Inter-transcriber reliability for two systems of prosodic annotation: ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch)

Mara Breen[1], Laura C. Dilley[2], John Kraemer[3], Edward Gibson[3]

[1]University of Massachusetts Amherst
[2]Michigan State University
[3]Massachusetts Institute of Technology

October 5, 2010
Please direct correspondence to:

Mara Breen
Department of Psychology
University of Massachusetts
Amherst, MA 01003
mbreen@psych.umass.edu

ABSTRACT

Speech researchers often rely on human annotation of prosody to generate data to test hypotheses and generate models. We present an overview of two prosodic annotation systems: ToBI (Tones and Break Indices) (Silverman et al., 1992), and RaP (Rhythm and Pitch) (Dilley & Brown, 2005), which was designed to address several limitations of ToBI. The paper reports two large-scale studies of inter-transcriber reliability for ToBI and RaP. Comparable reliability for both systems was obtained for a variety of prominence- and boundary-related agreement categories. These results help to establish RaP as an alternative to ToBI for research and technology applications.

Keywords: prosody, prosodic annotation, inter-transcriber reliability, ToBI, RaP

INTRODUCTION

Prosodic phenomena form the core of research questions in a wide variety of fields, including linguistics, speech technology, psychology, and computer science. Identifying instances of prosodic categories or events in speech corpora can be a useful research strategy; however, the question of how to automatically detect prosodic information from speech is still largely unsettled, in part because the acoustic factors that mediate the perception of accents and boundaries are complex and not fully understood (Choi, Hasegawa-Johnson, & Cole, 2005; Duez, 1993; Kochanski, Grabe, Coleman, & Rosner, 2005; Salverda, Dahan, & McQueen, 2003; Watson, Arnold, & Tanenhaus, 2008). Annotation systems by which human listeners code prosodic events in speech corpora have thus been essential tools for investigating questions of how prosody conveys information in spoken language.

In the early 1990s, the Tones and Break Indices (ToBI) system of prosodic annotation for mainstream American English (Beckman & Ayers Elam, 1997; Silverman et al., 1992) was developed by a group of speech researchers drawn from several disciplines. Since its development, ToBI has been widely adopted as the standard annotation system, but has also been shown to have certain limitations, which will be discussed in detail below. One concern about the ToBI system is the extent to which coders agree on the labels they apply to speech. We will argue that although several inter-transcriber reliability studies have been conducted, none has utilized enough speech or coders to effectively gauge ToBI's reliability. Therefore, one of the goals of the current paper is to conduct a large-scale evaluation of inter-transcriber reliability in ToBI.

Recently, Dilley and Brown (2005) developed the Rhythm and Pitch (RaP) prosodic annotation system to provide an option other than ToBI to speech prosody researchers and technologists. The RaP system differs from ToBI in that it takes into account developments in phonetics, phonology, and speech technology since the development of the original ToBI system. The second goal of this paper is to motivate the development of RaP and detail its distinct features and potential advantages compared with ToBI. In addition, we will present the first evaluation of inter-transcriber reliability in the RaP system; this evaluation is usefully accomplished in part through a comparison with agreement benchmarks for ToBI.

The paper is organized as follows: As background, we first describe the development and usage of ToBI and review work evaluating ToBI categories. We then describe the RaP system, and indicate how it was designed to address some of ToBI's perceived limitations. Finally, results of two studies of inter-transcriber reliability for both systems are presented, one using naïve labelers (Study One) and the other using expert labelers (Study Two).

The ToBI System

ToBI was developed in the early 1990s by researchers from linguistics, psychology, and computer science (See Beckman, Hirschberg, & Shattuck-Hufnagel, 2005 for discussion of the development of ToBI.). ToBI is based on a phonological approach to prosody, that of autosegmental-metrical theory (Beckman & Pierrehumbert, 1986; Dilley, 2005; Goldsmith, 1976; Liberman, 1975; Pierrehumbert, 1980), and its labeling conventions are largely based on the theoretical work of Pierrehumbert and colleagues (Beckman & Pierrehumbert, 1986; Pierrehumbert, 1980; Pierrehumbert & Beckman, 1988), as well as earlier labeling systems (Price, Ostendorf, Shattuck-Hufnagel, & Fong, 1991; Wightman, Shattuck-Hufnagel, Ostendorf, & Price, 1992). Variants of ToBI have been proposed for a number of languages and dialects (Jun, 2005);

in the present document, we will use "ToBI" to refer exclusively to the version that Beckman et al. (2005) proposed for Mainstream American English.

A standard ToBI transcription consists of four tiers of labels time-aligned with the speech signal: (1) an *orthographic tier*, on which text is labeled; (2) a *tonal tier* for labeling pitch events; (3) a *break-index* tier for labeling perceived disjuncture between words; and (4) a *miscellaneous tier* for additional information. Recently, a fifth *alternative* (or *alt*) tier for alternative label choices has been proposed (Brugos, Veilleux, Breen, & Shattuck-Hufnagel, 2008). Tonal and break index tiers form the core of a ToBI transcription; an example of an utterance annotated with ToBI is shown in Figure 1.

<u>*Tone tier*</u>. The set of symbols used in ToBI is presented in Table 1. Labels are determined jointly by a labelers' auditory perception of prosodic events and by visual inspection of the fundamental frequency (F0) contour. Both prominence (i.e., pitch accent) information, as well as phrase tone information are captured on this tier. The pitch accent notation (*) is used to indicate syllables that are perceived to have prominence relative to other words in the current phrase, a property which is unpredictable from the lexical stress characteristics of words (Beckman & Ayers Elam, 1997). ToBI conventions ensure that prominent, pitch-accented syllables are usually, though not always, accompanied by a pitch change or excursion. ToBI conventions require that each syllable be either pitch-accented or unaccented. There are five basic pitch accent types, which are single-toned (H*, L*) or bitonal (L+H*, L*+H, and H+!H*) with three downstepped variants (!H*, L+!H* and L*+!H).[1] A complex, many-to-many mapping exists between F0

[1] Downstepped variants of pitch accents (i.e., those marked with '!') are identical to non-downstepped variants except that the pitch of the high tone of the accent is judged lower than that of a preceding high tone in the same intermediate intonational phrase.

characteristics and pitch accent types (Dilley, 2005; Dilley & Brown, 2007; Ladd, 2000, 2009; Pierrehumbert, 1980).

The second type of tonal label, edge tones, marks the right edges of prosodic phrase units. Two sizes of phrasal unit are assumed; the smaller and larger units are the intermediate phrase (ip) and the intonational phrase (IP), respectively, and these are hierarchically structured, such that the former is nested within the latter (Beckman & Pierrehumbert, 1986; Pierrehumbert & Beckman, 1988). The right edge of each syllable sequence that is judged to correspond to an ip is obligatorily marked with one of three tonal labels, termed phrase accents: H-, !H-, or L-. Moreover, the right edge of each syllable sequence judged to correspond to an IP is obligatorily marked by a combination of a phrase accent plus an additional label (H% or L%) termed a boundary tone; the two boundary tones and three phrase accents produce six possible phrase accent-boundary tone combinations at IP boundaries.[2] The mapping assumed from edge tone labels to F0 is complex and depends on both the shape of the phrase-final F0 contour and its position in the pitch range (Beckman & Ayers Elam, 1997). The mapping includes certain exceptional treatments; for example, both H-L% and !H-L% correspond to unidirectional, falling F0 movement relative to the preceding tone, but L-H%, which also involves both L and H tones, generally corresponds to bidirectional falling-rising F0 movement relative to the preceding tone (Beckman & Ayers Elam, 1997).

Break index tier. The break index tier is used for labeling break indices: numbers from 0 to 4 which specify the perceived degree of disjuncture between words. For example, a ‘0’ indicates the smallest degree of disjuncture, as for that associated with fast speech or

[2] Only five of these phrase accent plus boundary tone combinations are referenced in the original ToBI annotation conventions (H-H%, L-L%, H-L%, !H-L% and L-H%) (Beckman & Hirschberg, 1994). However, the sixth combination (!H-H%) is logically possible, and is referenced in more recent training materials (Brugos, Shattuck-Hufnagel, & Veilleux, 2006).

cliticization processes, e.g. *didja* for *did you*. A ‘1’ indicates normal disjuncture between words in fluent speech. Beyond ‘0’ and ‘1’, however, increasing values of break indices do not necessarily correspond to monotonically increasing degrees of perceived disjuncture. At the highest end of the scale, ‘4’ typically indicates maximal disjuncture corresponding to an intonational phrase boundary (i.e., IP), often signaled by durational lengthening, pitch movement, and/or silence. However, a ‘4’ is also prescribed anytime the F0 contour warrants labeling a phrase accent-boundary tone combination on the tonal tier, even if the perceived degree of disjuncture is less than for a typical ‘4’ (Beckman & Ayers Elam, 1997). Similarly, the label ‘3’ typically indicates moderate disjuncture corresponding to an intermediate phrase boundary (i.e., ip), which is greater than that of ‘1’ but less than that of ‘4’. However, a ‘3’ must also be labeled anytime a phrase accent (H-, L-, or !H-) is labeled on the tone tier, even if the boundary lacks any significant sense of disjuncture. The diacritic ‘-’ can be added to break indices to indicate less perceived disjuncture than normally associated with a given break index or to indicate uncertainty between two consecutive break index levels. For example, ‘4-‘ could indicate either that the degree of disjuncture was less than a full ‘4’ or else uncertainty between break indices ‘3’ and ‘4’. Finally, a break index ‘2’ may be used either to indicate the presence of lengthening or sizeable disjuncture when a phrase accent or boundary tone is judged not to be present, or else the absence of lengthening when a phrase accent or boundary tone *is* judged to be present (Beckman & Hirschberg, 1994). This dual definition of ‘2’ is a significant component of non-monotonicity for the relationship between increasing break index values and perceived disjuncture.[3]

[3] The duality of the definition of ‘2’ has also led to this break index being colloquially termed the “garbage pail category”.

The ToBI system has several strengths. For example, it was designed as a standard and has been in international use for over a decade. Moreover, modified versions of ToBI have been created for numerous languages and dialects (Jun, 2005). In addition, there is broad empirical support for several aspects of the theoretical framework on which ToBI is based (Arvaniti, Ladd, & Mennen, 1998; Ladd, 1996, 2000; Ladd, Faulkner, Faulkner, & Schepman, 1999).

