the 'phonetic' data in cases where one or the other claim is not fully justified, as can be appreciated by reading, for example, Coleman (1998) versus Dell and Elmedlaoui (1998) regarding the segmental status of Berber stop releases.

This characterisation of symbolic transcription as a necessarily phonological act holds even more strongly for tonal categories, since even varieties of the same language can differ markedly in their bases for segmenting the tone contour. For example, in a language with lexical tone, the pitch fall in a disyllabic word specified for a high tone on the initial syllable followed by low tone on the second in one variety might sound like a falling lexical tone in another variety where tone contrasts are specified over the whole word. Or, when the word is uttered in isolation, it might sound exactly like a disyllabic phrase with high lexical tone on the initial syllable followed by a tonally unspecified second syllable before a low boundary tone. Similarly, in a language where pitch accents are a relevant descriptive notion, a rising accent in phrase-final position in one variety might sound like a rise from a low pitch accent to a high boundary tone in another variety. Or it might be confused with a complex boundary tone in a third variety where the initial target is not anchored to any syllable because of influence from a substrate language that does not have accent. Transcribing the 'same' HL or LH pattern for all of the types in each set of cases cannot elucidate the differences among the varieties. There is no substitute for a detailed comparison of the F0 contours in a controlled examination of the 'same' tone sequences in other phrasal contexts and other pragmatic conditions.

A related point is that the fundamental frequency contour is an obligatory phonetic representation in the ToBI framework. Psychoacoustics research might eventually yield a better phonetic representation of the pitch contour, but a continuous phonetic representation can never be replaced by even the most detailed symbolic encoding of pitch events. To put it another way, symbolic tone labels in the ToBI framework are intended to 'tag' the intonation contour and not to 'encode' it. A tag is a pointer for retrieving phonologically relevant portions of the fundamental frequency and audio signals. It is not a symbolic representation of purportedly language-neutral pitch levels (Chao 1920) or pitch movements (Hirst, Di Cristo, Le Besnerais, Najim, Nicolas and Roméas1993). This is an especially useful way to understand the relationship between a MAE_ToBI transcription and the signal, because of the long history of research on the inventory of contrasts in varieties to which MAE_ToBI is known to be applicable. Someone who subscribes to a slightly different model of MAE or SBE intonation (such as the model of Crystal 1969 or Gussenhoven 1984) and who wants to investigate the phonetics of a category that is explicit in the alternative model but not in MAE_ToBI should still be able to use a database that has been labelled with MAE_ToBI tags, because there probably will be a correspondence between the tags that MAE_ToBI provides and the tags that the

researcher would have used if labelling in the alternative model. Although it is not likely to be a simple one-to-one correspondence (cf. Roach 1994), this correspondence probably will be lawful and hence useful. For example, anyone who wants to look at the factors determining the slope or extent of the pitch movements within a 'sliding head' (i.e., a pre-nuclear stretch with a peaky alternation of rise and fall instead of the gently downstepping trend that was the most common pattern for the head in Crystal's 1969 SBE corpus) can search for MAE_ToBI sequences such as H* L+H*, L+H* L+H*, and so on, to identify relevant instances of pre-nuclear pitch falls in a MAE_ToBI-labelled database. After listening to the utterance transcribed with the MAE_ToBI Tones-tier labels shown in (1)[9], the researcher who subscribes to Crystal's (1969) analysis, is free to replace the MAE_ToBI tones transcription in (1) with the tonetic stress marks in (2). Note that while these two transcriptions differ from each other in the phonological analysis assumed, with (1) describing the tune as a sequence of rising accents and (2) describing it as a sequence of falling accents, both differ even more fundamentally from Hirst's (1999: 73) INTSINT 'narrow phonetic' transcription of the sliding head in (3), which only notes the alternation of rise and fall without analyzing any part of the tune as the contrastive pitch event that is distinctively anchored to the accented syllable.

(1) There's a lovely yellowish old one.
L+H* L+!H* L+!H* L-L%

(2) There's a `lovely `yellowish `old one.

(3) There's a LOVEly YELlowish OLD one.
[⇑ ↓ ↑ ↓ ↑ ⇓]

The status of the symbolic tags in a ToBI framework system thus differs qualitatively from the status of the labels in the INTSINT system (Hirst et al. 1993; Hirst and di Christo 1999), as well as from the labels on the 'pitch movement' tier in IViE (Grabe, Nolan and Farrar 1998; Grabe 2000). The formulators of INTSINT wanted a tool for making the 'equivalent of a narrow phonetic transcription' (Hirst and di Christo 1999: 14), something that could be used to record 'pitch points' or 'targets' in intonation contours of a language even before the relevant categories for anchoring tune to text can be known. Thus, the INSTINT symbols ⇓ and ↓ are used to transcribe a fall in pitch, whether it is the necessarily steep interpolation from the peak of a L+H* rising accent to an immediately following L- phrase accent in the English utterance in (3), or the variably steep interpolation from the peak of a LH accent at the beginning of one word to the valley of the LH accent at the beginning the next word in the Finnish example in (4). (The latter is Hirst and Di Cristo's interpretation of an untranscribed Finnish example in Iivonen 1999. See Välimaa-Blum 1988 for our analysis of the fall as an interpolation between word-initial LH pitch accents, which will be more or less steep depending on the lengths of the words.)

(4) LAIna LAInaa LAInalle LAInan.
[⇑ ↓ ↑ ↓ ↑ ↓ ↑ ⇓]
'Laina lends Laina a loan.' (= Example (6) in Hirst and Di Cristo 1999: 16.)

The IViE 'pitch movement' labels similarly were formulated to be an 'auditory phonetic transcription' of pitch movements which the labeller uses to 'capture the phonetic realisation of a pitch accent in F0, at least as far as that is possible with a set of discrete tone labels' (Grabe 2000). *The IViE Labelling Guide* justifies this on pedagogical grounds (new labellers tend to 'rely too heavily on F0' and do not develop skills for 'careful listening'), and on the grounds that the discrete symbols make for easier compilation and comparison of phonetic realisation patterns ('H*+L, for instance, can be realised as hM-l, as mH-l, as hH-l, as mH-l (peak lag) or lH (truncation)'). In other words, like the INTSINT arrows and braces in (3) and (4), the IViE

'pitch movement' labels are 'phonetic' because they attempt to encode a downsampled F0 contour.

Calling a symbolic tier in the ToBI framework a 'Phonetic tier', by contrast, signifies something quite different. Rather than an attempt at a language-neutral encoding of the F0 pattern, a Phonetic tier label in the ToBI framework is a variety-specific tag, a way to keep track of possible phonological analyses of tune and of tune-text association at a stage when a research team has looked at enough data to make plausible guesses but does not yet have enough knowledge to specify a definitive inventory of contrasting intonations and prosodic patterns. The labels are discrete not because they are a grossly downsampled encoding of the F0 contour, but because they are hypotheses about discrete phonological categories. They still bear the same relationship to continuous phonetic representations that the final set of tags on the Tones tier will bear. That is, they will not replace the audio and F0 signals, but merely provide a convenient way to retrieve all instances where a particular hypothesis was made (along with instances of relevant potentially contrasting categories) so as to make possible a subsequent more penetrating examination of the phonetic record and of the associated phonological, syntactic, and pragmatic contexts. For this reason, our own advice for how best to fill the needs that *The IViE Labelling Guide* identifies is to use the F0 contour to the full, at least until a better representation of pitch is developed. The pedagogical ends, which are important, can be satisfied by exercises designed to teach new transcribers to effectively 'listen' to the F0 contour (e.g., McGory 2000). The more seasoned researcher, similarly, cannot hope to avoid the labor of finding effective ways to control the materials, of devising good F0 and other phonetic measures, and of creating new experimental tasks to test competing analyses, as in Liberman and Pierrehumbert (1984), Pierrehumbert and Steele (1989), Silverman and Pierrehumbert (1990), Hirschberg and Ward (1991), Shattuck-Hufnagel, Ostendorf and Ross (1996), and a host of other studies of tone alignment, tone scaling, and other aspects of MAE intonational categories.

