The ToBI Annotation Conventions

by Julia Hirschberg and Mary E. Beckman

1 Synopsis

A ToBI transcription for an utterance consists minimally of a recording of the speech, an associated record of the fundamental frequency contour, and (the transcription proper) symbolic labels for events on the following four parallel tiers:

1. an orthographic tier
2. a tone tier
3. a break-index tier
4. a miscellaneous tier

Conventions are specified for both simple text-based transcription using this system and for $WAVES^{TM}$ label files and formats to accompany a speech file and associated time-aligned analysis records for the utterance. We first summarize the conventions assuming a computer-based labeling system such as $WAVES^{TM}$ label files and formats. A final section (Section 9), provided by Jacques Terken and Mari Ostendorf, summarizes the added guidelines for adapting the conventions to simple text-based transcription.

2 The Orthographic Tier

The orthographic tier will be used only for the transcription of orthographic words. In the $WAVES^{TM}$ label file, each word's orthographic form should be marked at the end of the final segment in the word, as determined by the labeler from the waveform or spectrogram record. That is, each orthographic word will be marked at its right 'edge'. Individual transcribers will also determine whether and how to transcribe phenomena such as filled pauses (e.g., "um","uh") and whether to use contractions (e.g., "gotta") or not. There are several existing orthographic conventions for transcribing such phenomena, which labellers may want to consult. For example, the ATIS corpus conventions specify "er", "mm", "uh", and "um" as the allowable transcriptions for filled pauses.

3 The Break Index Tier

Break indices are to be marked at the right edges of the words that have been transcribed in the orthographic tier (on or slightly to the left of each word mark), resulting in a rating for the degree of juncture perceived between each pair of words and between the final word and the silence at the end of the utterance. All junctures must be assigned an explicit break index value; there is no default juncture type.

3.1 Break Index Values

Values for the break index are chosen from the following set:

0 for cases of clear phonetic marks of clitic groups; e.g. the medial affricate in contractions of 'did you' or a flap as in 'got it'.

1 most phrase-medial word boundaries.

2 a strong disjuncture marked by a pause or virtual pause, but with no tonal marks; i.e. a well-formed tune continues across the juncture.
OR

a disjuncture that is weaker than expected at what is tonally a clear intermediate or full intonation phrase boundary.

3 intermediate intonation phrase boundary; i.e. marked by a single phrase tone affecting the region from the last pitch accent to the boundary.

4 full intonation phrase boundary; i.e. marked by a final boundary tone after the last phrase tone.

For example, a typical fluent utterance of the following sentence:

```
Did you want an example?
```

might have a '0' between 'Did' and 'you' indicating palatalization of the /d j/ sequence across the boundary between these words. Similarly, the break index value between 'want' and 'an' might again be '0' indicating deletion of /t/ and subsequent flapping of /n/. The remaining break index values would probably be '1' between 'you' and 'want' and between 'an' and 'example', indicating the presence of a mere word boundary, and '4' at the end of the utterance, indicating the end of a well-formed intonation phrase.

In the $WAVES^{TM}$ label file, the number should be associated with a point in time at the end of each word, as indicated in the orthographic tier (Section 2). It should be located exactly at, or slightly to the left, of this word marker, so that break indices can be unambiguously associated with other tiers.

3.2 Uncertainty and Underspecification

Transcriber uncertainty about break-index strength is to be indicated with a '-' affixed directly to the right of the break index (e.g. '1-' to indicate uncertainty between '0' and '1'; '2-' to indicate uncertainty between '2' and '1'; and so on).

The full ToBI transcription must include both break index values and tone values. However, to accommodate backward compatibility with previously labelled databases or to allow intermediate stages in the labelling process, a partial ToBI transcription may have only break index values or only tone values assigned. Underspecification of break index values may be indicated by a value of 'X' at the word boundary in the break index tier.