ToBI Limitations. It has been argued in the literature that ToBI has certain drawbacks, which collectively have motivated the development of the RaP system as an alternative prosodic annotation system. First, several lines of evidence suggest that listeners hear more or fewer perceptual categories than ToBI models. For example, ToBI allows for substantial phonetic variability in the H* accent such that tokens which differ only in terms of the timing of the highest F0 point (i.e., F0 peak) within a syllable are labeled as H* (Beckman & Ayers Elam, 1997; Silverman & Pierrehumbert, 1990). However, work by Dilley (2005; submitted; Redi, 2003) has demonstrated that listeners hear distinct perceptual categories for different patterns of F0 peak alignment classified as ToBI H*. This variability in peak timing can be captured in ToBI, but there are multiple ways to do so, leading to issues with inter-transcriber reliability (Shattuck-Hufnagel, Dilley, Veilleux, Brugos, & Speer, 2004). Moreover, evidence from a variety of experimental paradigms suggests that listeners do not treat contours corresponding to H* and L+H* as categorically distinct (Bartels & Kingston, 1994; Braun, 2006; Calhoun, 2006; Dilley, 2007, 2010; Watson, Tanenhaus, & Gunlogson, 2008).

A second, related drawback of ToBI is its lack of a consistent, transparent mapping between labeling distinctions and phonetic and/or perceptual events. For example, ToBI has been criticized for not capturing the perceptual experiences of labelers (Wightman, 2002). In particular, ToBI requires that labelers select break indices of ‘3’ or

‘4’ in absence of a perceptual sense of disjuncture when an edge tone (phrase accent and/or boundary tone) is labeled on the tonal tier. In addition, ToBI guidelines dictate that some syllables lacking any local pitch excursion should be labeled as pitch accents (e.g., a string of L* accents) (Beckman & Ayers Elam, 1997; Pierrehumbert, 1980). Finally, there are inconsistent assumptions across languages in the phonetic correlates of ToBI pitch accent types, posing significant problems both for theoretical treatment of the phonetics-phonology interface, as well as applicability of the system to not-yet-described languages (Ladd, 2000, 2009). These aspects of ToBI make it challenging for labelers to learn and use the ToBI system.

Third and finally, ToBI does not provide a way of labeling certain distinctions that are important for many speech prosody researchers. For instance, ToBI allows only for coding a binary prominence distinction: as pitch-accented vs. unaccented. This dichotomy requires that labelers must designate any syllable that sounds perceptually prominent to be “pitch-accented”, even if there is no evidence of pitch change in the vicinity of the syllable (Beckman & Ayers Elam, 1997; Wightman, 2002). In addition, evidence exists suggesting that three accent categories or prominence levels can describe the variation in speech better than two (Greenberg, Carvey, & Hitchcock, 2002), consistent with other annotation systems (e.g., Halliday, 1967). ToBI also does not allow for coding rhythmic patterns of speech, which are vital to both language acquisition and mature language perception (Cutler & Norris, 1988; Nazzi & Ramus, 2003). The importance of rhythmic (or, metrical) prominence as distinct from “pitch accent” is highlighted by recent work by Beaver et al. (2007) who showed that meaningful distinctions (in this case in second-occurrence focus) are signaled using non-pitch related prominence cues.

The RaP System

The RaP (Rhythm and Pitch) system (Dilley & Brown, 2005) is based, like ToBI, on autosegmental-metrical theory (Beckman & Pierrehumbert, 1986; Dilley, 2005; Goldsmith, 1976; Liberman, 1975; Pierrehumbert, 1980). It was developed to meet the needs of the speech prosody research community by addressing certain perceived weaknesses of ToBI described above. A RaP transcription is based on labelers' perception of prosodic events; unlike ToBI, a visual display of the signal is considered an aid for annotation rather than a requirement. A transcription comprises four tiers of acoustically time-aligned symbolic labels: (1) a *words tier* for syllables; (2) a *rhythm tier* for speech rhythm; (3) a *tonal tier* for tonal information,; and (4) a *miscellaneous tier* for additional information; rhythm and tonal tiers constitute the core of a RaP transcription.

Rhythm tier. The rhythm tier permits information about prominence and phrasal boundaries to be captured (cf. the tonal tier in ToBI). Considering first the labeling of prominence, prominent syllables are designated as "beat" syllables ('X' or 'x'). The label 'X' indicates that a syllable is a strong metrical prominence, while 'x' indicates that a syllable is a moderate-to-weak metrical prominence. As with ToBI, RaP conventions indicate that syllables are to be labeled as prominent when this prominence is perceived in the phrasal context and is not predictable from lexical stress characteristics alone.

Moreover, phrasal boundary information is indicated on the rhythm tier on the basis of a labeler's perception of degree of disjuncture; ')' indicates a minor phrase boundary approximately equivalent to a ToBI break index of '3', and '))' indicates a major phrase boundary approximately equivalent to a ToBI break index of '4'. Unlike

ToBI, tonal labels are not obligatorily marked at phrasal boundaries. Rather, tonal labels are indicated only when the F0 evidence warrants it.

Pitch tier. All tonal events are indicated with a label which captures the pitch of the labeled syllable in relation to that of the preceding labeled tone; specifically, a syllable can be labeled with ‘H’, ‘L’, or ‘E’ to indicate that it is higher than, lower than, or equal to, respectively, the preceding labeled tone.[4] This relative tone labeling scheme thus bears similarities to the International Transcription System for Intonation (INTSINT) developed by Hirst and colleagues (Hirst & Di Cristo, 1998).

The tonal primitives ‘H’, ‘L’, and ‘E’ can be marked with a variety of diacritics to provide information about the relationship between tonal and rhythmic characteristics of speech. First, in the RaP system, pitch accents correspond to those tonal targets which are associated with metrically prominent (strong or weak beat) syllables (‘X’ or ‘x’ in the metrical tier); these are termed “starred tones” and are marked with an asterisk (e.g., ‘H*’). As a result, pitch-accented syllables in RaP correspond to a subset of prominent syllables, giving rise to a three-way basic prominence-based distinction (non-prominent, prominent but not pitch-accented, prominent plus pitch-accented), compared with the two-way prominence distinction in ToBI (pitch-accented vs. unaccented). Second, tonal targets that occur on non-metrically prominent syllables are termed unstarred tones; these are tones that precede or follow a starred tone, and are labeled with a ‘+’ (before or after the tonal label depending on the location of the starred tone) indicating their association with the adjacent starred tone (e.g., L+ in L+H*). Moreover, RaP distinguishes small and large pitch changes, capturing possibly meaningful (e.g., focus-related) distinctions in pitch excursion size (cf. Bartels & Kingston, 1994; Braun, 2006). Small pitch changes are

[4] If the tone is initial in an utterance, the diacritic “:” is used and the tonal label (H, L or E) indicates the pitch of the labeled syllable in relation to the labeled tone to the right.

indicated for syllables with a pitch change of less than three semitones from the previous syllable, and are indicated with the '!' diacritic (e.g., !H, !L). Finally, tonal labels ('>>' and '<<') can be indicated on phrase-final syllables which demonstrate a change in pitch to the highest or lowest part of the speaker's pitch range, respectively.

The tonal labeling inventories of ToBI and RaP are comparable in many ways, including the fact that tone labeling is sparse in RaP and describes the overall pattern of pitch changes in speech (Beckman & Ayers Elam, 1997; Dilley & Brown, 2005) (e.g. see Figure 1). For example, RaP permits a means of capturing all the distinctions that have been investigated as the basis of meaning differences in the ToBI framework (e.g., L*+H vs. L+H*, Pierrehumbert & Hirschberg, 1990; Pierrehumbert & Steele, 1989). An important difference between them, however, is the fact that the phonetic mapping from acoustic correlates to tonal labels is simpler and more consistent in RaP, for several reasons. First, the choice of tonal primitives in RaP, i.e., H, L or E tones, is based uniformly on the tone's relation to the preceding labeled tone, whereas the choice of tonal primitives in ToBI, e.g., H* vs. L*, is based on a variety of factors: the tone's relation to the preceding labeled tone, the size of its pitch excursion relative to other syllables, and/or its position in the speaker's pitch range (Pierrehumbert, 1980; Beckman & Ayers Elam, 1997). Second, RaP allows for a more consistent treatment of certain meaningful pitch scaling variables than ToBI, such as the size of pitch excursions and the position of a tone in the speaker's overall pitch range. Third, turning points are uniformly analyzed as arising from tones in RaP, consistent with recent research (Dilley, 2005, 2010; Ladd, 2000, 2008; Ladd & Schepman, 2003), unlike in ToBI. For example, while a dip between two peaks is often treated as a non-phonological "sagging transition" in ToBI, and receives no label, such a dip will always be labeled with 'L' in RaP, reflecting the presence of a low tonal target. The correspondence between labels for pitch events in

ToBI and in RaP is shown in Table 1. Moreover, Figure 1 illustrates how the phonetically simpler tonal inventory in RaP permits more contours to be distinguished than in ToBI, potentially permitting more meaningful intonation distinctions to be captured.

Figure 1 provides an illustration of speech annotated with RaP. Note that while metrical prominences are labeled on the rhythm tier, pitch accents (i.e. starred tones) are labeled on the tonal tier, and so pitch accents are indicated if and only if there is a local pitch excursion in the immediate phonetic vicinity. In this way, RaP distinguishes syllables that are prominent due to a pitch excursion — "true" pitch accents — from syllables that are prominent for other (e.g., rhythmic) reasons, and it contrasts with ToBI, where syllables which are heard as prominent for any reason (including non-tonal reasons) are labeled as having pitch accents.

In sum, RaP presents a number of strengths as a prosodic annotation system. First, RaP was designed to be easier to learn and use than ToBI, in part because it embodies a consistent relationship between prosodic labels and phonetic and perceptual characteristics. Second, like ToBI, the RaP system builds on the well-established autosegmental-metrical theory (Beckman & Pierrehumbert, 1986; Pierrehumbert, 1980; Pierrehumbert & Beckman, 1988). In addition, RaP labels are based only on listeners' perceptions of events, rather than being determined by prior theoretical assumptions. Third, RaP uses a prosodic label set for capturing tonal information in which unstarred tones and starred tones are selected independently and associated with syllables. This independent selection is consistent with recent phonetic evidence that unstarred tones show are anchored to and are closely coordinated with a given syllable, rather than with the starred tone of a bitonal pitch accent (Arvaniti et al., 1998; Arvaniti, Ladd, & Mennen, 2000; Dilley, Ladd, & Schepman, 2005; Ladd, Mennen, & Schepman, 2000). Moreover, RaP permits a wider variety of potentially meaningful contours to be distinguished, as illustrated in Figure 2. Finally, RaP fills a gap in the speech research community by presenting a method of labeling rhythmic information as distinct from pitch, and allowing multiple levels of prominence to be represented.