Another way to understand the difference between these two approaches is to consider how vowel systems might be compared across two varieties of the same language. The formulators of IViE opt for a narrow symbolic transcription: 'We can describe all varieties of English as having three (historically) short front vowels i, e, and a. But if we want to describe the difference between the English of New Zealand and Yorkshire we need the phonetic categories [ɪ e ɛ] and [i̽ ɛ a̲].' (Nolan and Grabe 1997: 260). In the view that guided the development of the MAE_ToBI standard, a different approach would be adopted. Decades of research in sociolinguistics suggests that a more illuminating way to capture the difference in timbre between the corresponding New Zealand and Yorkshire vowel categories is to represent the vowels in the two dialects in terms of the distribution of formant values in appropriately large and varied databases (cf. Hindle 1979; Labov 1994; Watson, Harrington and Evans 1998; Docherty and Foulkes 1999). The question then reduces to the mechanics of how to tag the databases. Vowel formant values can be extracted from a database however the vowels are tagged, so long as they are tagged consistently within each database. That is, it does not matter whether the tags are [ɪ], [e], and [ɛ] for New Zealand versus [i̽], [ɛ], and [a̲] for Yorkshire, or 'i', 'e' and 'a' (or 'I', 'E' and '@', or 'ih', 'eh', and 'ae') for both. What matters is that all of the instances of any one of these three historically short front vowels be tagged in the same way within a given database and differentiated from the historically long vowel counterparts, so that researchers can retrieve all instances of the different categories and compare their formant values between the two varieties.

Of course, this characterisation of the ToBI approach begs the question of how one decides that two vowels belong to the same or to different categories within any one variety, or how vowel categories might correspond across two different varieties. In the comparison between New Zealand and Yorkshire, we know the correspondences because these three vowels are phonemes which contrast a large number of words that are common to both dialects, and because a substantial body of philological research shows that (especially by comparison to historically back vowels) these short front vowel phonemes developed in a very uniform way across the lexicon in the dialects of English that were married in making the New Zealand variety. Lexical tone categories and their correspondences in cognate words across varieties of Mandarin are somewhat messier, but still fairly easy to establish by comparison to the tones of English, where the categories typically do not contrast sets of morphemes but constitute pragmatic morphemes in their own right (see, e.g., Ladd 1980; Gussenhoven 1984; Ward and Hirschberg 1985; Pierrehumbert and Hirschberg 1990). However, a large body of research using a variety of techniques has given us a good idea of what the tonal categories are for SBE and MAE, and there is a large body of research on intonation systems of other languages that can be tapped as a source of ideas about what questions to ask in deciding what the categories are for an unstudied (or under-studied) language variety.

In looking at a relatively unstudied variety of English, for example, an obvious first set of questions to ask would be: 'What is the history of this variety? Was there contact with a substrate language that might lead us to expect it to lack such MAE and SBE categories as stress and nuclear pitch accent? If so, how can we characterise patterns that seem to correspond to these categories in MAE and SBE?' These are especially valid questions to ask for varieties such as Hawaiian English, Singapore English, or West African English (see, e.g., Vanderslice and Pearson 1967; Lim 1997; Gut 2000); it would be a mistake to begin one's analysis of these varieties by having MAE or SBE speakers tag the syllables they perceive to be stressed. On the other hand, if the understudied variety clearly is related to MAE and SBE with respect to the applicability of the notions of stress and accent, then one can proceed differently. The researcher who is a speaker of SBE or MAE might enlist a native speaker of the other variety to collaborate in asking questions such as the following. 'Do we consistently hear nuclear accent as falling in the same place in utterances elicited in the same controlled contexts in both of our varieties? If so, what is the F0 pattern around the nuclear-accented syllable in a broad focus SVO declarative utterance in the other variety? Is the F0 pattern (and the syntax) the same in a context that puts narrow focus on the object NP? What about on the subject NP, or the verb? Can we elicit subject narrow focus in longer declarative sentences, to see what happens when there are few versus many words following the nuclear accent? What is the F0 pattern around the nuclear accented syllable in the inverted broad focus yes-no question that is the counterpart to the short SVO statement and which does not presuppose its answer? If the F0 contour in this case is a rise from the nuclear syllable, do we see the same rising pattern in a yes-no question when the speaker expects a 'yes' answer, or does the rise begin at a higher level? What happens if the yes-no question focuses on the subject NP or on the verb?'

Some of these questions can be addressed also by looking at a suitable corpus that one eventually intends to tag in a ToBI framework system, particularly if one can collaborate with a native speaker consultant, as in Daly and Warren (in press). Developing elicitation protocols for recording suitable corpora, such as the Map task (Anderson et al. 1991), is thus another important research endeavor in its own right. Whatever the materials, however, the hard slogging work of addressing such questions cannot be circumvented by trying to force the analysis of tunes in the other variety into a transcription system designed for SBE and MAE.

Researchers approaching a previously undescribed variety should not try to use the original MAE_ToBI system (or Crystal's system, or Halliday's system, or any other transcription system designed for SBE or MAE) as if it were a variety-neutral phonetic transcription system for intonation and prosody. Couching the research program within a transcription framework that has been used to describe these better-studied varieties can help formulate relevant questions for the initial analysis and for later comparison across the varieties, but it cannot do more than that. A transcription system and the framework for developing it are two separate things.

In sum, applying the ToBI framework to develop a transcription system for a new variety presumes that the development team has access to an established body of research specific to the variety for which the transcription system is intended. If there is no such body of research, the development team must do the necessary research. Knowing a great deal about several other varieties of the same language, and being able to state that knowledge in a common framework, can help to establish relevant controls from early in the research endeavor. However, there is no shortcut around actually doing the research. A ToBI framework system devised for one language variety cannot be assumed to be applicable even to other varieties of the same language, without first establishing appropriate intonational and prosodic analyses for each variety. Any claim that the symbolic tags are comparable across varieties must be based on a thorough variety-specific analysis of each of the varieties involved, as in Fletcher and Warren's (2000) study of F0 contours for high rising terminals in both Australasian varieties of English. This is so because the Tones-tier labels in a ToBI framework system are comparable to a broad phonemic representation of consonants and vowels, and not to a narrow phonetic one.

2.7. The work ahead for MAE_ToBI

That said, we can think of several useful ends that a more 'allophonic' transcription might serve. One is to record features of productions in regions where dialect contact has established a 'mixed code' in which the different distributions of phonetic values for a common phonological category come to have a kind of distinctive status as badges of contrasting social affiliation. As more and more utterances of Mainstream American English are transcribed with MAE_ToBI, it is likely that we will discover systematic differences among different communities. We would be surprised if no one finds mixed codes incorporating, say, features of the African-American Vernacular English intonation system into a predominately MAE utterance for stylistic effect (cf. Hay, Jannedy and Mendoza-Denton 1999). It is our hope that a **Code** tier (as proposed in the Pan-Mandarin ToBI system — see Peng et al. this volume) will provide the right approach for capturing such phenomena, but further experience with the MAE_ToBI system in a variety of contexts will be necessary to test its appropriateness.

A number of other issues might be illuminated by an allophonic transcription of consonant and vowel segments, of the sort that Veillieux and Shattuck-Hufnagel (1998) and Jurafsky, Bell and Girand (in press) are already using to study the prosody of function words. The question of whether to add a Stress tier to MAE_ToBI (as in the proposed Pan-Mandarin system) probably is one that will require a closer study of the consonant and vowel qualities that are associated with stressed syllables at different levels of the prominence hierarchy. Some instances of low inter-transcriber agreement regarding the presence or absence of a pitch accent suggest that speakers can use segmental rhythms to create the sense of accentual prominence in the absence of any tonal markings of pitch accent. Some instances of break index 2 (as in the utterance in Figure 2.2) similarly suggest that speakers can use the rhythms of intermediate phrasing to set off words as focally prominent without ending the current phrase by pronouncing a phrase accent. In general, MAE_ToBI is vague about the segmental effects relevant for

differentiating break index levels when there are no tonal marks. Wightman et al. (1992) and others have examined durational correlates of juncture, particularly at the higher break index levels. Pierrehumbert and Talkin (1992) and Dilley, Shattuck-Hufnagel and Ostendorf (1995), similarly, have shown that vowels and sonorant consonants tend to have more or less glottalised variants when they occur at the beginning of an intermediate phrase or an intonational phrase, particularly if the initial syllable also is accented. However, tagging of break index level 0 (e.g., prosodic-word internal, as in *gimme*, *doncha*) versus break index level 1 (normal word boundary, as in *give me*, *don't you*) tends to be less consistent across teams of experienced MAE_ToBI transcribers, because we do not have a good understanding of the phonetic bases for break index 0 in cases where the segmental correlates are less obvious than in the almost lexicalised examples given here (see Syrdal et al., 2001). A good study of where tap variants of /t/ and /d/ occur in MAE and the Australasian varieties would help clarify whether the MAE *ToBI Annotation Conventions* were too sanguine in citing this as a straightforward clue to the creation of a constituent smaller than a "normal" word in these varieties.