3.3 Disfluencies

The perception of an audible hesitation (for example, an abrupt cutoff or a prolongation) can be marked by the diacritic 'p' immediately to the right of the break index (e.g. '3p'). This diacritic should be applied only to break indices of 1, 2, or 3. We expect that '1p' will be used for abrupt cutoffs, and '2p' and '3p' will be used to indicate prolongation, with '3p' suggesting hesitation after the onset of the tonal marks for an intermediate phrase. (See also Section 5.)

4 The Tone Tier

Two types of tones are marked in the tonal tier: pitch events associated with intonational boundaries (phrasal tones) and pitch events associated with accented syllables (pitch accents). The basic tone levels are high (H) in the local pitch range versus low (L) in the local pitch range.

4.1 Phrasal Tones

Phrasal tones will be assigned at every intermediate or intonation phrase:

L- or H- phrase accent, which occurs at an intermediate phrase boundary (level 3 and above); note that this represents a return to the notation in Pierrehumbert (1980)

L% or H% (final) boundary tone, which occurs at every full intonation phrase boundary (level 4)

%H high initial boundary tone; marks a phrase that begins relatively high in the speaker's pitch range; the default initial boundary is in the middle of the range or lower, and will be left unmarked in the transcription. Transcribers should use %H only when a high pitch at the beginning of an utterance cannot be attributed to a H accent (H* or H+!H*) on the first or second syllable in the utterance (i.e., when the first word itself does not appear to be accented, or when its accented syllable occurs too far into the word to account for the initial H), and where the utterance contrasts with a possible rendition with a lower-pitched onset.

Note that, since intonation phrases are composed of one or more intermediate phrases plus a boundary tone, full intonation phrase boundaries will have two final tones, e.g.:

L- L% for a full intonation phrase with a L phrase accent ending its final intermediate phrase and a L% boundary tone falling to a point low in the speaker's range, as in the standard 'declarative' contour of American English.

L- H% for a full intonation phrase with a L phrase accent closing the last intermediate phrase, followed by a H boundary tone, as in so-called 'continuation rise'.

H- H% for an intonation phrase with a final intermediate phrase ending in a H phrase accent and a subsequent H boundary tone, as in the canonical 'yes-no question' contour. Note that the H- phrase accent causes 'upstep' on the following boundary tone, so that the H% after a H- rises to a very high value.

H- L% for an intonation phrase in which the H phrase accent of the final intermediate phrase upsteps the L% to a value in the middle of the speaker's range, producing a final level 'plateau'.

For convenience, labelers may prefer to mark the tones at a break index with value '4' in a single step, with H-H%, L-L%, H-L%, or L-H%. We recommend that ToBI label menus include these symbols in addition to the separate symbols for phrasal and boundary tones described above; two additional symbols for downstepped phrase accent/boundary tone combinations will be described below in Section 4.3.

In the $WAVES^{TM}$ label file, the phrase accent and/or boundary tone associated with a phrase should be marked at a point at or just before the end of the last segment in the word ending the intermediate or full intonation phrase; high initial boundary tones should be marked at the beginning of the phrase, where the H tone is observed and should always be located after the break-index marker for any preceding phrase.

%r Will be used to mark the left edge of an intonation phrase which begins after a hesitation or disfluency. The '%r' notation is used to indicate a 'contour restart' — i.e. the initiation of a new intonational contour after a disruption. This diacritic should be used only in cases where the disfluency has caused a clear contour discontinuity.

4.2 Pitch Accents

Pitch accent tones will be marked at every accented syllable. Lack of pitch accent assignment for a syllable will be interpreted as meaning that the syllable is NOT accented. The ToBI transcription allows for the following five types of pitch accents. (Transcribers labelling utterances in dialects other than standard American English, standard Australian English, or RP British English may need to add additional types. These should be described in a general introduction to the transcribed database.)

H* 'peak accent' — an apparent tone target on the accented syllable which is in the upper part of the speaker's pitch range for the phrase. This include tones in the middle of the pitch range, but precludes very low F0 targets. [Corresponds to H* and H*+L in Pierrehumbert's six-accent inventory.]