The previous section provides the theoretical motivation for RaP, including the important ways in which it addresses the perceived limitations of the ToBI system. However, another important characteristic of a useful prosodic annotation system is that labelers agree on its use. As stated earlier, there is no published inter-transcriber reliability data on RaP (but see preliminary data in Dilley, Breen, Bolivar, Kraemer, & Gibson, 2006), and inter-transcriber reliability studies of ToBI have been weak for several reasons, as will be described below. The current studies were designed, therefore, to generate data about the agreement of multiple coders on a large corpus of speech for both systems.

Previous studies of inter-transcriber reliability for ToBI

There have been three previous studies of inter-transcriber reliability using ToBI, each of which has had empirical limitations. In an initial study by Pitrelli et al. (1994), 26 labelers applied the ToBI system to 489 words taken from both read and spontaneous speech corpora. Although there were a large number of labelers, there were two major limitations of this study: (1) the agreement metric used did not take into account the possibility of chance agreement; and (2) the corpus that was labeled was very small, and probably not representative of typical speech. A more recent study by Syrdal & McGory (2000) employed six labelers who annotated 645 words. Although this study did take into account chance agreement by using a chance-adjusted kappa metric (Carletta, 1996), the speech corpus was very small and was comprised of only two speakers reading the same words, so the results may not generalize to other speakers or to spontaneous speech.

Finally, Yoon et al. (2005) investigated inter-transcriber reliability in ToBI; whereas this study used both a chance-adjusted agreement metric and a larger corpus of spontaneous speech (including 79 speakers and 1600 words), the speech was annotated by only two labelers.

Despite limitations in the design of these prior investigations of ToBI agreement, they have revealed some consistent findings. First, all prior studies of ToBI agreement have demonstrated high agreement on the presence of a pitch accent (>80%), and moderate agreement on pitch accent type (>60%). Moreover, all three studies have demonstrated high agreement with regard to the presence vs. absence of intonational boundaries (>89%). However, there are shortcomings in the designs of these previous studies which necessitate a larger scale ToBI agreement study. Given that ToBI is considered the current standard prosodic labeling system, for RaP to show comparable or better agreement to ToBI would help to establish RaP as a viable alternative available to researchers and technologists for purposes of prosody labeling.

STUDY ONE

The motivations for the first study were: (1) to conduct a more complete test of inter-transcriber reliability for the ToBI system than had been previously performed, using multiple trained labelers with no previous annotation experience, a sizeable corpus of speech, and appropriate statistical measures of agreement between labeler; and (2) to assess inter-transcriber reliability for RaP using the same labelers, corpus and statistical measures.

Method

Participants. Four MIT undergraduates served as labelers. Each received course credit or monetary compensation at a rate of $8.75/hour for the duration of the project. Three of the labelers had taken an introductory linguistics course; none had any knowledge of prosody research, nor any experience with prosodic annotation.

Materials. To ensure representation of diverse speech styles, materials were drawn from two speech corpora: the Boston Radio News (BRN) corpus of read professional broadcast news speech (Ostendorf, Price, & Shattuck-Hufnagel, 1995), and the CallHome corpus of spontaneous nonprofessional speech from telephone conversations (Linguistic Data Consortium, 1997). The amount of speech from each corpus that was annotated in each system is shown in Table 2. Materials were divided into 60 sound files, with a mean duration of 35 seconds (SD 18 seconds).

Procedure. Training and testing on the prosodic systems occurred in three successive phases. In the first phase, labelers trained and were tested on ToBI; they then applied this system to the speech corpora. In the second phase, the labelers trained and were tested on the RaP system; they then applied it to a subset of the corpus which had already been annotated with ToBI. In the third phase, the labelers annotated a smaller corpus of speech with the ToBI system, which had not been previously annotated. Inclusion of a second period of ToBI labeling permitted testing whether higher agreement between labelers might result from more labeling experience in general, regardless of the identity of the prosodic labeling system.[5] Details about training and labeling of the test materials are given below.

During the initial phase of the project, labelers were trained on ToBI through the manual and computerized exercises in Beckman and Ayers Elam (1997), as well as receiving one-on-one feedback from an expert labeler (author MB) and participating in weekly meetings with four expert ToBI labelers throughout the project (author MB and three other ToBI experts in the MIT speech community). After initial training, labelers

[5] This order of system application was used due to the unavailability of materials for RaP training at the beginning of the first study period.

received feedback from the experts on two 60-second practice annotations. Next, they completed a ToBI test, in which their annotations of a 90-second mini-corpus (approximately 60 seconds read speech, 30 seconds spontaneous) were graded by three experts, including authors MB and LD.[6] None of the annotated speech materials used during the training and feedback phase were included in subsequent agreement analyses. Labelers subsequently spent four weeks annotating 26.7 minutes of the corpus with ToBI (11 spontaneous, 15.7 read). The order of files was pseudo-randomly determined so that approximately equal amounts of read and spontaneous speech would be annotated, and so that successive files came from different speakers. The order of files in the corpus was the same for every labeler.

During the second phase of the project, labelers spent two weeks learning RaP via written guidelines and computerized exercises (Dilley & Brown, 2005).[7] After an initial week of intensive group training with the manual, labelers received extensive feedback on two 60-second practice annotations from author LD. Next, they completed a RaP "test," in which they were graded on annotations of a 60-second mini-corpus comprised of approximately 30 seconds of read speech, and 30 seconds of spontaneous speech. All labelers passed this test and were cleared to begin annotating the corpus according to the RaP conventions. Labelers spent the next four weeks annotating 19.2 minutes of the

[6] The annotations were evaluated using the following system: One or two points were deducted for each label with which the expert mildly or moderately disagreed, respectively. Three points were deducted when a label was strongly disagreed with and/or presented incorrect ToBI syntax. Experts also employed a subjective grading system ranging from excellent (5) to poor (1), indicating their overall impression of the labels. Three coders received average grades of 4 or higher from all three expert evaluators on both test files and began annotating the corpus. The other two coders received average grades of 3 from the experts, and were instructed to go back through the guidelines, paying attention to the labels they had misused in the test labels. After another week of training, they too began corpus annotation.

[7] Available at http://tedlab.mit.edu/tedlab_website/RaPHome.html.

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

Finally, during the third phase labelers annotated 9.4 minutes of the corpus (4.2 spontaneous, 5.2 read) using ToBI.

Data analysis: Agreement measures. Raw- and chance-corrected agreement scores were calculated for all comparisons. Raw agreement was calculated using the "labeler-agreement-pair" approach proposed by Silverman et al. (1992) and used subsequently in other studies (Pitrelli, Beckman, & Hirschberg, 1994; Syrdal & McGory, 2000; Yoon, Chavarria, Cole, & Hasegawa-Johnson, 2004). We refer to this metric as the *transcriber-agreement-pair* (TAP). Our TAP approach uses syllables and words as the units of agreement for prominence-based and phrase-based metrics, respectively. Note that in previous studies of ToBI inter-transcriber reliability, prominence labels were taken to be aligned with whole words, so that e.g., two labelers were said to agree on the presence of a pitch accent if both labeled a pitch accent on a given word, even if each labeled that pitch accent on a different syllable. In contrast, labelers in the present study aligned prominence labels with an individual syllable. This alignment scheme is both a more faithful representation of perceived prosody (since prominences are aligned to specific syllables), and it allows direct comparison between prominence placement in both systems. Overall, the higher specificity required for agreement means that this measure may result in lower agreement than in previous studies.

Note that the original TAP method of Pitrelli et al. (1994) did not adjust for expected chance agreement rates, which varies with the number and distribution of available labels (Carletta, 1996). For instance, if there are only two labels, and these are used with equal frequency, the probability that two labelers will agree by chance on these labels is 50%, while if there are five equally-probable labels, then chance agreement is 20%. However, if there are two categories, one of which is used 90% of the time, chance agreement is not 50%, but 82%. To adjust for chance agreement, Carletta recommends the kappa (κ) metric (1), where A_E is expected agreement based on chance and A_O is observed (or actual) agreement:

(1) $$\kappa = \frac{A_O - A_E}{1 - A_E}$$

Values of κ were computed as follows in the current studies: First, specific labels were grouped into *label equivalence relations*, depending on what labels counted as equivalent for a particular analysis. These groupings were necessary because the same label could be treated differently depending on what it was being compared to. For example, a label of H* from one labeler and L* from another could agree in some cases, but not others. Specifically, the calculation of agreement on presence vs. absence of a pitch accent was performed on the data where (1) the five labeling options H*, L+H*, !H*, L*, L*+H[8] were taken as equivalent and (2) the two labeling options ?* or *no label* are taken as equivalent. Alternatively, the calculation of type of pitch accent were performed on the data where (1) !H*, H+!H*, L+!H*, and X* were taken as equivalent, (2) L* and L*+H were taken as equivalent, and (3) ?* or *no label* are taken as equivalent.. Therefore, H* and L* would agree under the first comparison, but not the second.

A *transcriber-agreement pair* refers to a pair of labels, one from each of two labelers, which are both assigned to the same syllable or word from the same recording. If the pair of labels is drawn from the same group based on the equivalence relation (e.g., H* and L*, in the presence of pitch accent comparison), then it counts as agreement; otherwise, it counts as disagreement.

[8] The two ToBI downstepped labels, L+!H* and L*+!H, were treated as equivalent to their non-downstepped versions, L+H* and L*+H, respectively.

Overall agreement for a given category of prosodic agreement analysis (see Table 3) was calculated as follows. First, raw agreement values were calculated from labels assigned to each recording from all transcriber agreement pairs for that recording, based on the label equivalence relation for that category of agreement analysis. Next, observed agreement, A_o, was calculated by determining the weighted arithmetic mean of the raw observed agreement values (transcriber agreement pairs) of each recording; weights corresponded to the product of the number of transcriber pairs and the number of units of comparison (*viz.*, words or syllables) for that recording. Next, chance agreement was calculated by determining the maximum posterior likelihood estimate of the chance agreement rate given the distribution of labels in the recording's annotations. That is, we determined, for each label equivalence relation, the probability that two labelers would agree, given the frequency with which the labels in that label equivalence relation occurred. A label equivalence relation consisting of only two groups, one of which is very frequent, would thus result in higher chance agreement than a label equivalence relation with multiple groups which occurred with similar frequency. For example, the label equivalence relation of *boundary vs. no-boundary* would be an example of a relation with only two groups—the 'boundary' group consisting of break index labels of 3 or 4, and the 'no boundary' group, consisting of all other break index labels—where one group (the second group in this example) occurs with greater frequency (since most words do not correspond to a boundary). As such, this label equivalence relation would have high chance agreement. Conversely, the label equivalence relation *type of pitch accent*, with groups of label equivalence relations occurring with varying frequency, would have lower chance agreement. Overall chance agreement for each label equivalence relation (A_E) was then computed across the entire corpus by computing a weighted mean as described above, and κ was then computed as in (1).