Of course, these issues will not be addressed adequately just by adding a more allophonic tagging of consonants and vowels in the relevant cases; research aimed at finding a quantitative phonetic representation also is needed. The F0 contour is relatively easy to calculate as a phonetic representation of pitch, but other phonetic properties, such as degree of glottalisation (or creak) and degree of breathiness are not as well studied. The field would be well served by the development of phonetic representations of these voice qualities that can be applied to continuous speech, and not just to sustained vowels. We are encouraged about this direction of research by the recent increase in attention to the problem, with closer interaction among researchers who study voice quality from a variety of viewpoints, including its role in phonological contrast, its systematic variation with prosodic structure, and its range and mechanisms of variation in pathology (cf. Shattuck-Hufnagel, Kreiman and Gerratt, in press.) Other important efforts involve the attempt to develop classification systems for laryngealisation events in continuous speech (Batliner, Burger, Johne and Kiessling 1993), and to characterise their distribution (Kohler 1994; Hagen 1997). Work on accent in Dutch (e.g., Sluijter and van Heuven 1996) suggests to us that phonetic measures of voice quality could illuminate some of the cases of perceived phrase-level prominence in the absence of pitch accent. Such measures clearly are needed also to represent some intonational contrasts, such as the two different interpretations (incredulity versus uncertainty) of the rise-fall-rise tune (Hirschberg and Ward 1992). Current acoustic phonetic representations of timing (e.g., Campbell and Isard 1991) and 'rhythmicity' (e.g., Ramus, Nespor and Mehler 1999; Grabe and Low, in press) also seem quite crude by comparison to our understanding of articulatory dynamics (e.g., Browman and Goldstein 1990; Munhall, Kawato and Vatikotis-Bateson 2000; Beckman and Cohen 2000). Basic research to devise acoustic measures of timing phenomena that are as good as our phonetic representation of pitch would be useful. For example, better phonetic representations of timing and rhythmicity should illuminate our understanding of effects such as the 'chanted' or 'stylised' variant of H* !H- L%, which constitutes the 'calling contour' (Ladd 1980). Better measures of timing and of voice quality also might increase inter-transcriber reliability for break index 3 versus break index 4 in the tonally ambiguous cases of L- versus L-L% and H- vs. H-L%.

As the above suggests, we think that comparing points of greater and lesser inter-labeller reliability is useful for suggesting avenues of necessary further research. This means that the ToBI endeavor could also benefit from more work on inter-transcriber consistency, and from the development of other metrics for establishing correspondences between the criteria that different transcribers use to distinguish categories that they perceive as more or less similar (see McGory,

Herman and Syrdal 1999). In the same vein, it would be useful to establish the correspondences between MAE_ToBI transcriptions and transcriptions made using other prosody annotation schema. Comparing points of high 'inter-system reliability' with points where correspondences are more difficult to establish could illuminate areas where the underlying models of intonational phonology need work. For example, the IViE 'Phonological tier' labels, like Gussenhoven's (1984) model of SBE and all of the traditional British systems on which it is based, differs from MAE_ToBI in allowing no leading tones, such as the leading L tone that distinguishes MAE_ToBI L+H* from H*, or the leading H tone that distinguishes MAE_ToBI H+!H* (= Pierrehumbert's H+L*) from !H* or L*. Since IViE is based on British systems that are designed for SBE, we might ask whether this difference between the transcription systems can be attributed to a more substantive difference between the two dialects. If this is the case, Grice (1995) shows that inter-dialect differences cannot be the sole explanation, since SBE also clearly has an accent category with a leading tone falling onto a lower tone (which Grice transcribes as a separate H+L* pitch accent type). The MAE_ToBI distinction between H* and L+H*, on the other hand, is more controversial. Work by Ladd and colleagues (Ladd 1993; Ladd and Morton 1997) suggests that, in SBE at least, the H* versus L+H* contrast might be a gradient difference in prominence associated with 'normal' versus 'expanded' pitch ranges rather than a strictly categorical binary distinction. The contrast also causes more inter-transcriber disagreement in transcribing MAE than any other accent pair (see Silverman et al. 1992; Pitrelli, Beckman and Hirschberg 1994; Syrdal et al. 2001). In these two cases of accents with leading tones, comparing across the MAE_ToBI and IViE transcription systems augments the comparison across transcribers to highlight places where both systems might be revised. Establishing correspondences across transcription schemes also has a more practical utility. Implementing known correspondences into automatic translation algorithms would expand the repertoire of databases available to all researchers, whatever their theoretical orientations.

In short, as members of the original development team, we would characterise the original ToBI system as we would characterise the ToBI framework as a whole: it is an ongoing research program rather than a set of 'rules' cast in stone for all time. Despite its 'unfinished' nature, however, the effort that went into its development clearly has been worthwhile. In addition to what MAE_ToBI has taught us about the process of transcription system development, a number of results from using the system signal its value. These results include the overall high level of transcriber agreement, the system's productiveness in encouraging and guiding the development of similar systems for other languages, and its usefulness as a communal corpus creation tool even in its current state. There are now many MAE_ToBI corpora, and a good number of these are publicly available. For example, a large part of the BU FM Radionews database has been annotated using MAE_ToBI by teams at Boston University and MIT, and these tags been used in a large variety of research projects ranging from the training of an intonation recogniser and an intonation synthesis system (Ostendorf and Ross 1997; Ostendorf and Ross 1999) to the definitive study of stress shift as early accent placement (Shattuck-Hufnagel, Ostendorf and Ross 1994). Large parts of the native speaker Map Task dialogs in the Australian National Database of Spoken Language (ANDOSL) have been annotated with MAE_ToBI labels by teams at Macquarie University and University of Melbourne, and this database also has been used for a variety of purposes such as training duration rules for an Australian English TTS system (Fletcher and McVeigh 1993) and identifying spontaneous speech materials for use in psycholinguistics experiments testing the role of accent in pronoun resolution (Stirling, Fletcher, Mushin and Wales 2001). A large portion of the MAGIC corpus at Columbia University has been MAE_ToBI labelled and used

for exploring the relationship of various syntactic, semantic, and discourse features to intonational features such as accent and phrasing, including a test of Bolinger's idea of semantic weight as a determinant of accentuation (Pan and McKeown 1999).

Thus, the vision which inspired Victor Zue to convene the initial workshop, and the willingness of the various sets of participants to persevere through negotiations whose intensity would make Kofi Annan shudder, has resulted in the creation of large labelled databases, a better understanding of the strengths and weaknesses of the underlying phonological theories, the development of ToBI-like systems for other languages, and even in the development of contrasting systems inspired by the concreteness and explicit assumptions of the ToBI approach. We look forward to further developments in our understanding of spoken prosody, such as the ToBI-framework systems described in this book.

References

Ainsworth, H. (2000), 'Telling Tales in Taranaki: Evidence of Regional Variation in New Zealand English.' Paper presented at the Workshop on Varieties of English intonation and prosody, Victoria University of Wellington, 12-14 December 2000.

Allen, J. and Core, M. (1997), *DAMSL: Dialog Act Markup in Several Layers*. Online ms available at http://www.cs.rochester.edu/research/trains/annotation/RevisedManual.

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S., and Weinert, R. (1991), 'The HCRC Map Task Corpus'. *Language and Speech*, 34: 351-366.

Anderson, S. R. (1979), 'Tone Features', in V. A. Fromkin (ed.), *Tone: A Linguistic Survey* (New York: Academic Press), 133-176.

Armstrong, L. E. and Ward, I. C. (1926). *A Handbook of English Intonation* (Cambridge, UK: Heffer and Sons).

Arvaniti, A., & Baltazani, M. (this volume), 'Intonational Analysis and Prosodic Annotation of Greek Spoken Corpora'.

Ayers, G. M. (1994), 'Discourse Functions of Pitch Range in Spontaneous and Read Speech', *Ohio State University Working Papers in Linguistics*, 44: 1-49.

Batliner, A., Burger, S., Johne, B., and Kiessling, A. (1993), 'MUESLI: A Classification Scheme for Laryngealizations', in D. House and P. Touati (eds.), *Proceedings of an ESCA Workshop on Prosody, Lund 1993. (Working Papers in Linguistics & Phonetics, No. ?).* Lund University. ROUND OUT.