L* 'low accent' — an apparent tone target on the accented syllable which is in the lowest part of the speaker's pitch range.

L*+H ‘scooped accent’ — a low tone target on the accented syllable which is immediately followed by relatively sharp rise to a peak in the upper part of the speaker’s pitch range.

L+H* ‘rising peak accent’ — a high peak target on the accented syllable which is immediately preceded by relatively sharp rise from a valley in the lowest part of the speaker’s pitch range.

H+!H* a clear step down onto the accented syllable from a high pitch which itself cannot be accounted for by a H phrasal tone ending the preceding phrase or by a preceding H pitch accent in the same phrase; should only be used when the preceding material is clearly high-pitched and unaccented. (Otherwise the accent is a simple !H*.)

In a $WAVES^{TM}$ label file, the pitch accent tone label should be placed within the nucleus of the accented syllable (i.e. the syllable that is phonologically associated to the starred tone of the accent).

If the F0 peak or valley for the starred H or L tone does not occur within the accented syllable, labellers who so wish may mark the early (or late) F0 event with ‘>’ (or ‘<’) pointing to the following (or preceding) pitch accent label. Thus, for example, if the F0 maximum for a L+H* occurs after the end of the accented syllable, a labeller may mark the time of the F0 peak with a ‘<’ pointing back to the L+H* label.

Implicit in our discussion of the five pitch accents is the notion that H* is the ‘default’ accent type. So, if there is any uncertainty about how low the F0 is before the peak, as in some cases of possible L+H* near the beginning of an utterance, the transcriber should mark ‘H*’ rather than ‘L+H*’.

4.3 Downstep Diacritic for Pitch Accents and Phrase Accents

Downstepped (high) tones will be marked explicitly using:

! preceding the downstepped pitch accent peak or downstepped H phrase accent. Transcribers familiar with Pierrehumbert’s full system should note that this eliminates the H*+L accent as a necessary downstep trigger within the system, since now the contrast between H* and H*+L will be marked by the absence versus presence of ‘!’ on the following H tone.

Note that, since it is the H tone in each case that is affected by the downstep, the ‘!’ diacritic should immediately precede the affected H tone in a pitch accent or phrase accent. Note also that this diacritic is NEVER applied to the first H tone in a phrase.

Some example uses of the downstep diacritic are:

H* !H- L% for the downstepped high phrase tone in the “calling contour” that in Pierrehumbert’s original system was analyzed as H*+L H- L%

H* !H* L- L% for the “staircase” pattern that in Pierrehumbert’s original system was analyzed as H*+L H* L- L%

L*+H L*+!H L*+!H for the succession of downstepped peaks that would occur in a succession of scooped accents

In light of our recommendations above that the possible tones of a level 4 break should be included as separate menu items, the possibility of the downstepped H phrase accent at full intonation phrase boundaries means that ‘!H-L%’ and ‘!H-H%’ should also be included as menu items.

4.4 Underspecification and Uncertainty

The full ToBI transcription must include both break index values and tone values. However, to accommodate backward compatibility with previously labelled databases or to allow intermediate stages in the labelling process, a partial ToBI transcription may have only break index values or only tone values assigned. Underspecification of tonal values may be indicated by ‘*’, ‘-’, and ‘%’ for a tonally unspecified pitch accent, phrase accent, and boundary tone, respectively. Note that this does not indicate uncertainty about the tonal value, but rather that the tonal values have yet to be assigned.

On the tonal tier, two kinds of uncertainty may be indicated: uncertainty over whether an event of a particular type has occurred, and uncertainty over the tonal value of an event that clearly has occurred. Thus, for example, the labeler may be unsure whether a particular syllable is accented, or, knowing that it is accented, may be uncertain of the accent type. Uncertainty of the first sort (whether the event has occurred) is indicated by '*?', '-?', and '%?' for pitch accents, phrase accents, and boundary tones, respectively. Uncertainty of the second sort (over the tonal value of a clearly occurring event) is indicated by 'X*?', 'X-?', and 'X%?'. Thus, for example:

***** means 'This syllable is accented but the database does not yet have accent type transcribed.'

***?** means 'I'm not sure whether this syllable is accented or not.'

X*? means 'I believe this syllable is accented but I am uncertain what accent type to assign.'