Results and Discussion

Values of chance-corrected κ over .40 and over .60 have previously been taken to indicate moderate and substantial agreement, respectively (Landis & Koch, 1977; Rietveld & van Hout, 1993). Syrdal and McGory (2000) furthermore interpreted κ values of .6 or higher as reliable. Moreover, κ values between .67 and .80 have been taken to be 'tentatively conclusive', while values above .8 have been described as 'conclusive' (Krippendorff, 1980). Following this work, we therefore interpret the degree of agreement according to the stratification of κ values shown in Table 4. Note that absolute values for the TAP metric cannot be similarly stratified concerning reliability, since TAP values are not corrected for chance; instead, they are useful as relative reliability measures, given comparable numbers of labeling categories, as well as for comparison with previous studies of ToBI reliability.

In the following, we focus on agreement within and across transcription systems for the current study. Comparison of agreement values with other studies is considered in the General Discussion.

Agreement for ToBI and RaP is given in Table 5. We first consider agreement for ToBI. Inspection of the TAP metric (first column), which is not chance-corrected, suggests that labelers agreed best on the binary prominence vs. non-prominence judgment; they also agreed well on the ternary high vs. low (vs. absent) accent type distinction, which merges several ToBI categories. Chance-corrected κ values for ToBI (third column) are more informative. The highest κ agreement results from the binary categories of prominence vs. nonprominence and the ternary high vs. low (vs. absent) accent type distinction, consistent with TAP scores; agreement for both these categories was substantial and reliable (κ > .60). However, three of the five agreement categories for

ToBI--type of pitch accent (all distinct), phrasal boundary present/absent, and size of phrasal boundary--score below 0.60, indicating questionable reliability with only moderate agreement. It is perhaps not surprising that the type of pitch accent was not reliably indicated, given the multiple pitch accent categories that ToBI employs. What is perhaps more surprising is that phrasal boundaries were labeled with questionable reliability, both in terms of the binary boundary present vs. absent judgment, as well as in the three-way IP (large boundary) vs. intermediate IP (medium boundary) vs. no-boundary distinction. Finally, agreement across categories was generally higher for read speech than for spontaneous speech, as expected, since read speech is considered to have clearer prosodic cues (e.g., Rouas, Farinas, Pelligrino, & Andre-Obrecht, 2003). The exceptions were the two categories dealing with phrasal boundaries, for which agreement was slightly higher for spontaneous than read speech.

Tables 6 and 7 display confusion matrices for ToBI labels of pitch accent and break index, respectively. Table 6 suggests, among other things, that labelers often disagreed on H* and L+H* labels, and that the H+!H* label was rarely agreed upon by two labelers. Moreover, Table 7 demonstrates that the ‘2’ label was rarely agreed upon by pairs of labelers. In general, both tables show a substantial number of off-diagonal pairings and provide detail to supplement Table 5 about the specific nature of disagreements that arose in this study within the ToBI system.

Next, we consider agreement for RaP. Inspection of TAP values (second column of Table 5) suggests that labelers agreed best on the binary present vs. absent judgments for the presence of a phrasal boundary and for the prominence vs. non-prominence distinction, respectively. Moreover, inspection of chance-corrected κ values (fourth column) likewise shows the highest agreement for binary judgments for the presence of a phrasal boundary and of prominence vs. non-prominence, respectively. In addition, five of six agreement categories for RaP show κ values above 0.60, indicating that most prosodic distinctions examined were made reliably and with substantial agreement. Finally, agreement was uniformly higher for read speech than spontaneous speech, for all agreement categories examined. Tables 8, 9, and 10 display confusion matrices for RaP labels. These tables reveal disagreement concerning whether a syllable is prominent and how prominent it is (Table 8), whether a syllable has a starred tone and what type (Table 9), and whether a phrasal boundary is present and what its size is (Table 10). Together, these tables provide substantial detail to supplement Table 5 about the specific nature of disagreements that arose in this study within the RaP system.

To investigate variability in overall agreement across labelers, we also calculated the pairwise agreement between pairs of annotators in ToBI and in RaP, averaged across values of Kappa shown in Table 5. The result is shown in Table 11. For the pair of labelers represented by each cell, the first and second numbers indicate the average Kappa values for ToBI and for RaP, respectively. It can be observed that average pairwise Kappa agreement across labelers for ToBI ranges from .51 to .62, while for RaP agreement across labelers ranges from .62 to .65.

Agreement across the two systems can also be directly compared for several categories. First, the binary prominence/non-prominence judgment can be compared in RaP (cf. beat vs. nonbeat) and in ToBI (cf. pitch-accented vs. unaccented). For this distinction, comparable agreement is observed across systems in TAP values (87% in ToBI vs. 89% in RaP), and RaP shows a numerical agreement advantage in terms of chance-corrected κ values (0.71 for ToBI and 0.77 for RaP).[9] The statistical reliability of

[9] Note that the pitch-accented vs. unaccented distinction cannot be directly compared across ToBI and RaP. This is because ToBI entails only the two-way pitch-accented vs. unaccented prominence distinction in which a labeler must judge simultaneously both whether a syllable is prominent and/or has a salient pitch excursion. In contrast, RaP

these differences was assessed using a Monte Carlo simulation involving 50 samples of randomly selected files subsets consisting of half of the files annotated by two or more labelers from the corpus of labeled ToBI and RaP data and computing κ for each sample. The mean values of κ for ToBI and RaP based on this simulation were 0.707 and 0.773; this difference was significant in an independent samples *t*-test, *t*(98) = 14.982, *p* < .0001.

Next, agreement on the presence of a phrasal boundary can be compared in the two systems. For this metric, RaP agreement exceeds that of ToBI by 8% for the TAP metric (92% vs. 84%, respectively) and 0.26 for the chance-adjusted κ metric (.78 vs. .52, respectively). The statistical reliability of these differences was again assessed using a Monte Carlo simulation following the procedure outlined above. The mean values of κ for ToBI and RaP from simulations were 0.517 and 0.780; this difference was significant in an independent samples *t*-test, *t*(98) = 31.679, *p* < .0001.

Finally, with respect to size of a phrasal boundary, RaP agreement again exceeds ToBI agreement by 5% for TAP (86% vs. 81%, respectively) and 0.21 for κ (.68 vs .47, respectively). The statistical reliability of these differences was again assessed using the procedure outlined above. The mean values of κ for ToBI and RaP from simulations were 0.474 and 0.683; this difference was significant in an independent samples *t*-test, *t*(98) = 28.155, *p* < .0001. Comparably high agreement for RaP compared with ToBI is apparent when separately examining read vs. spontaneous speech across the two phrasal boundary agreement categories. Together, these comparisons highlight overall higher agreement for RaP in labeling prominence and phrasal boundaries; however,

distinguishes nonprominent and prominent syllables, where prominent syllables further may have a starred tone (i.e., a pitch excursion in the vicinity of a stressed syllable) or not; thus, the judgment of whether a syllable has a salient pitch excursion is separated from the judgment of whether the syllable is prominent.

interpretation of these agreement levels must be made cautiously due to possible order or practice effects.

In particular, one issue which must be considered in interpreting these agreement differences is the order of application of the two transcription systems. Since RaP was applied after ToBI, then higher agreement for RaP could arguably have resulted from a practice effect – greater overall proficiency with prosody labeling as time elapsed – rather than greater reliability for RaP *per se*. To investigate this possibility, we compared agreement for ToBI transcriptions annotated during the project's initial phase (26.7 min. of speech), with that for RaP transcriptions annotated during the second phase (19.2 minutes of speech) and for ToBI transcriptions for a smaller corpus in the project's third phase (9.4 minutes of speech) (Table 12).

Two aspects of agreement trends over time bear on the issue of whether greater labeling proficiency led to greater labeling agreement. The first is the extent to which agreement is different (i.e., higher) for the last phase of ToBI labeling (third phase) compared with the first phase. The results in Table 12 show no significant difference between ToBI agreement across these five measures for the first phase vs. the third phase under a two-tailed, paired-samples test for TAP values, $t(4) = .125$, $p = 0.91$, or for κ values, $t(4) = 1.39$, $p = 0.24$.

The second aspect of the data that bears on the issue of order of system application is the agreement trend across the three labeling periods. If the higher agreement for phrasal boundaries for RaP compared with ToBI is due to a general increase in labeling proficiency over time, then levels of agreement should either rise steadily across the three phases, or else asymptote at a high level across the second and third phases. As stated above, the three agreement categories for ToBI and RaP which can be directly compared are: (1) the prominent vs. non-prominent distinction, (2) the presence of a phrasal boundary, and (3) the size of a phrasal boundary. Inspection of these categories, listed in rows 1, 4, and 5 of Table 12, respectively, demonstrate no evidence of a consistent increase in proficiency level across the three time periods, nor is there evidence of an asymptote at a high level across the second and third phases. Instead, agreement is roughly flat over time for presence of pitch accent (row 1) for both TAP and κ; moreover, data for presence and size of phrasal boundary (rows 4 and 5) show the highest agreement for the second phase, i.e. RaP labeling. Therefore, this data pattern suggests that higher agreement on RaP for phrasal boundaries is not due entirely to an increase in general labeling proficiency over time.

It should be noted that there was an increase in phrasal boundary agreement for ToBI comparing the first and last time periods. While TAP values rise 3-5% in these measures, still larger increases are observed for κ. Inspection of the data revealed that the increase in κ was due almost entirely to a decrease chance agreement (A_E), rather than an increase in observed agreement (A_O). This finding indicates that labelers were using a more varied and equal distribution of break indices during the last ToBI phase compared with the first. The improved κ scores for ToBI in the third phase relative to the first thus reflects an increase in label diversity and may reflect an increase in general proficiency. However, any such apparent increase in proficiency is not so great as to wholly account for higher agreement for RaP on phrasal boundary characteristics.

Overall, the results of the first study demonstrate that both ToBI and RaP labelers achieved levels of agreement which were substantial and reliable across almost all categories examined. Notably, agreement for RaP was comparable to, or in some cases higher than, agreement for ToBI, which is considered the current standard for prosodic labeling.