Beckman, M. E. (1986), *Stress and Non-Stress Accent* (Dordrecht: Foris).

Beckman, M. E. and Ayers, G. M. (1994), *Guidelines for ToBI Labelling*. Online MS and accompanying files. Available at http://www.ling.ohio-state.edu/phonetics/E_ToBI.

Beckman, M. E. and Cohen, K. B. (2000), 'Modeling the Articulatory Dynamics of Two Levels of Stress Contrast', in M. Horne (ed.), *Prosody: Theory and Experiment. Studies Presented to Gösta Bruce* (Dordrecht: Kluwer), 169-200.

Beckman, M. E. and Hirschberg, J. (1994), *The ToBI Annotation Conventions.* Online MS. Available at http://www.ling.ohio-state.edu/~tobi/ame_tobi/annotation_conventions.html.

Beckman, M. E. and Pierrehumbert, J. B. (1986), 'Intonational Structure in Japanese and English', *Phonology Yearbook*, 3: 255-309.

Bolinger, D. L. (1951), 'Intonation: Levels versus Configurations', *Word*, 7: 199-210.

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

Bolinger, D. L. (1964), 'Around the Edge of Language: Intonation', *Harvard Educational Review*, 34: 282-293.

Browman, C. P. and Goldstein, L. (1990), 'Tiers in Articulatory Phonology, With Some Implications for Casual Speech', in J. Kingston and M. Beckman (eds.), *Papers in Laboratory Phonology I* (Cambridge: Cambridge University Press), 341-376.

Bruce, G. (1977), *Swedish Word Accents in Sentence Perspective* (Lund: Lund University).

Bruce, G. (1982), 'Developing the Swedish Intonation Model. *Working Papers in Phonetics, Lund University* [CHECK WP NO], 22: 51-117.

Campbell, W. N. and Isard, S. D. (1991), 'Segment Durations in a Syllable Frame', *Journal of Phonetics*, 19: 37-48.

Carletta, J. , Isard, A. , Isard, S. , Kowtko, J. and Doherty-Sneddon, G. (1996), *HCRC Dialogue Structure Coding Manual. Technical Report HCRC/TR-82*. Available online at http://www.hcrc.ed.ac.uk/publications/tr-82.ps.gz.

Cedergren, H. J. and Perrault, H. (1994), 'Speech Rate and Syllable Timing in Spontaneous Speech', *Proceedings of the 1994 International Conference on Spoken Language Processing*, 1087-1090.

Chao, Y. R. (1920), 'A System of Tone Letters', *Le maître phonétique*, 45: 24-27.

Charniak, E. (2000). 'A Maximum-Entropy-Inspired Parser', *Proceedings of the ANLP-NAACL 2000*, 132-139.

Coleman, J. (1996), 'Declarative Syllabification in Tashlhit Berber', in J. Durand and B. Laks (eds.), *Current Trends in Phonology: Models and Methods* (Salford: European Studies Research Institute, University of Salford), vol. 1, 177-218.

Collins, M. (1999), 'Head-Driven Statistical Models for Natural Language Parsing'. Ph.D. dissertation (University of Pennsylvania).

Cooper, W. E. (1976), 'Syntactic Control of Timing in Speech Production: A Study of Complement Clauses', *Journal of Phonetics*, 4: 151-171.

Crystal, D. (1969), *Prosodic Systems and Intonation in English* (Cambridge, UK: Cambridge University Press).

Daly, N. and Warren, P. (in press), 'Pitching it Differently in New Zealand English: Some Gender Differences in Intonation Patterns', *Journal of Sociolinguistics*.

Dell, F. and Elmedlaoui, M. (1996), 'Nonsyllabic Transitional Cocoid in Imdlawn Tashlhiyt', in J. Durand and B. Laks (eds.), *Current Trends in Phonology: Models and Methods* (Salford: European Studies Research Institute, University of Salford), vol. 1, 217-244.

Dilley, L., Shattuck-Hufnagel, S. and Ostendorf, M. (1995) 'Individual differences in the glottalization of vowel-initial syllables', *Journal of the Acoustical Society of America*, 97: 3418-3419.

Docherty, G. J. and Foulkes, P. (1999), 'Instrumental Phonetics and Phonological Variation: Case Studies from Newcastle upon Tyne and Derby', in P. Foulkes and G. J. Docherty (eds.), *Urban Voices: Accent Studies in the British Isles* (London: Arnold), 47-71.

Fletcher, J. and Bishop., J. (this volume), 'Intonation in Six Dialects of Bininj Gun-Wok'.

Fletcher, J., Grabe, E. and Warren, P. (this volume), 'Intonational Variation in Four Dialects of English: The High Rising Tune'.

Fletcher, J. and Harrington, J. (1996), 'Timing of Intonational Events in Australian English', *Proceedings of the Sixth Australian International Conference on Speech Science and Speech Technology*, 611-615.

Fletcher, J. and McVeigh, A. (1993), 'Syllable and Segment Duration in Australian English', *Speech Communication*, 13: 355-365.

Fletcher, J. and Warren, P. (2000), 'Variation in Rises and Rises in Varieties', Paper presented at the Workshop on Varieties of English intonation and prosody (Victoria University of Wellington, 12-14 December 2000).

Gee, J. P. and Grosjean, F. (1983), 'Performance Structures: A Psycholinguistic and Linguistic Appraisal', *Cognitive Psychology*, 14: 411-458.

Grabe, E. (2000), *The IViE Labelling Guide, Version 2*. Online MS available at http://www.mml.cam.ac.uk/ling/ivyweb/guide.html.

Grabe, E. and Low, E. L. (in press), 'Acoustic Correlates of Rhythm Classes,' in C. Gussenhoven and N. Warner (eds.), *Papers in Laboratory Phonology VII* (Berlin: Mouton de Gruyter).

Grabe, E., Nolan, F. and Farrar, K. (1998), 'IViE — a Comparative Transcription System for Intonational Variation in English', *Proceedings of the 5th International Conference on Spoken Language Processing.* Sydney: Australian Speech Science and Technology Association (CDROM). Grice, M. (1995), 'Leading Tones and Downstep in English', *Phonology*, 12: 183-233.

Grice, M., Reyelt, M., Benzmüller, R., Mayer, J. and Batliner, A. (1996), 'Consistency in Transcription and Labelling of German Intonation with GToBI', *Proceedings of the 1996 International Conference on Spoken Language Processing* (New Castle, Delaware : Citation Delaware), 1716-1719.

Grice, M., Ladd, D. R. and Arvaniti, A. (2000), 'On the Place of Phrase Accents in Intonational Phonology', *Phonology*, 17: 145-187.

Grosz, B. and Hirschberg, J. (1992), 'Some Intonational Characteristics of Discourse Structure', *Proceedings of the 1992 International Conference on Spoken Language Processing* (Banff: University of Alberta), 429-432.

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

Gussenhoven, C. (1984), *On the Grammar and Semantics of Sentence Accents* (Dordrecht: Foris).

Gussenhoven, C. (1990), 'Tonal Association Domains and the Prosodic Hierarchy in English', in S. Ramsaran (ed.), *Studies in the Pronunciation of English* (London: Routledge), PP?.

Gut, U. (2000), Session on West African Englishes. Workshop on Varieties of English intonation and prosody, Victoria University of Wellington, 12-14 December 2000.

Hagen, A. (1997), 'Linguistic Functions of Glottalizations and their Language Specific use in English and German', M.A. thesis (Erlangen University).

Hay, J., Jannedy, S. and Mendoza-Denton, N. (1999), 'Oprah and /ay/: Lexical Frequency, Referee Design and Style', *Proceedings of the 14th International Congress of Phonetic Sciences* (CD-Rom distributed by the Regents of the University of California).

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

Hindle, D. M. (1979), 'The Social and Situational Conditioning of Phonetic Variation', Ph.D. dissertation (University of Pennsylvania).

Hirschberg, J. (1993), 'Pitch Accent in Context: Predicting Intonational Prominence from Text', *Artificial Intelligence*, 63: 305-340.

Hirschberg, J. and Rambow, O. (2001), 'Learning Prosodic Features using a Tree Representation', *Proceedings of Eurospeech 2001* (Alborg: Center for Personkommunication).

Hirschberg, J. and Nakatani, C. (1998), Acoustic indicators of topic segmentation. *Proceedings of the 1998 International Conference on Spoken Language Processing*, Sydney (Distributed by the Australian Speech Science and Technology Association on CDROM).