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.

5 Miscellaneous Tier

The miscellaneous tier will be used for any comments or markings (e.g., silence, audible breaths, laughter, disfluencies, and so on) desired by particular transcription groups. The only conventions TOBI specifies for this tier are that events should be labeled at their temporal beginnings and endings with labels of the form:

```
event< ... event>.
```

These labels should be placed in the text transcription or in the $WAVES^{TM}$ label file to correspond as closely as possible to the temporal beginning and endings of the phenomena begin described. So, a period of laughter plus speech might be indicated by marking the beginning and end of the laughter with:

```
laughter< ... laughter>
```

In general, it is the assumption of the participants in the common transcription group that silences should be automatically detectable, at least to a first approximation, and that transcriber time should not be spent marking these by hand. Disfluencies, by contrast, are not automatically detectable, and the absence of markings for them makes it difficult to parse the tone and break index tiers. For these reasons, transcribers are urged to mark disfluencies on the miscellaneous tier using 'disfluent<' and 'disfluent>', and to provide these marks in the miscellaneous tier menu when using $WAVES^{TM}$. Since demarcating a disfluent region is considerably more difficult than merely recognizing its presence, the marks 'disfluent<' and 'disfluent>' should be interpreted as rough pointers to the disfluent region and transcribers should not agonize over placing them precisely. Suggested conventions for further specification of particular types of disfluencies and their labels are provided in the "Guidelines for ToBI Labelling".

6 Pitch Range

HiF0 In transcriptions using $WAVES^{TM}$ label files, local pitch range will be marked for each intermediate phrase (interval between level 3 boundaries) with this diacritic. To estimate a phrase's pitch range, mark a point within the pitch accent in the phrase which includes a 'H' tone and which contains the F0 maximum for the phrase. That is, the accent containing the HiF0 mark should be one of H*, L+H*, L*+H, or H+!H*. Thus if an intermediate phrase contains only L* accents, HiF0 will NOT be marked for that phrase. Transcribers should take reasonable care to choose a point in time that reflects the target of the H for the accent. In several cases this will mean choosing some point other than the actual F0 maximum. For example, sometimes the highest F0 value in an accented syllable reflects the 'intrinsic' effect of a voiceless consonant and will thus be a poor estimate of the speaker's choice of pitch range. More seriously, in a phrase where the highest accent-related F0 occurs in a H* H- H% sequence, choosing the absolutely highest value for HiF0 will artifactually inflate the pitch range estimate by the amount of the upstep on the H%. In such cases, we recommend that the syllable's amplitude contour be used to pinpoint HiF0 within the candidate region.

7 Redundancy Among Tiers

There is some redundancy among tiers. For example, break index locations are redundant to the orthographic tier. Also, the occurrence of phrase accents and boundary tones on the tone tier is redundant to the presence of break index values '3' and '4' on the break index tier. In tonally underspecified databases, the marks '-' and '%' will be completely redundant to break index values '3' and '4'. Even in tonally specified databases, '-?' and '%?' will be redundant to break index values '3-' and '4-'. In order to save time and improve intertranscriber consistency, we recommend that labellers who use $WAVES^{TM}$ avail themselves of routines for automatically inserting redundant labels on either tier.

8 Files Associated with the Transcriptions

8.1 Speech File Formats

Since utterances will be recorded and transcribed at different sites, and for different immediate research purposes, it seems unlikely that we can arrive at any simple guidelines for such matters as sampling rate. We recommend adoption of formats compatible with other corpora insofar as possible.