While these results provide a useful quantitative benchmark of RaP agreement levels, it may be noted that the same corpus was used in the first two phases of the study, i.e. the same set of recordings was labeled first in ToBI, then in RaP. The increased familiarity with materials may have contributed to higher agreement in some categories for RaP than ToBI; however, it is unlikely that this factor is entirely responsible for absolute agreement levels observed for RaP during the second study phase. Nevertheless, a second agreement study was undertaken which provided strict control for possible practice effects, as well as ordering effects, while independently examining, across the two studies, the effect of degree of labeling proficiency on agreement in the two prosodic transcription systems. The observation of high agreement for ToBI and RaP with the imposition of these additional controls will provide additional evidence of the inter-transcriber reliability of these systems.

STUDY TWO

For Study Two, four expert labelers were recruited to label a new corpus of speech using both ToBI and RaP. This made possible an estimate of the effects of degree of labeling proficiency on agreement in the two systems (moderately high proficiency in Study 1 vs. expert in Study 2). Moreover, we counterbalanced both the order of speech each labeler annotated, as well as the order of application of the two annotation systems, thereby minimizing possible practice effects.

Method

Participants. Four labelers who were experts in both ToBI and RaP participated in the present study. Two were undergraduates from the first study, who continued to receive either course credit or monetary compensation at a rate of $8.75/hr for the duration of the project. The other two labelers were authors M.B. and L.D.

Materials. All materials used in the second study were new (i.e., they had not been annotated as part of the first study). A total of approximately six minutes of speech was selected for the study (178 seconds of read speech from the BRN corpus and 181 seconds of spontaneous speech from the CallHome corpus). The speech was from 7 talkers, roughly equally balanced between male (183 sec) and female talkers (177 sec), and contained a total of 1533 syllables and 1072 words.

Procedure. Each labeler annotated the entire corpus with both systems. The order of speech files and order of application of systems (ToBI vs. RaP) were counterbalanced for each labeler. This required each labeler to switch from coding in one system to coding in the other system at several points during the study. Prior to each switch, each labeler annotated one or more practice speech files in the new system and received feedback on his/her labels from L.D. Labelers annotated individually, and never discussed their labels at any point during the study.

Analyses. Agreement analyses were calculated as in Study One.

Results and Discussion

Agreement results are shown in Table 13. Considering first ToBI, inspection of TAP values (first column) reveals the highest agreement for presence of phrasal boundary, with presence of pitch accent and type of pitch accent (H vs. L) following closely; agreement for TAP values ranges from 80% to 91% across categories. Inspecting chance-corrected κ values (third column), agreement is substantial and reliable (i.e., κ > .60) for four out of five categories, with agreement highest in analyses of presence of pitch accent and presence of phrasal boundary, consistent with TAP values. Finally, agreement across categories was higher for read speech than for spontaneous speech for both agreement metrics. Tables 14 and 15 display confusion matrices of ToBI pitch accents and break indices from Study Two, respectively. In general, both tables show a substantial number of off-diagonal pairings on pitch accents and break indices and provide detail supplementing Table 12 about the specific nature of disagreements that arose in this study within the ToBI system.

Considering next RaP, inspection of TAP values (second column) shows the highest agreement in analyses of presence of beat and presence of phrasal boundary, with agreement ranging from 75% to 90% across categories. Inspecting chance-corrected κ values (fourth column), agreement is substantial and reliable for five of six categories, with agreement highest in analyses of presence of beat and presence of phrasal boundary, consistent with TAP values for RaP. Finally, note that agreement was higher for read speech than spontaneous speech for most agreement categories; however, for phrase boundary-related agreement categories, higher agreement was observed for spontaneous speech than for read speech. Tables 16, 17, and 18 display confusion matrices of RaP labels from Study Two. These tables reveal disagreement concerning whether a syllable is prominent and how prominent it is (Table 8), whether a syllable has a starred tone and what type (Table 9), and whether a phrasal boundary is present and what its size is (Table 10). Together, these tables provide significant detail to supplement Table 12 about the specific nature of disagreements that arose in this study within the RaP system.

To investigate variability in overall agreement across labelers, we also calculated the pairwise agreement between pairs of annotators in ToBI and in RaP, averaged across values of Kappa shown in Table 13. The result is shown in Table 19 For the pair of labelers represented by each cell, the first and second numbers indicate the average Kappa values for ToBI and for RaP, respectively. It can be observed that average pairwise Kappa agreement across labelers for ToBI ranges from .64 to .70, while for RaP agreement across labelers ranges from .58 to .69. Note that labelers L1 and L2 were authors MB and LD, respectively. It can be observed that these two labelers produced the highest ToBI agreement of any pair, but not the highest RaP agreement. This suggests that these authors contributed high agreement levels to ToBI, and furthermore, that overall high agreement levels for RaP were not merely a function of the authors' participation in the study.

Agreement can be compared across systems for this second study for the categories of prominent vs. non-prominent, presence of phrasal boundary, and size of phrasal boundary as in Study 1. Considering first the prominent vs. non-prominent distinction, agreement between the two systems is comparable under the TAP metric (ToBI: 88, RaP: 89) and slightly higher for RaP compared with ToBI under the κ metric (ToBI: .74, RaP: .78). The statistical reliability of these differences was assessed using a Monte Carlo simulation involving 50 random samples of eight of the 16 files annotated for ToBI and RaP and computing κ for each. The mean values of κ for ToBI and RaP based on this simulation were 0.738 and 0.783; this difference was significant in an independent samples *t*-test, $t(98) = 7.224$, $p < .001$.

Considering next presence of a phrasal boundary, agreement for ToBI and RaP is comparable under the TAP metric (ToBI: 91%, RaP: 90%) and slightly higher for ToBI than RaP under the κ metric (ToBI: .77, RaP: .75). The statistical reliability of these differences was again assessed using a Monte Carlo simulation following the procedure outlined above. The mean values of κ for ToBI and RaP from simulations were 0.767 and 0.765; this difference was not significant in an independent samples *t*-test, $t(98) = 0.336$, $p > .05$.

Finally, considering size of a phrasal boundary, agreement between ToBI and RaP is similar, with ToBI slightly higher agreement under both TAP (ToBI: 87%, RaP: 85%) and Kappa (ToBI: .68, RaP: .67) metrics. However, agreement is again mediated by style of speech, with RaP showing higher agreement than ToBI for spontaneous speech, with the opposite pattern for read speech. The statistical reliability of these differences was again assessed using the procedure outlined above. The mean values of κ for ToBI and RaP from simulations were 0.683 and 0.673; this difference was not significant in an independent samples *t*-test, *t*(98) = 1.92, *p* > .05.

Finally, we can consider how the agreement from Study 2 compares with that of Study 1. Inspection of values across categories from Tables 6 and 7 reveals that differences in agreement levels across the two studies are generally quite small, suggesting substantial reliability and consistency in the use of both systems over labelers and data sets. For example, if we consider all agreement categories except for those relating to phrasal boundaries, then the average difference (Study 2 – Study 1) in TAP values is +2% for ToBI and +1.5% for RaP, while the average difference in κ values is +.04 for ToBI and 0 for RaP. With respect to the two phrasal boundary agreement categories, however, substantial improvement is seen for ToBI for Study 2 compared with Study 1; the average difference for ToBI in TAP and κ values, respectively, is +6.5 and +0.23. In comparison, there is almost no difference in phrasal boundary agreement for RaP; the average difference for RaP in TAP and κ values, respectively, is -1.5 and -0.02.

The fact that the disparity in observed agreement on phrasal boundaries between ToBI and RaP in Study One disappeared in Study Two has two explanations. First, given that the primary difference between Studies One and Two was that the labelers were naïve and expert, respectively, in the use of the ToBI labeling system, the lower agreement observed for ToBI compared to RaP in phrasal boundaries for Study One likely reflects the generally higher proficiency of the expert labelers in identifying and classifying phrasal boundary events. Relatedly, the higher observed agreement for RaP in Study One was likely due to the fact that coders already had experience both with prosodic labeling, and with the speech they were labeling, as they had previously labeled it with ToBI. However, the consistent, high agreement for phrasal boundary agreement seen for RaP in Study 2 compared with Study 1 serves to rule out the possibility that high agreement observed for RaP in Study 1 was solely the result of practice or order effects. This is because quantitatively similar performance is observed in Study Two when controlling for these variables. These results therefore demonstrate that high agreement is achieved on labeling prosodic categories in RaP, replicating results across two studies using different labelers and data sets. In addition, these results replicate previous studies demonstrating high agreement using ToBI.

General Discussion

There were two main purposes of this paper. The first was to present an overview of the ToBI (Silverman, et al.,1992) and RaP annotation systems (Dilley & Brown, 2005), including background on the motivation for development of RaP as an alternative to ToBI. The second goal was to present large-scale inter-transcriber reliability data for ToBI and RaP, and in particular, to establish whether labelers could achieve agreement in RaP at a level comparable to ToBI.

Two inter-transcriber reliability studies were conducted using both naïve (Study One) and expert (Study Two) labelers. In Study One, naïve labelers learned and applied both systems to a varied speech corpus in the order *ToBI-RaP-ToBI*. In Study Two, labelers who were expert in both ToBI and RaP applied each system to a different, varied speech corpus, counterbalancing the order of application of the two systems. Across both studies, consistently high, substantial reliability was demonstrated for the RaP system. Critically for the present paper, levels of agreement were comparable to those for ToBI for all agreement comparisons; these comparisons represented a sampling of the most substantive aspects of each system. Importantly, the fact that quantitatively similar agreement was obtained in both studies across all categories indicates that (1) high agreement for ToBI and RaP replicates across speech materials, labelers, and study conditions; and (2) high agreement for ToBI and RaP demonstrated in Study 1 was not merely attributable to practice and/or order effects. In addition, the high agreement observed for RaP in both studies therefore serves to establish the reliability of this new prosody labeling system. Given that the RaP system was designed to address certain recognized weaknesses of ToBI (e.g., greater phonetic transparency, greater ease of learning and use, capability of labeling multiple levels of prominence, etc.), the present results suggest that RaP is an alternative to ToBI suitable for a variety of speech prosody research and technology applications.