Hirschberg, J. and Prieto, P. (1994), 'Training Intonational Phrasing Rules Automatically for English and Spanish Text-to-Speech', *Proceedings of the Second ESCA/IEEE Workshop on Speech Synthesis*, 159-162.

Hirschberg, J. and Ward, G. (1991), 'The Influence of Pitch Range, Duration, Amplitude, and Spectral Features on the Interpretation of L*+H L H%', *Journal of Phonetics*, 20: 241-251.

Hirst, D. (1999), 'Intonation in British English', in D. Hirst and A. Di Cristo (eds.), *Intonation Systems* (Cambridge, UK: Cambridge University Press), 56-77.

Hirst, D. and Di Cristo, A. (1999), 'A Survey of Intonation Systems', in D. Hirst and A. Di Cristo (eds.), *Intonation Systems* (Cambridge, UK: Cambridge University Press), 1-44.

Hirst, D. J., Di Cristo, A., Le Besnerais, M., Najim, Z., Nicolas, P. and Roméas, P. (1993), 'Multi-Lingual Modelling of Intonation Patterns, *Proceedings of the ESCA Workshop on Prosody (Lund Working Papers in Linguistics and Phonetics, 41),* 204-207.

Hualde, J. I., Elordieta, G., Inaki, G. and Smiljanić, R. (in press). 'From Pitch Accent to Stress Accent in Basque and the Typology of Accentual Systems', in C. Gussenhoven and N. Warner (eds.), *Papers in Laboratory Phonology VII* (Berlin: Mouton de Gruyter).

Iivonen, A. (1999), 'Intonation in Finnish', in D. Hirst and A. Di Cristo (eds.), *Intonation Systems* (Cambridge, UK: Cambridge University Press), 311-327.

Jun, S.-A. (this volume, ch. 9), 'Korean Intonation and Prosodic Transcription'.

Jun, S.-A. (this volume, ch. 17), 'Prosodic Typology'.

Jurafsky, D., Bell, A. and Girand, C. (in press), 'Phonological Variation as Evidence for Lexical Representation of Homonyms', in C. Gussenhoven and N. Warner (eds.), *Papers in Laboratory Phonology VII* (Berlin: Mouton de Gruyter).

Kingdon, R. (1939), 'Tonetic Stress Markers for English', *Maître Phonétique*, 54: 60-64.

Koehn, P., Abney, S. Hirschberg, J. and Collins, M. (2000), 'Improving Intonational Phrasing with Syntactic Information', *Proceedings of ICASSP 2000* (Istanbul).

Kohler, K.J., (1994), 'Glottal Stops and Glottalization in German', *Phonetica*, 51: 38-51.

Labov, W. (1994), *Principles of Linguistic Change* (Cambridge, MA: Blackwell).

Ladd, D. R. (1980), *The Structure of Intonational Meaning: Evidence from English* (Bloomington, IN: Indiana University Press).

Ladd, D. R. (1983), 'Phonological Features of Intonational Peaks', *Language*, 59: 721-759.

Ladd, D. R. (1990), 'The Metrical Representation of Pitch Register', in J. Kingston and M. Beckman (eds.), *Papers in Laboratory Phonology I* (Cambridge, UK: Cambridge University Press), 35-57.

Ladd, D. R. (1993), 'Constraints on the Gradient Variability of Pitch Range (or) Pitch Level 4 Lives!' on P. Keating (ed.), *Papers in Laboratory Phonology III* (Cambridge, UK: Cambridge University Press), 43-63.

Ladd, D. R. (1996), *Intonational Phonology* (Cambridge, UK: Cambridge University Press).

Ladd, D. R. and Morton, R. (1997), 'The Perception of Intonational Emphasis: Continuous or Categorical?' *Journal of Phonetics*, 25: 313-342.

Leben, W. (1973), 'Suprasegmental Phonology'. Ph.D. dissertation (Massachusetts Institute of Technology).

Lehiste, I. (1960), *An Acoustic-Phonetic Study of Internal Open Juncture (Phonetica Supplement)*.

Lehiste, I. (1975), 'The Phonetic Structure of Paragraphs', in A. Cohen and S. Nooteboom (eds.), *Structure and Process in Speech Perception* (Berlin: Springer-Verlag), 195-206.

Liberman, M. and Pierrehumbert, J. (1984), 'Intonational Invariance under Changes in Pitch Range and Length', in M. Aronoff and R. Oehrle (eds.), *Language Sound Structure* (Cambridge, MA: MIT Press), 157-233.

Lim, L. (1997), 'Intonation Patterns Characterising Three Ethnic Varieties of English in Singapore: Observations and Implications, *Proceedings of the ESCA Workshop on Intonation: Theory, Models and Applications, Athens, Greece, 18-20 September 1997*, 207-210.

MacDonald, M. C., Pearlmutter, N. J. and Seidenberg, M. S. (1994), 'The Lexical Nature of Syntactic Ambiguity Resolution', *Psychological Review*, 101: 676-703.

Marcus, M. P., Santorini, B. and Marcinkiewicz, M. A. (1993), 'Building a Large Annotated Corpus of English: The Penn Treebank', *Computational Linguistics*, 19: 313-330.

Mayo, C., Aylett, M. and Ladd, D. R. (1997), 'Prosodic Transcription of Glasgow English: An Evaluation Study of GlaToBI', *Proceedings of the ESCA Workshop on Intonation: Theory, Models and Applications, Athens, Greece, 18-20 September 1997*, 231-234.

McCawley, J. D. (1968), *The Phonological Component of a Grammar of Japanese* (The Hague: Mouton).

McGory, J. T. (2000), 'Linguistics 795T: Practicum in English Intonation', course materials (Ohio State University).

McGory, J. T., Herman, R. and Syrdal, A. (1999), 'Using Tone Similarity Judgments in Tests of Intertranscriber Reliability', *Journal of the Acoustical Society of America*, 106: 2242.

Munhall, K. G., Kawato, M. and Vatikiotis-Bateson, E. (2000), 'Coarticulation and Physical Models of Speech Production', in M. Broe and J. Pierrehumbert (eds.), *Papers in Laboratory Phonology V* (Cambridge, UK: Cambridge University Press), 9-39.

Nakatani, C. H., Groz, B. J., Ahn, D. D. and Hirschberg, J. (1995), *Instructions for Annotating Discourse. Technical Report Number TR-21-95* (Cambridge, MA: Center for Research in Computing Technology), available online at ftp://ftp.pitt.edu/dept/lrdc/edtech/jmoore/emd/nakatani-etal-guide.ps.Z

Nakatani, L., O'Conner, K. and Aston, C. (1981), 'Prosodic Aspects of American English Speech Rhythms', *Phonetica*: 38: 84-106

Nespor, M. and Vogel, I. (1986), *Prosodic Phonology* (Dordrecht: Foris).

Nolan, F. and Grabe, E. (1997), 'Can "ToBI" Transcribe Intonational Variation in British English?' *Proceedings of the ESCA Workshop on Intonation: Theory, Models and Applications, Athens, Greece, 18-20 September 1997*, 259-262.

O'Malley, M. M., Kloker, D. and Dara-Abrams, B. (1973), 'Recovering Parentheses from Spoken Algebraic Expressions', *IEEE Transactions in Audio and Electroacoustics*, AU-21: 217-220.

Ostendorf, M. and Ross, K. (1997), 'A Multi-Level Model for Recognition of Intonation Labels', in Y. Sagisaka, N. Campbell and N. Higuchi (eds.), *Computing Prosody* (New York: Springer-Verlag), 291-308.

Ostendorf, M. and Ross, K. (1999), 'A Dynamical System Model for Generating Fundamental Frequency for Speech Synthesis', *IEEE Transactions on Speech and Audio Processing*, 7: 295-309.

Ostendorf, M. and Veilleux, N. (1994), 'A Hierarchical Stochastic Model for Automatic Prediction of Prosodic Boundary Location', *Computational Linguistics*, 20: 27-54.

Palmer, H. E. (1922), *English Intonation with Systematic Exercises* (Cambridge, UK: Heffer).

Pan, S. and McKeown, K. (1999), 'Word Informativeness and Automatic Pitch Accent Modeling', *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Peng, S., Chan, M. K.-M. Tseng, C., Huang, T., Lee, O. and Beckman, M. E. (this volume). 'Towards a Pan-Mandarin prosodic annotation system'.