8.2 Transcription Label Files

Each tier of TOBI should be representable in a simple text-based transcription, and as a separate label file in the $WAVES^{TM}$ label format. So, there will be separate label files for the orthographic tier, the break index tier, the tonal tier, and the miscellaneous tier. Such modularity allows partial transcription to be done and allows sites to add additional tiers as additional label files. All label files are of course aligned temporally via the waveform they label. This approach should also allow variation in display and access to different types of information. It is easy to provide software that supports labeling in such a format and that will generate summaries of prosodic information from such label files in a variety of formats.

8.3 Other Associated Analyses

We recommend conventions for additional analyses, e.g., recommended algorithms for pitch tracking and formant tracking, insofar as possible, although these do not form part of the transcription system we have proposed.

9 Conventions for Non-$WAVES^{TM}$ Format

by Jacques Terqen and Mari Ostendorf

Each line contains a number of fields. Fields are separated by markers to facilitate extraction of information. The format is as follows.

```
field_1 ^field_2 $field_3 @field_4 ;field_5
```

The contents of the fields are as follows.

Field_1 contains the orthographic transcription. The syllable(s) containing a pitch accent is/are marked by an asterisk (*) before the vowel.

Field_2 contains the tonal transcription, including pitch accents, phrase accents and boundary tones. If a word contains more than one pitch accent, the association with asterisk-marked syllables in Field_1 is from left to right. Uncertainty about the occurrence or type of pitch accent is indicated in this field using the conventions described in Section 4.4. In addition, the accented syllable having highest pitch within an intermediate phrase can be marked by HiF0. The convention is that HiF0 is associated with an accented syllable containing an H (either H*, L+H*, L*+H or H+!H*) — see Section 6. Finally, the convention with respect to phrasal accents is that a phrasal accent should be associated with the last word in the phrase, and that it is assumed to extend backwards until the last accented syllable in the phrase. This association convention is needed because the break index tier may not unambiguously indicate the location of an intermediate phrase boundary: "At the break index tier, a 2 may signal a disjuncture that is weaker than expected at what is tonally a clear intermediate .. phrase boundary" (See Section 3.1.)

Field_3 contains the break index value. This value gives the strength of the break between the word on the current line and the word on the next line. By definition, the beginning of a file is the beginning of the utterance; that is, there is an implied line before the first line only containing a 4 in the $-field, i.e. the tone field.

Field_4 contains the time markers associated with the break indices. Since in $WAVES^{TM}$ each tonal marker also has a time stamp, a possible extension is to have a list of time stamps rather than a single time stamp. Since phrasal accents and boundary tones are by definition associated with word boundaries, separate time stamps would be needed only for tonal markers containing an asterisk. The convention would be to associate the list of time stamps from left to right with tonal markers and the word boundary. If there are tonal markers associated with the word but only one time specified in Field_4, the time marked is by default the word boundary time.

Field_5 contains miscellaneous information comments. For comments continuing on the next line there is an obligatory continuation marker ";" at the beginning of the line, as follows:

```
than          ^           $1   @500      ;this is so much comment about all kinds
                                         ;of things that it continues on the next line
*eight        ^h*         $1   @600      ;
```

Thus, a typical line may be abstractly represented as:

```
w(*)ord ^tonal_marker $break_index @(time_stamp) time_stamp ;comment
```

An example of the fields of a non-$WAVES^{TM}$ transcription is shown below. (The neat organization in columns is purely for reading convenience and is not a requirement.) The waveform, F0 contour, and associated labels in a $WAVES^{TM}$ transcription are given in Appendix 10.

```
it's          ^               $1      @1.924903      ;
l*ovely       ^L+H* HiF0      $1      @2.303698      ;
and           ^               $1      @2.556273      ;
y*ellowish    ^L+!H* L-       $3      @3.118653      ;
and           ^               $1      @3.234365      ;
it's          ^               $1      @3.406066      ;
an            ^               $1      @3.514313      ;
*old          ^X*? HiF0       $1      @3.733797      ;
one           ^L-L%           $4      @4.074712      ;
```

10 Sample Utterance