We have proposed that the comparably high agreement for RaP and ToBI shown here constitutes a demonstration of the overall viability of the RaP system for prosody labeling as an alternative to ToBI. Moreover, the present assessments of ToBI agreement offered a number of strengths relative to those of previous studies, including more valid agreement metrics, larger corpora, and/or a greater number of coders. Still, it is important to consider whether the levels of agreement demonstrated in the present studies for ToBI are in line with those of previous studies, which is important for evaluating the generality of the present findings. Considering first agreement categories related to pitch accent, the prominent vs. non-prominent distinction yielded TAP values of 87% (Study 1) and 88%

(Study 2), and chance-corrected κ values of .71 (Study 1) and .74 (Study 2). These numbers compare well with those of Yoon et al. (2005) who reported a κ of .75 for the prominent vs. non-prominent (i.e., pitch-accented vs. unaccented) comparison in ToBI, as well as Pitrelli et al. (1994), who reported a TAP value of 81% for this metric, though are slightly lower than those of and Syrdal and McGory (2000), who reported raw agreement of 91-92%. Next, for the H vs. L type of pitch accent distinction, the observed κ values of .69 (Study 1) and .73 (Study 2) are considerably larger than the value of .51 reported by Yoon et al. (2004). The difference is mostly likely due to the fact that Yoon et al. used a small sample size (1594 words), which could have lead to a higher chance estimate in their sample than would be expected in a larger corpus. Finally, for the 6-way type of pitch accent distinction with all accents distinct, the TAP values of 77% (Study 1) and 80% (Study 2) are again higher than the 64% reported by Pitrelli et al. However, it is hard to know how to interpret these differences, since TAP values are not corrected for chance; thus, labelers in our study may simply have been using ToBI pitch accent categories with different relative frequencies than those in Pitrelli et al. (1994). Moreover, the fact that Pitrelli et al. collapsed L+H* and H* (and their downstepped counterparts), while we did not, would also be expected to affect agreement levels across the two studies.

Next, we compare agreement for phrasal boundary-related categories in the present studies to agreement for phrasal boundary categories in previous studies. Considering first agreement for the presence of a phrasal boundary (collapsing across phrase accents and boundary tones), we found TAP values of 84% (Study 1) and 91% (Study 2), as well as Kappa values of .52 (Study 1) and .77 (Study 2). These values are not out of line with previous studies: Yoon et al. (2004) reported κ = .67 for phrase accent presence, and κ = .58 for boundary tone presence, while Pitrelli et al. (1994)

reported 90% for phrase accent presence, 93% for boundary tone presence. Finally, Syrdal and McGory (2000) report raw agreement of 85% on phrase accents and boundary tones combined. The comparisons above demonstrate that the ToBI labelers in our studies were at least as proficient as those who contributed data to previous studies of ToBI agreement, and, therefore, that their ToBI labels can be validly compared to their RaP labels.

Overall, these results clearly establish that labelers achieve a substantial, comparable degree of reliability in the use of both the ToBI and RaP systems. Given that RaP was designed to provide an additional option permitting greater phonetic transparency of labels and greater ease of use than ToBI, the present findings help to establish RaP as a viable alternative to ToBI for a variety of corpus-based speech prosody research and technology applications.

References

ARVANITI, A., LADD, D. R., & MENNEN, I. (1998). Stability of tonal alignment: the case of Greek prenuclear accents. *Journal of Phonetics*, *26*, 3-25.

ARVANITI, A., LADD, D. R., & MENNEN, I. (2000). What is a starred tone? Evidence from Greek. In *Papers in Laboratory Phonology V* (pp. 119-130): Cambridge University Press.

BARTELS, C., & KINGSTON, J. (1994). Salient pitch cues in the perception of contrastive focus. In P. Bosch & R. van der Sandt (Eds.), *Focus and Natural Language Processing: Proceedings of the Journal of Semantics Conference on Focus*: IBM Working Papers, TR-80.94-006.

BEAVER, D. I., CLARK, B. Z., FLEMMING, E. S., JAEGER, T. F., & WOLTERS, M. K. (2007). When semantics meets phonetics: Acoustical studies of second occurence focus. *Language*.

BECKMAN, M., & AYERS ELAM, G. (1997). Guidelines for ToBI labeling, version 3: Ohio State University.

BECKMAN, M., & HIRSCHBERG, J. (1994). The ToBI annotation conventions. The Ohio State University and AT&T Bell Laboratories, unpublished manuscript. Available at ftp://ftp.ling.ohio-state.edu/pub/phonetics/TOBI/ToBI/ToBI.6.html.

BECKMAN, M., HIRSCHBERG, J., & SHATTUCK-HUFNAGEL, S. (2005). The original ToBI system and the evolution of the ToBI framework. In S.-A. Jun (Ed.), *Prosodic Typology: The Phonology of Intonation and Phrasing* (pp. 9-54): Oxford University Press.

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

BRAUN, B. (2006). Phonetics and phonology of thematic contrast in German. *Language and Speech, 49*(4), 451-493.

BRUGOS, A., SHATTUCK-HUFNAGEL, S., & VEILLEUX, N. (2006). Transcribing prosodic structure of spoken utterances with ToBI (MIT open course ware), http://ocw.mit.edu.

BRUGOS, A., VEILLEUX, N., BREEN, M., & SHATTUCK-HUFNAGEL, S. (2008). The Alternatives (Alt) tier for ToBI: advantages of capturing prosodic ambiguity. In *Proceedings of Speech Prosody*. Campinas, Brazil.

CALHOUN, S. (2006). *Information structure and the prosodic structure of English: A probabilistic relationship.* Unpublished Ph.D. dissertation, University of Edinburgh.

CARLETTA, J. (1996). Assessing agreement on classification tasks: the Kappa statistic. *Computational Linguistics, 22*(2), 249-254.

CHOI, J.-Y., HASEGAWA-JOHNSON, M., & COLE, J. (2005). Finding intonational boundaries using acoustic cues related to the voice source. *Journal of Acoustical Society of America, 118*(4), 2579-2587.

CUTLER, A., & NORRIS, D. G. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance, 14*, 113-121.

DILLEY, L. C. (2005). *The phonetics and phonology of tonal systems.* Unpublished Ph.D. dissertation, MIT, Cambridge, MA.

DILLEY, L. C. (2007). Pitch range variation in English tonal contrasts: Continuous or categorical? In *Proceedings of the International Congress of Phonetic Sciences*. Saarbruecken, Germany.

DILLEY, L. C. (2010). Pitch range variation in English tonal contrasts is continuous, not categorical. *Phonetica, 67*, 63-81.

DILLEY, L. C. (submitted). The role of F0 alignment in distinguishing categories in American English intonation. *Journal of Phonetics*.

DILLEY, L. C., BREEN, M., BOLIVAR, M., KRAEMER, J., & GIBSON, E. (2006). A comparison of inter-transcriber reliability for two systems of prosodic annotation: RaP (Rhythm and Pitch) and ToBI (Tones and Break Indices). In *Proceedings of Interspeech 2006*. Pittsburgh.

DILLEY, L. C., & BROWN, M. (2005). The RaP (Rhythm and Pitch) Labeling System, Version 1.0: Available at http://tedlab.mit.edu/rap.html.

DILLEY, L. C., & BROWN, M. (2007). Effects of pitch range variation on F0 extrema in an imitation task. *Journal of Phonetics, 35*, 523-551.

DILLEY, L. C., LADD, D. R., & SCHEPMAN, A. (2005). Alignment of L and H in bitonal pitch accents: Testing two hypotheses. *Journal of Phonetics, 33*(1), 115-119.

DUEZ, D. (1993). Acoustic correlates of subjective pauses. *Journal of Psycholinguistic Research, 22*, 21-39.

GOLDSMITH, J. (1976). *Autosegmental phonology*. Unpublished Ph.D. dissertation, MIT, Cambridge, MA.

GREENBERG, S., CARVEY, H., & HITCHCOCK, L. (2002). The relationship between stress accent and pronunciation variation in spontaneous American English discourse. In *Proceedings of the ISCA Workshop on Prosody and Speech Processing* (pp. 56-61).

HALLIDAY, M. A. K. (1967). *Intonation and Grammar in British English*. Paris: Mouton.

HIRST, D., & DI CRISTO, A. (1998). *Intonation systems. A survey of twenty languages*. Cambridge: Cambridge University Press.

JUN, S.-A. (Ed.). (2005). *Prosodic typology: The phonology of intonation and phrasing*. Oxford: Oxford University Press.

KOCHANSKI, G., GRABE, E., COLEMAN, J., & ROSNER, B. (2005). Loudness predicts prominence: fundamental frequency lends little. *Journal of Acoustical Society of America, 118*, 1038-1054.

KRIPPENDORFF, K. (1980). *Content analysis: An introduction to its methodology*: Sage Publications.

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

LADD, D. R. (2000). Tones and turning points: Bruce, Pierrehumbert, and the elements of intonational phonology. In M. Horne (Ed.), *Prosody: Theory and Experiment - Studies presented to Gosta Bruce* (pp. 37-50). Dordrecht: Kluwer.

LADD, D. R. (2008). *Intonational Phonology* (2nd ed.). Cambridge: Cambridge University Press.

LADD, D. R. (2009). *Intonational Phonology* (2nd ed.). Cambridge: Cambridge University Press.

LADD, D. R., FAULKNER, D., FAULKNER, H., & SCHEPMAN, A. (1999). Constant "segmental anchoring" of F0 movements under changes in speech rate. *Journal of the Acoustical Society of America, 106*(3), 1543-1554.

LADD, D. R., MENNEN, I., & SCHEPMAN, A. (2000). Phonological conditioning of peak alignment in rising pitch accents in Dutch. *Journal of the Acoustical Society of America, 107*(5), 2685-2696.

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

LANDIS, J., & KOCH, G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

LIBERMAN, M. (1975). *The intonation system of English.* Unpublished Ph.D. dissertation, MIT, Cambridge, MA.

NAZZI, T., & RAMUS, F. (2003). Perception and acquisition of linguistic rhythm by infants. *Speech Communication, 41*, 233-243.

PIERREHUMBERT, J. (1980). *The phonology and phonetics of English intonation.* Unpublished Ph.D. dissertation, MIT, Cambridge, MA.

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

PIERREHUMBERT, J., & HIRSCHBERG, J. (1990). The meaning of intonational contours in the interpretation of discourse. In P. Cohen, J. Morgan & M. Pollack (Eds.), *Intentions in Communications* (pp. 271-311). Cambridge, MA: MIT Press.

PIERREHUMBERT, J., & STEELE, S. A. (1989). Categories of tonal alignment in English. *Phonetica, 46*, 181-196.

PITRELLI, J. F., BECKMAN, M., & HIRSCHBERG, J. (1994). Evaluation of prosodic transcription labeling reliability in the ToBI framework. In *Proceedings of the International Conference on Spoken Language Processing* (pp. 123-126).

PRICE, P. J., OSTENDORF, M., SHATTUCK-HUFNAGEL, S., & FONG, C. (1991). The use of prosody in syntactic disambiguation. *Journal of the Acoustical Society of America, 90*(6), 2956-2970.