Pierrehumbert, J. B. (1980), 'The Phonology and Phonetics of English Intonation', Ph.D. dissertation (Massachusetts Institute of Technology).

Pierrehumbert, J. B. (2000), 'Tonal Elements and their Alignment', in M. Horne (ed.), *Prosody: Theory and Experiment. Studies Presented to Gösta Bruce* (Dordrecht: Kluwer), 11-36.

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

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

Pierrehumbert, J. and Steele, S. (1989), 'Categories of Tonal Alignment in English', *Phonetica*, 46: 181-196.

Pierrehumbert, J. and Talkin, D. (1992), 'Lenition of /h/ and Glottal Stop', in G. Docherty and D. R. Ladd (eds.), *Papers in Laboratory Phonology II* (Cambridge University Press), 90-116.

Pitrelli, J. F., Beckman, M. E. and Hirschberg, J. (1994), 'Evaluation of Prosodic Transcription Labelling Reliability in the ToBI Framework', *Proceedings of the 1994 International Conference on Spoken Language Processing* (PLACE), 123-126.

Price, P., Ostendorf, M., Shattuck-Hufnagel, S. and C. Fong, C. (1991), 'The Use of Prosody in Syntactic Disambiguation', *Journal of the Acoustic Society of America*, 90: 2956-2970.

Ramus, F., Nespor, M. and Mehler, J. (1999), 'Correlates of Linguistic Rhythm in the Speech Signal', *Cognition*, 73: 265-292.

Roach, P. (1994), 'Conversion between Prosodic Transcription Systems: "Standard British" and "ToBI"', *Speech Communication*, 15: 91-99.

Selkirk, E. O. (1978). *On Prosodic Structure and its Relation to Syntactic Structure* (Bloomington, IN: Indiana University Linguistics Club).

Shattuck-Hufnagel, S., Kreiman, J. and Gerratt, B. (eds.) (in press), Special issue of the *Journal of Phonetics* on voice quality.

Shattuck-Hufnagel, S., Ostendorf, M. and Ross, K. (1994), 'Stress Shift and Early Pitch Accent Placement in Lexical Items in American English', *Journal of Phonetics*, 22: 357-388.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. and Hirschberg, J. (1992), 'TOBI: a Standard for Labeling English Prosody', *Proceedings of the 1992 International Conference on Spoken Language Processing* (PLACE), 867-870.

Silverman, K. and Pierrehumbert, J. (1990), 'The Timing of Prenuclear High Accents in English', in J. Kingston and M. Beckman (eds.), *Papers in Laboratory Phonology I* (Cambridge, UK: Cambridge University Press), 72-106.

Sluijter, A. M. C. and Van Heuven, V. J. (1996), 'Spectral Balance as an Acoustic Correlate of Linguistic Stress', *Journal of the Acoustical Society of America*, 100: 2471-2485.

Sproat, R. (1994), 'English Noun-Phrase Accent Prediction for Text-to-Speech', *Computer Speech and Language*, 8: 79-94.

Srinivas, B. and Joshi, A. K. (1999), 'Supertagging: An Approach to Almost Parsing', *Computational Linguistics*, *25*: 237-265.

Stirling, L., Fletcher, J. Mushin, I. and Wales, R. (2001), 'Representational Issues in Annotation. Using the Australian Map Task Corpus to Relate Prosody and Discourse Structure', *Speech Communication,* 33: PP??.

Swerts, M. and Ostendorf, M. (1997), 'Prosodic and Lexical Indications of Discourse Structure in Human-Machine Interactions', *Speech Communication*, 22: 25-41.

Syrdal, A. K., Hirschberg, J., McGory, J. T. and Beckman, M. (2001), 'Automatic ToBI Prediction and Alignment to Speed Manual Labeling of Prosody', *Speech Communication*, 33: PP??

Trager, G. L. and Smith, D. L. (1951), *An Outline of English Structure* (Norman, OK: Battenburg Press).

Trim, J. K. M. (1959), 'Minor and major tone groups in English', *Le maître phonétique*, 112: 26-29.

Välimaa-Blum, R. M. (1988), 'Finnish Existential Clauses — Their Syntax, Pragmatics and Intonation'. Ph.D. dissertation (Ohio State University).

Vanderslice, R. and Pearson, L. S. (1967), 'Prosodic Features of Hawaian English', *Quarterly Journal of Speech*, 53: 156-166.

Veilleux, N. and Shattuck-Hufnagel, S. (1998), 'TITLE', *Proceedings of the 1998 International Conference on Spoken Language Processing* (CD-Rom distributed by ????).

Venditti, J. J. (1997), 'Japanese ToBI Labelling Guidelines', *Ohio State University Working Papers in Linguistics*, 50: 62-72.

Venditti, J. J. (2000), 'Discourse Structure and Attentional Salience Effects in Japanese Intonation'. Ph.D. dissertation (Ohio State University).

Venditti, J. J. (this volume), 'Sun-Ah, could you fill in the final title for this chapter'.

Ward, G. and Hirschberg, J. (1985), 'Implicating Uncertainty: The Pragmatics of Fall-Rise Intonation', *Language*, 51: 747-776.

Watson, C. I., Harrington, J. and Evans, Z. (1998), 'An Acoustic Comparison between New Zealand and Australian English Vowels', *Australian Journal of Linguistics*, 18: 185-207.

Wightman, C., Shattuck-Hufnagel, S., Ostendorf, M. and Price, P. J. (1992), 'Segmental Durations in the Vicinity of Prosodic Phrase Boundaries', *Journal of the Acoustical Society of America*, 91: 1707-1717.

Wightman, C. and Talkin, D. (1994), 'The Aligner: Text to Speech Alignment using Markov Models and a Pronunciation Dictionary', *Proceedings of the second ESCA/IEEE workshop on Speech Synthesis*.

Wong, P. W.-Y., Chan, M. K.-M., and Beckman, M. E. (this volume), 'An Autosegmental-Metrical analysis and prosodic annotation conventions for Cantonese'.

Table 2.1. The six obligatory parts of a MAE_ToBI record

audio	An audio recording of the utterance in some form.
F0	An electronic and/or paper record of the fundamental frequency contour.
Tones	An autosegmental transcription of the intonation contour; other tone related tags.
Words	An orthographic transcription of each word in the utterance, placed at the word's end, which is marked with a time index.
Break-Indices	Numeric index of the perceived degree of juncture after each orthographic word.
Misc	Markers for disfluencies, comments, and other miscellaneous events.

Table 2.2. The inventory of MAE_ToBI Tones-tier labels

basic tones:
- phrase accents: H- (!H-), L- (obligatorily placed at every BI = 3 and higher)
- boundary tones: H%, L% (obligatory at every 4)
  %H (marginal, at beginnings of some intonational phrases after pause)
- pitch accents: L*, H* (!H*), L+H* (L+!H*), L*+H (L*+!H), H+!H*

other labels:
- downstep: e.g., !H*, L+!H*, !H- (the ! diacritic marks the beginning of compressed pitch range)
- uncertainty: *?, -?, %? (uncertainty about occurrence); X*?, X-?, X%? (about tone type)
- phonetic events transcribed in careful labelling:
  < (delayed peak);
  HiF0 (maximum F0 associated with H of an accent within an intermediate phrase)
- restart: %r (see the Misc tier)

Table 2.3. The inventory of MAE_ToBI Break-Indices tier labels.

basic break index values:
- 0 (very close inter-word juncture)
- 1 (ordinary phrase-internal word end)
- 3 (intermediate phrase end, with phrase accent)
- 4 (intonational phrase end, with boundary tone)

diacritics:
- - (uncertainty) — e.g., 4- (intermediate between 3 and 4)
- p (perceived hesitation) — 1p for "cutoff", 2p and 3p for "prolongation"

tones-breaks mismatch:
- 2 (perceived 1 with unexpected tonal marker, or lengthening etc., suitable for break index 3 or 4 without the phrase accent and/or boundary tone)

Abstract The **ToBI conventions** are a consensus system for labelling spoken utterances that segregates tags for different types of phonological events and structures into parallel quasi-independent tiers. Most notably, the conventions specify a way to mark the phonologically contrastive intonational events (**Tones**) separately from the hierarchy of inter-word junctures (**Break-Indices**) with which some of these pitch events are associated. The original ToBI conventions are language-specific; they were intended to cover the phonologically contrastive tones of Mainstream American English. However, other annotation conventions based on the same general design principles have now been proposed for several other English varieties and for a number of other languages. This function of the original ToBI system as a general model for developing language-specific annotation conventions makes it possible to compare prosodic systems across languages using a common vocabulary, and to search for universals. This chapter is an overview of the original ToBI system. It reviews the design of the original system and its foundations in basic and applied research. It describes the inter-disciplinary community of users and uses for which the system was intended, and it outlines how the consensus model of American English intonation and inter-word juncture was achieved by finding points of useful intersection among the research interests and knowledge embodied in this community. It thus identifies the practical principles for designing prosodic annotation conventions that emerged in the course of developing, testing, and using this particular system.