REDI, L. C. (2003). Categorical effects in production of pitch contours in English. In *Proceedings of the 15th International Congress of the Phonetic Sciences* (pp. 1647-1650). Barcelona.

RIETVELD, T., & VAN HOUT, R. (1993). *Statistical techniques for the study of language and language behavior*: Mouton de Gruyter.

ROUAS, J.-L., FARINAS, J., PELLIGRINO, F., & ANDRE-OBRECHT, R. (2003). Modeling prosody for language identification on read and spontaneous speech. In *International conference on Multimedia and Expo* (pp. 753-756).

SALVERDA, A. P., DAHAN, D., & MCQUEEN, J. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition, 90*, 51-89.

SHATTUCK-HUFNAGEL, S., DILLEY, L. C., VEILLEUX, N., BRUGOS, A., & SPEER, R. (2004). F0 peaks and valleys aligned with non-prominent syllables can influence perceived prominence in adjacent syllables. In *Proceedings of Speech Prosody 2004*. Nara, Japan.

SILVERMAN, K., BECKMAN, M., PIERREHUMBERT, J., OSTENDORF, M., WIGHTMAN, C. W. S., PRICE, P., et al. (1992). ToBI: A standard scheme for labeling prosody. In *Proceedings of the 2nd International Conference on Spoken Language Processing* (pp. 867-879). Banff.

SILVERMAN, K., & PIERREHUMBERT, J. (1990). The timing of prenuclear high accents in English. In J. Kingston & M. Beckman (Eds.), *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech* (pp. 71-106). Cambridge: Cambridge University Press.

SYRDAL, A. K., & MCGORY, J. (2000). *Inter-transcriber reliability of ToBI prosodic labeling.* Paper presented at the International Conference on Spoken Language Processing, Beijing, China.

WATSON, D. G., ARNOLD, J. A., & TANENHAUS, M. K. (2008). Tic Tac TOE: Effects of predictability and importance on acoustic prominence in language production. *Cognition, 106*, 1548-1557.

WATSON, D. G., TANENHAUS, M. K., & GUNLOGSON, C. A. (2008). Interpreting pitch accents in on-line comprehension: H* vs. L+H*. *Cognitive Science, 32*, 1232-1244.

WIGHTMAN, C. W. (2002). ToBI or not ToBI? In *Proceedings of Speech Prosody 2002*. Aix-en-Provence, France.

WIGHTMAN, C. W., SHATTUCK-HUFNAGEL, S., OSTENDORF, M., & PRICE, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America, 91*(3), 1707-1717.

YOON, T.-J., CHAVARRIA, S., COLE, J., & HASEGAWA-JOHNSON, M. (2004). *Intertranscriber reliability of prosodic labeling on telephone conversation using ToBI.* Paper presented at the Interspeech 2004 Conference, Jeju Island, Korea.

Label Type	Intended to capture	**ToBI**	**RaP**
Metrical RaP: Rhythm tier ToBI: N/A	strong beat:	N/A	X, X?
	weak beat:	N/A	x, x?
	no beat:		no label
Tonal ToBI: Tones tier RaP: Pitch tier	Prominent and/or nonprominent syllables:	H*, L+H*, H+!H*, !H*, L+!H* L*, L+H*, L*+H	H*, L*, E*, H, L, E
	Major boundary:	L-L%, H-H%, L-H%, H-L%, !H-L%	H, L, E
	Minor boundary:	L-, H-, !H-	
Phrasal ToBI: Break index tier RaP: Rhythm tier	Major boundary:	4	)),))?
	Minor boundary:	3	),)?
	No boundary:	2, 1, 0	no label

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

System	Corpus	Minutes	Syllables	Labelers/File	Unique Speakers
ToBI	CallHome	15.2	3680	3.5	6
	BRNC	20.9	5939	3.4	6
RaP	CallHome	9.6	2638	3.8	6
	BRNC	9.6	2889	3.8	6
	Total	55.2	15146		12

Table 2. Amount of speech (in minutes and syllables) from each corpus annotated in each system during Study One, including number of labelers per file. Speakers are different for the two corpora, but are the same for both ToBI and RaP-annotated files.

<table>
<tr><th></th><th colspan="6">ToBI</th><th colspan="12">RaP</th></tr>
<tr><th rowspan="2">Prominence v. non-prominence</th><th colspan="3">Prominent (i.e., Pitch Accented)</th><th colspan="3">Nonprominent (i.e., Unaccented)</th><th colspan="6">Prominent</th><th colspan="6">Non-prominent</th></tr>
<tr><td colspan="3">(!)H*, L+(!)H*, L*, L*+(!)H, H+!H*</td><td colspan="3">*? or no label</td><td colspan="6">X, X?, x</td><td colspan="6">x?, no label</td></tr>
<tr><th rowspan="2">Level of metrical prominence</th><td colspan="6" rowspan="2">N/A</td><th colspan="4">Strong</th><th colspan="4">Moderate</th><th colspan="4">None</th></tr>
<tr><td colspan="4">X, X?</td><td colspan="4">x</td><td colspan="4">x?, no label</td></tr>
<tr><th rowspan="2">Pitch accent status</th><td colspan="6" rowspan="2">N/A</td><th colspan="6">RaP Starred Tone Present</th><th colspan="6">RaP Starred Tone Absent</th></tr>
<tr><td colspan="6">H*(?), E*(?), L*(?), !H*(?), !L*(?)</td><td colspan="6">H, L, E, no label</td></tr>
<tr><th rowspan="2">Type of pitch accent (High vs. Low)</th><th colspan="2">High accent</th><th colspan="2">Low accent</th><th colspan="2">No accent</th><td colspan="12" rowspan="2">N/A</td></tr>
<tr><td colspan="2">(!)H*, H+!H*, L+(!)H*, X*</td><td colspan="2">L*, L*+(!)H*</td><td colspan="2">No label or *?</td></tr>
<tr><th rowspan="2">Type of pitch accent – All distinct</th><th>H*</th><th>L*</th><th>L+H*</th><th>L*+H</th><th>H+!H*</th><th>No accent</th><th colspan="3">High</th><th colspan="3">Low</th><th colspan="3">Equal</th><th colspan="3">None</th></tr>
<tr><td>H*, !H*</td><td>L*</td><td>L+H*, L+!H*</td><td>L*+H, L*+!H*</td><td>H+!H*</td><td>*? or no label</td><td colspan="3">(!)H*</td><td colspan="3">(!)L*</td><td colspan="3">E*</td><td colspan="3">No label or *?</td></tr>
<tr><th rowspan="2">Phrasal boundary present†</th><th colspan="3">Present</th><th colspan="3">Absent</th><th colspan="6">Present</th><th colspan="6">Absent</th></tr>
<tr><td colspan="3">4, 4?, 4-, 3, 3?, 3-</td><td colspan="3">2, 2?, 2-, 1, 1?, 1-, 0</td><td colspan="6">)),))?,),)?</td><td colspan="6">No label</td></tr>
<tr><th rowspan="2">Size of phrasal boundary†</th><th colspan="2">Large</th><th colspan="2">Medium</th><th colspan="2">No boundary</th><th colspan="4">Large</th><th colspan="4">Medium</th><th colspan="4">None</th></tr>
<tr><td colspan="2">4, 4?, 4-</td><td colspan="2">3, 3?, 3-</td><td colspan="2">2, 2?, 2-, 1, 1?, 1-, 0</td><td colspan="4">)),))?</td><td colspan="4">),)?</td><td colspan="4">No label</td></tr>
</table>

Table 3. Label categories (i.e., label-equivalence relations) for ToBI and RaP over which raw and chance-corrected agreement are calculated. The dagger (†) in the first column indicates that the transcriber-agreement-pair metric was based on agreement for an entire word, while absence of a dagger indicates that the transcriber-agreement-pair metric was based on individual syllables. A parenthesis around a label indicates that the label was ignored for the purposes of defining an agreement category.

Kappa (κ) value	Implications for reliability
$0 < \kappa \leq 0.40$	Unreliable distinction
$0.40 < \kappa \leq 0.60$	Questionably reliable distinction with only moderate agreement
$0.60 < \kappa \leq 0.80$	Reliable distinction with substantial agreement
$0.80 < \kappa \leq 1.0$	Highly reliable distinction

Table 4. Stratification of kappa values and implications for reliability based on Landis & Koch (1977), Rietveld & van Hout (1993), and Syrdal and McGory (2000).

	TAP (%)		**κ**	
Agreement Analysis	**ToBI**	**RaP**	**ToBI**	**RaP**
Prominence vs. non-prominence	**87** (89,84)	**89** (91,86)	**.71** (.75,.64)	**.77** (.82,.73)
Level of metrical prominence	N/A	**77** (79,74)	N/A	**.61** (.65,.57)
Pitch accent status	**N/A**	**85** (87,82)	**N/A**	**.68** (.73,.62)
Type of pitch accent: H vs. L	**86** (88,83)	N/A	**.69** (.74,.63)	N/A
Type of pitch accent: All distinct	**77** (78,75)	**72** (77,68)	**.54** (.56,.50)	**.54** (.60,.46)
Phrasal boundary present	**84** (83,85)	**92** (92,91)	**.52** (.50,.54)	**.78** (.79,.77)
Size of phrasal boundary	**81** (80,82)	**86** (87,85)	**.47** (.46,.49)	**.68** (.71,.66)

Table 5. Study One agreement results for RaP and ToBI across the entire study period. The left two columns indicate raw percent agreement using the TAP metric; the right two columns indicate kappa values. The bold number indicates overall agreement, while the numbers in parentheses indicate the agreement results for read and spontaneous speech, respectively. See the text for descriptions of agreement analyses.

	no label	H*	L*	L+H*	L*+H	!H*	H+!H*	L*+!H	L+!H*
no label	*23239								
H*	2351	*3720							
L*	155	60	*33						
L+H*	402	2099	12	*570					
L*+H	43	29	2	21	*0				
!H*	1590	871	117	245	19	*1101			
H+!H*	498	383	25	45	9	403	*317		
L*+!H	5	0	0	0	0	3	2	*0	
L+!H*	112	226	3	88	2	248	23	0	*53

Table 6. Confusion matrix for pitch accent choice in ToBI for Study One. Asterisks indicate category agreement.

	0	1-	1	1p	2-	2	2p	3-	3	3p	4-	4
0	*0											
1-	0	*0										
1	30	0	*17603									
1p	2	0	134	*160								
2-	0	0	1	2	*0							
2	0	0	864	17	0	*46						
2p	0	0	114	117	0	21	*42					
3-	0	0	20	0	0	2	0	*0				
3	0	0	2204	27	0	214	39	13	*589			
3p	0	0	93	71	0	17	173	0	50	*253		
4-	0	0	21	0	0	2	0	0	21	2	*0	
4	0	0	1379	36	0	37	34	9	551	143	20	*1772

Table 7. Confusion matrix for break index selections in ToBI for Study One.

	no label	x?	x	X?	X
no label	*14222				
x?	170	*2			
x	2471	134	*3762		
X?	305	24	1788	*491	
X	300	12	1741	1472	*2366

Table 8. Confusion matrix for metrical prominence (i.e., beat) selections in RaP for Study One.

	no label	H*	E*	L*	!H*	!L*
no label	*15337					
H*	1106	*3704				
E*	606	235	*187			
L*	1235	257	225	*1111		
!H*	785	1203	277	229	*592	
!L*	786	181	225	526	193	*277

Table 9. Confusion matrix for starred tone (i.e., pitch accent) selections in RaP for Study One.

	no label	)?	)	))?	))
no label	*13853				
)?	428	*26			
)	1066	185	*570		
))?	184	37	505	*188	
))	236	13	337	480	*2385

Table 10. Confusion matrix for phrasal boundary selections in RaP for Study One.

	L1	L2	L3
L2	.52, .63		
L3	.51, .65	.62, .65	
L4	.52, .64	.60, .62	.59, .65

Table 11. Pairwise agreement between labelers in Study One. For the pair of labelers represented by each cell, the first and second numbers indicate the average Kappa values for ToBI and RaP, respectively, averaged across agreement categories in Table X.

	First phase: ToBI		**Second phase: RaP**		**Third phase: ToBI**	
Agreement Class	**TAP (%)**	**κ**	**TAP (%)**	**κ**	**TAP (%)**	**κ**
Prominent vs. non-prominent	87	0.71	89	0.77	85	0.69
Type of pitch accent: High vs. Low	86	0.69	N/A	N/A	84	0.67
Type of pitch accent: All accents distinct	77	0.53	72	0.54	74	0.53
Presence of phrasal boundary	83	0.49	92	0.78	88	0.72
Size of phrasal boundary	81	0.45	86	0.68	84	0.66

Table 12: Agreement for Study One on ToBI corpus annotated before and after training on RaP.

	TAP (%)		**κ**	
Agreement Class	**ToBI**	**RaP**	**ToBI**	**RaP**
Prominent vs. non-prominent	**88** (91,85)	**89** (91,87)	**.74** (.79,.68)	**.78** (.80,.74)
Degree of prominence	N/A	**79** (80,78)	N/A	**.63** (.63,.61)
Starred tone vs. no starred tone	**N/A**	**86** (87,83)	**N/A**	**.67** (.71.60)
Type of pitch accent: H vs. L	**88** (91,85)	**N/A**	**.73** (.79,.67)	**N/A**
Type of pitch accent: All distinct	**80** (82,77)	**75** (78,70)	**.60** (.64,.55)	**.51** (.57,.42)
Presence of phrasal boundary	**91** (92,90)	**90** (89,92)	**.77** (.80,.73)	**.75** (.71,.80)
Size of phrasal boundary	**87** (89,85)	**85** (84,86)	**.68** (.71,.64)	**.67** (.65,.70)

Table 13. Agreement results from Study Two. The left two columns indicate raw percent agreement using the transcriber agreement pairs metric; the right two columns indicate kappa values. The bold number indicates overall agreement, while the numbers in parentheses indicate the agreement results for read and spontaneous speech, respectively. See the text for descriptions of agreement analyses.

	no label	H*	L*	L+H*	L*+H	!H*	H+!H*	L*+!H	L+!H*
no label	*5403								
H*	544	*849							
L*	39	13	*6						
L+H*	67	386	1	*360					
L*+H	1	0	0	0	*0				
!H*	284	133	21	43	2	*218			
H+!H*	95	104	3	16	0	104	*93		
L*+!H	0	0	0	0	0	0	0	*0	
L+!H*	11	32	1	36	0	45	2	0	*31

Table 14. Confusion matrix for pitch accent choice in ToBI for Study Two.

	0	1-	1	1p	2-	2	2p	3-	3	3p	4-	4
0	*0											
1-	0	*0										
1	23	0	*4194									
1p	0	0	11	*10								
2-	0	0	13	0	*0							
2	1	0	173	1	1	*21						
2p	0	0	30	14	0	7	*21					
3-	0	0	47	0	0	9	3	*3				
3	0	0	253	2	1	40	18	57	*266			
3p	0	0	7	5	0	3	46	2	20	*34		
4-	0	0	12	0	0	1	3	3	74	0	*17	
4	0	0	70	4	0	1	26	5	140	44	59	*637

Table 15. Confusion matrix for break index selections in ToBI for Study Two.

	no label	x?	x	X?	X
no label	*4454				
x?	123	*7			
x	523	84	*829		
X?	181	24	692	*355	
X	74	2	144	381	*470

Table 16. Confusion matrix for metrical prominence (i.e., beat) selections in RaP for Study Two.

	no label	H*	E*	L*	!H*	!L*
no label	*5060					
H*	286	*719				
E*	174	75	*60			
L*	238	37	42	*114		
!H*	264	332	83	36	*160	
!L*	244	48	48	140	63	*126

Table 17. Confusion matrix for starred tone (i.e., pitch accent) selections in RaP for Study Two.

	no label	)?	)	))?	))
no label	*3868				
)?	164	*12			
)	347	73	*189		
))?	49	14	118	*40	
))	83	6	103	98	*785

Table 18. Confusion matrix for phrasal boundary selections in RaP for Study Two.

	L1	L2	L3
L2	.70, .66		
L3	.63, .61	.68, .69	
L4	.64, .58	.64,.62	.68, .64

Table 19. Pairwise agreement between labelers in Study Two. For the pair of labelers represented by each cell, the first and second numbers indicate the average Kappa values for ToBI and RaP, respectively, averaged across agreement categories in Table X. Labelers L1 and L2 were authors MB and LD, respectively.

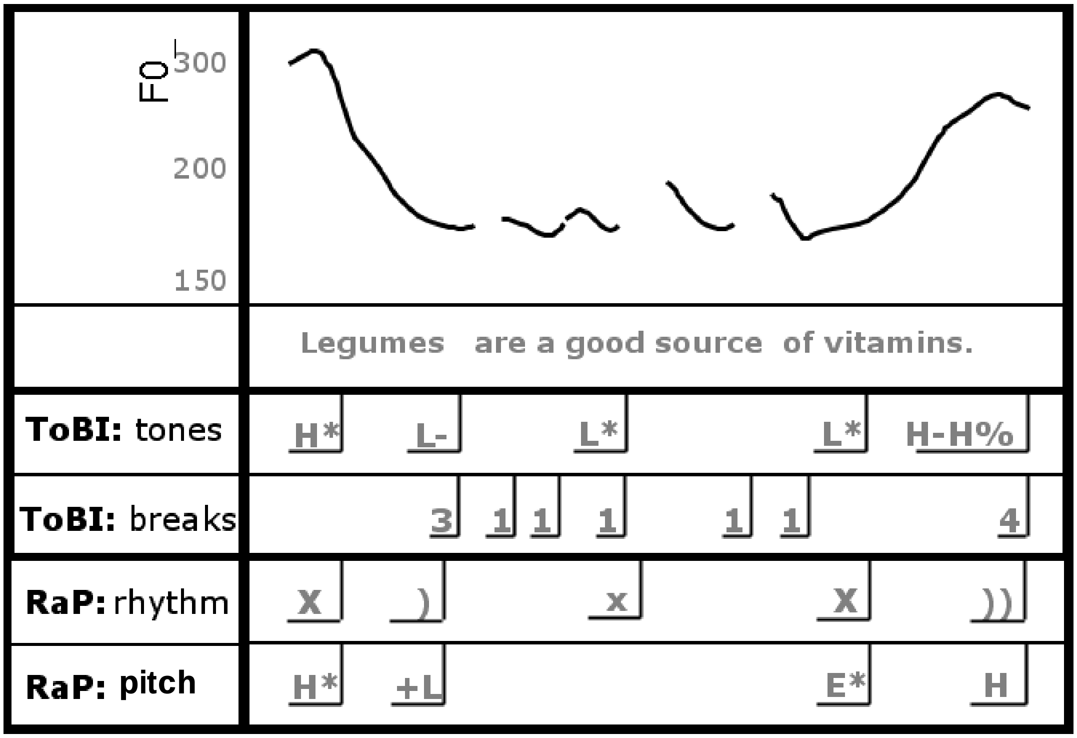


Figure 1. ToBI and RaP annotations for the same production of the sentence *Legumes are a good source of vitamins*. The critical differences in the annotations are: 1) In the ToBI annotation, *good* in labeled with a low pitch accent (L*); in RaP, this syllable is indicated as metrically prominent (x), but not pitch-accented, as it is not the locus of an F0 change; 2) In the ToBI annotation, the prominence on *vi-* is labeled with a low pitch accent (L*) because it is a prominence in the low part of the speaker’s range; in RaP, it is labeled as an equal pitch accent (E*), because it marks a prominence at the locus of a change from an equal to a rising F0. See text for more information.

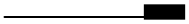
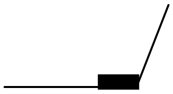
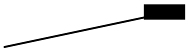
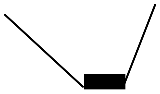
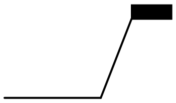
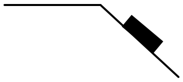
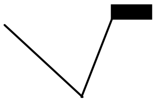
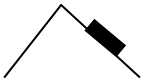
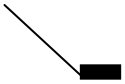
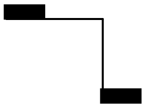
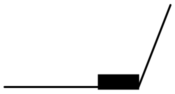
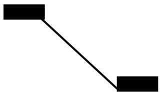
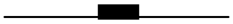
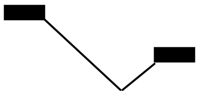
ToBI	RaP	Contour	ToBI	RaP	Contour
H*	:E+ E*		L*+H	:E+ E* +H	
H*	:!L+ !H*		L*+H	:H+ L* +H	
L+H*	:E E+ H*		H+!H*	:E E+ L*	
L+H*	:H L+ H*		H+!H*	:L H+ L*	
L*	:H+ L*		H* !H*	:E* E+ L*	
L*	:E+ E* +H		H* !H*	:H* L*	
L*	:E E		H* !H*	:H* L+ H*	

Figure 2. Comparison of select tone label sequences for ToBI and RaP.