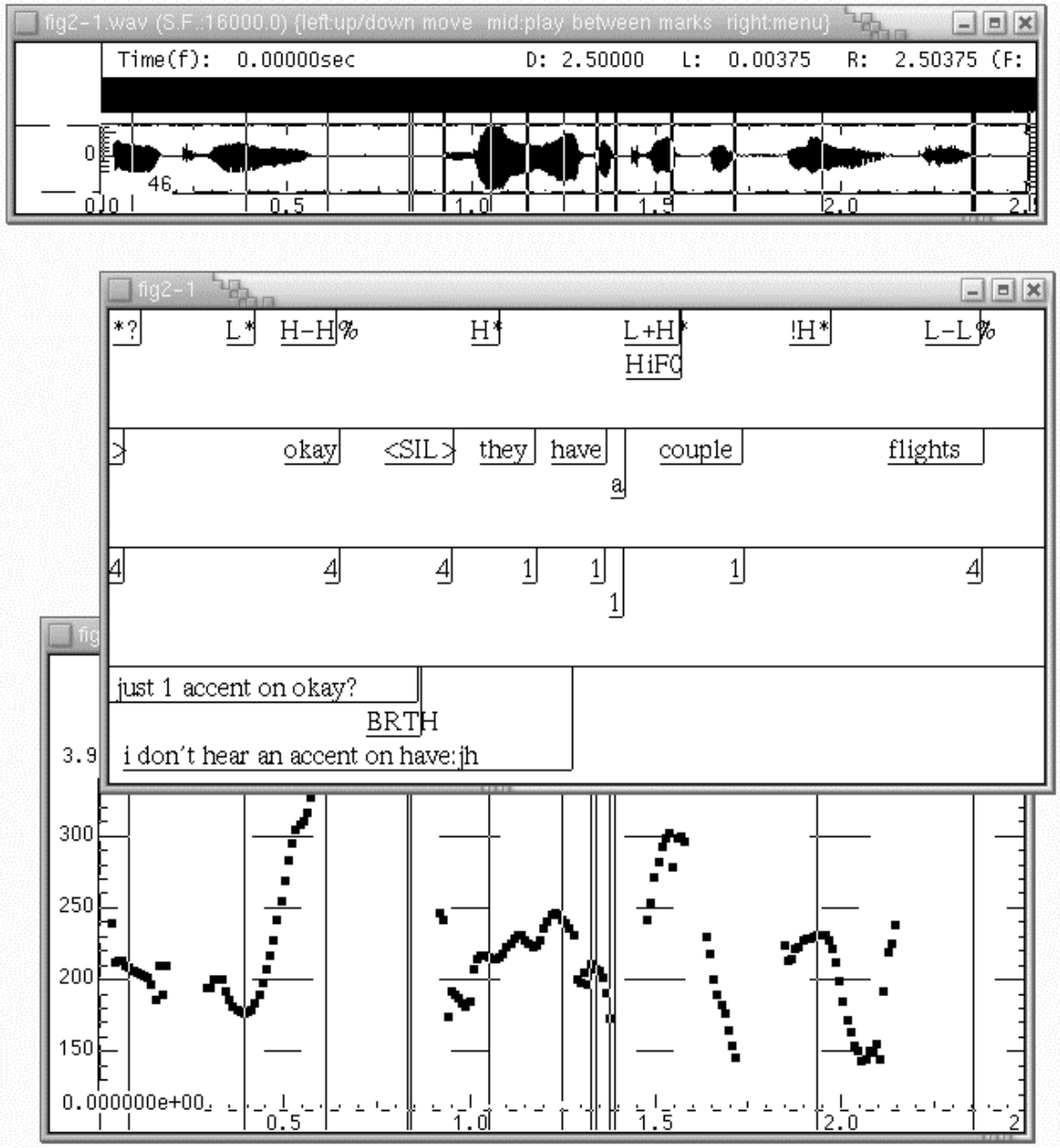

Figure 2.1. Audio waveform, F0 contour, and MAE_ToBI xlabel windows for utterance *Okay... They have a couple flights*.

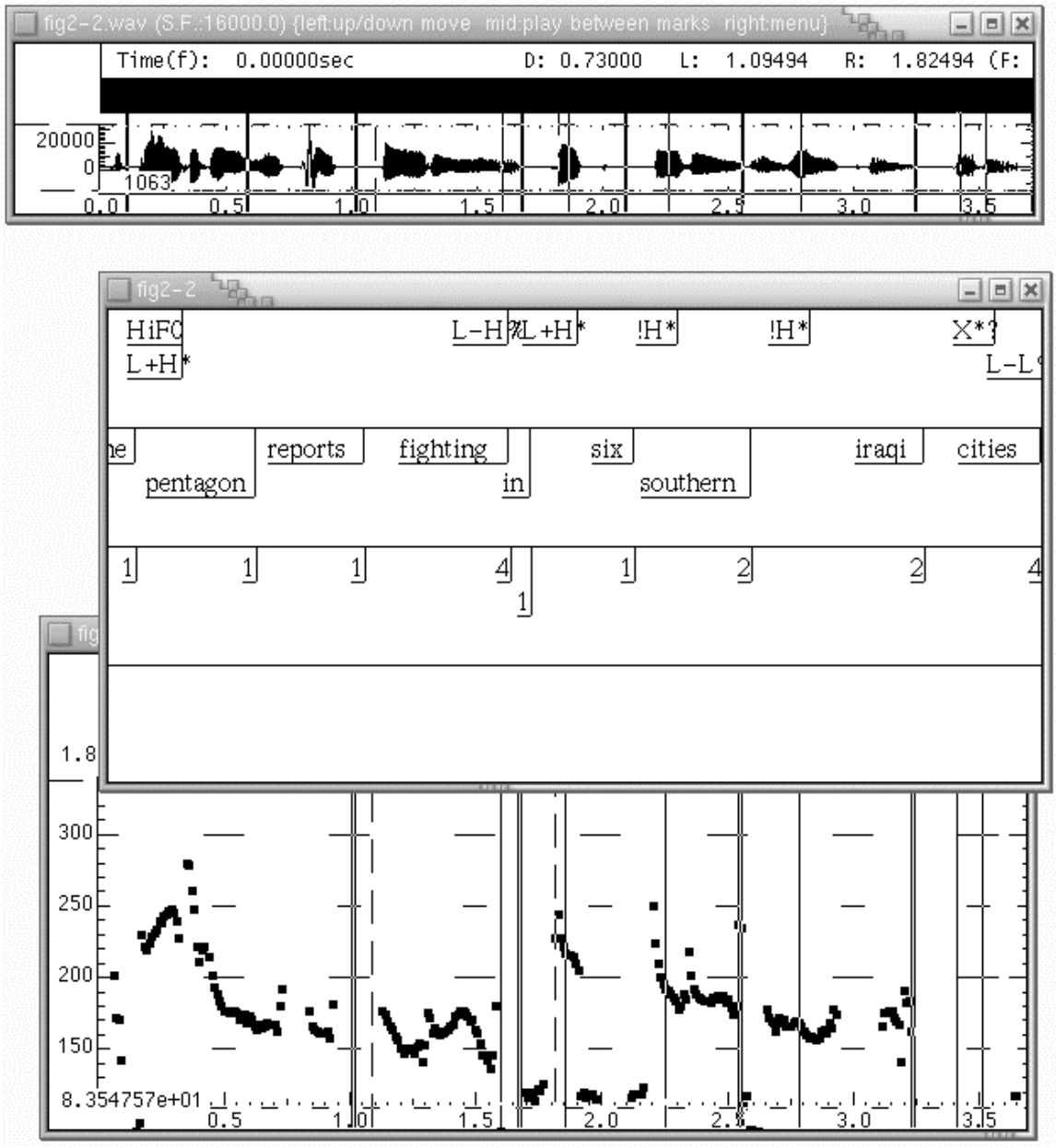

Figure 2.2. Audio waveform, F0 contour, and MAE_ToBI xlabel windows for utterance *The Pentagon reports fighting in six southern Iraqi cities*.

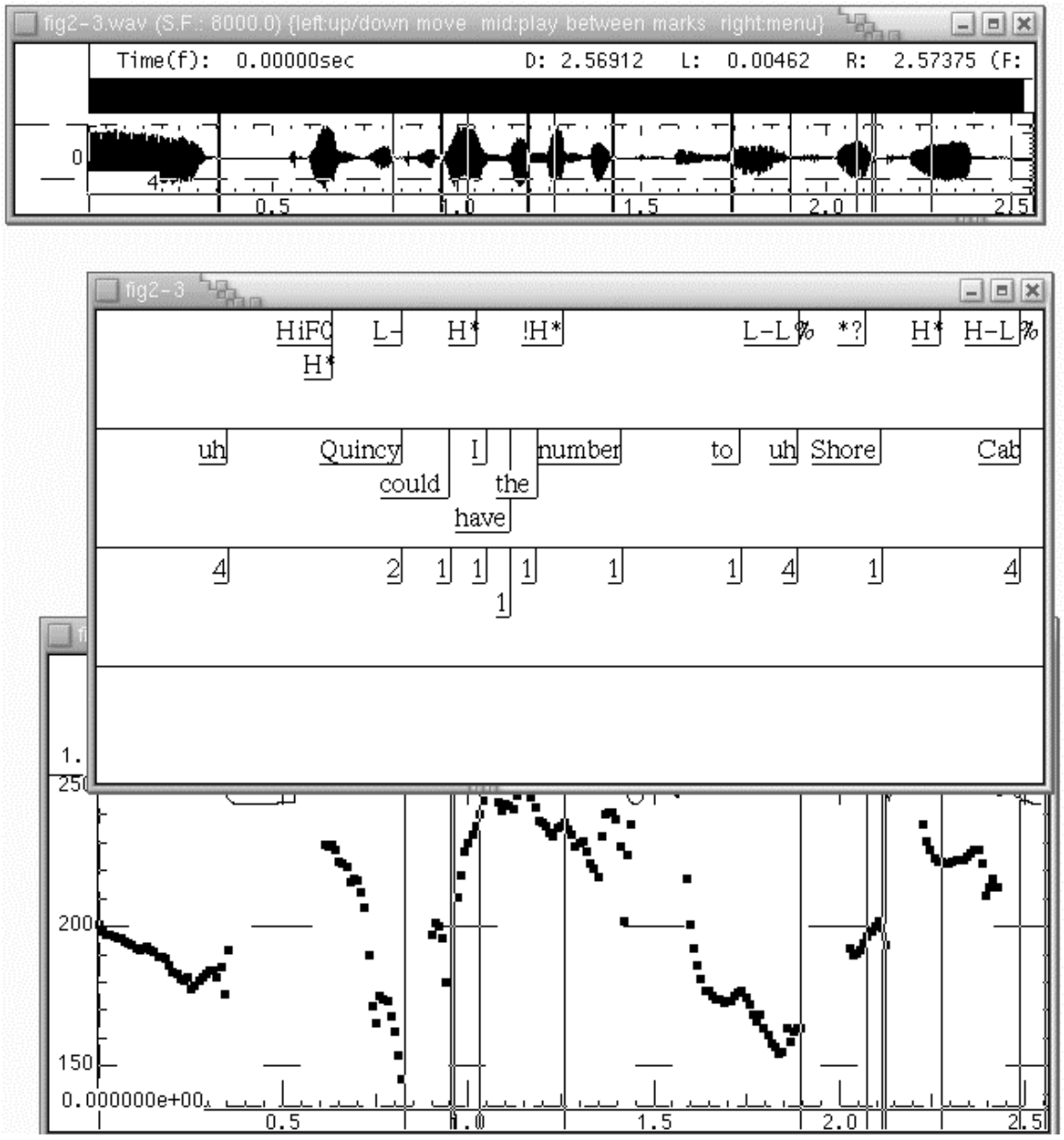

Figure 2.3. Audio waveform, F0 contour, and MAE_ToBI xlabel windows for utterance *Uhh... Quincy. Could I have the number to uh ... Shore Cab?*

Notes

[1] In accounts by British language teachers and phoneticians before the 1980s, the 'nucleus' of an intonation contour was modeled as a holistic dynamic tonal event governing the part of the contour beginning at the most stressed syllable. When this nucleus occurs far from the end of the contour, then, the pitch pattern on material after the nuclear stress is called the 'tail'. The general shape of the intonation contour over accented syllables before the nucleus is then the 'head'.

[2] Note that there are only four basic break index values, ordered from 0 to 4, with a "hole" at 2. In the original Price et al. (1991) use of break indices, the value 2 represented a perceived boundary strength intermediate between a normal word boundary and a larger phrase boundary, and was used to mark a number of imprecisely-defined phenomena. The ToBI system restricts the use of this label to an explicit subset of these phenomena — namely, inter-word junctures where there is ambiguity between a 1 and a 3 either because there is a phrase tone without the duration lengthening appropriate to a 3, or a lengthening appropriate to a 3 but no phrase tone. This means that ToBI labels do not recognise a prosodic constituent comparable to Selkirk's (1995) "Minor Phrase" unless this is equated with Beckman and Pierrehumbert's (1986) tonally marked "intermediate phrase". Labellers who postulate and perceive a constituent boundary that is larger than a "Prosodic Word" but smaller than the lowest intonationally marked constituent are encouraged to mark these events in a comments tier (see Section 2.5).

[3] The break index value '0' was intended to mark a boundary between two orthographic words which is perceived to be considerably reduced in strength from a "normal" word boundary. The MAE_ToBI conventions suggest that this sense of close grouping should be associated with such segmental sandhi phenomena as the flapping of final /t/ in utterances such as *Got a dime?*, the palatalisation of final /t/ in *We sent you the cheque*., and so on — i.e., phenomena that have been cited by phonologists as evidence of multi-word prosodic constituents such as the "Prosodic Word" or a "Clitic Group" (see Hayes 1989, Selkirk 1995, Peperkamp 1999, and the references they cite for discussion of different theoretical views of these constituents). A break index value of '1' is then a "normal" word boundary. A more precise definition of these levels is desirable, but not yet feasible, because corpus research on such phenomena as flapping and palatalisation lags considerably behind research on the phonetic correlates of prosodic grouping at the intermediate phrase and intonational phrase level.

[4] This meant omitting break indices 5 and 6 from the Price et al. (1991) model, since these two break index values could not be identified with a categorically marked level of prosodic structure such as the intonational phrase. Rather, they were intended to encode the percept of (possibly recursive) higher-level groupings above the intonational phrase.

[5] EMU is a set of tools for creating and analyzing speech databases. It includes a powerful search engine that can find segments and events based on their sequential and hierarchical contexts. For example, if a MAE spoken language database has associated word labels, and if those labels are hierarchically organised into intermediate phrases and intonation phrases, with associated MAE_ToBI labels, it is straightforward to query for every instance in the database of a word with an associated L+H* pitch accent that is also the last accent in its intermediate phrase and followed by a !H- phrase accent. The EMU readable version of the *Guidelines to ToBI Labelling* is available at http://www.shlrc.mq.edu.au/emu/emu-tobi.shtml.

---

[6] We note that no site seems to have rigorously adopted the practice envisioned by the original ToBI group of marking silences automatically, on the Misc tier.

[7] The intermediate phrase in Greek, like the intermediate phrase in English, is defined by the presence of a phrase accent after the nuclear pitch accent (see Grice, Ladd, & Arvaniti, 1999, for discussion of the cross-linguistic applicability of this concept). Thus, the use of 2 as a marker of two types of tones-breaks mismatch in English has resulted in different numbers correspond to levels that are defined in the same way in the two languages.

[8] See http://www.georgetown.edu/luperfoy/Discourse-Treebank/dri-home.html.

[9] See http://www.ling.ohio-state.edu/~tobi/ame_tobi/annotation_conventions.html for this utterance. Hirst (1999: 73) reports that the sliding head "has been described as typical of Scottish accents" and suggests that it "is probably gaining ground throughout England possibly due to the influence of American speech where the pattern is very common". Our impression is that it is more characteristic of Australian and New Zealand varieties, particularly those with a strong Scottish English substrate, than it is of mainstream American varieties — see, e.g., Fletcher and Harrington 1996; Ainsworth 2000.