Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 00 (2017) 1–26

www.elsevier.com/locate/csl

Crowd-sourcing prosodic annotation[☆]

Jennifer Cole[a,b,*], Timothy Mahrt[c], Joseph Roy[a]

[a] *University of Illinois at Urbana-Champaign, Department of Linguistics 4080 Foreign Language Building, 707 S Mathews Avenue, MC-168, Urbana, Illinois 61801, USA*

[b] *Northwestern University, Department of Linguistics, 2016 Sheridan Road Evanston, Illinois 60208, USA*

[c] *Laboratoire Parole et Langage, UMR 7309 CNRS, Aix-Marseille Université, 5 avenue Pasteur BP 80975, Aix-en-Provence 13604, France*

Received 30 August 2016; received in revised form 13 February 2017; accepted 14 February 2017

Abstract

Much of what is known about prosody is based on native speaker intuitions of idealized speech, or on prosodic annotations from trained annotators whose auditory impressions are augmented by visual evidence from speech waveforms, spectrograms and pitch tracks. Expanding the prosodic data currently available to cover more languages, and to cover a broader range of unscripted speech styles, is prohibitive due to the time, money and human expertise needed for prosodic annotation. We describe an alternative approach to prosodic data collection, with coarse-grained annotations from a cohort of untrained annotators performing rapid prosody transcription (RPT) using LMEDS, an open-source software tool we developed to enable large-scale, crowd-sourced data collection with RPT. Results from three RPT experiments are reported. The reliability of RPT is analysed comparing kappa statistics for lab-based and crowd-sourced annotations for American English, comparing annotators from the same (US) versus different (Indian) dialect groups, and comparing each RPT annotator with a ToBI annotation. Results show better reliability for same-dialect annotators (US), and the best overall reliability from crowd-sourced US annotators, though lab-based annotations are the most similar to ToBI annotations. A generalized additive mixed model is used to test differences among annotator groups in the factors that predict prosodic annotation. Results show that a common set of acoustic and contextual factors predict prosodic labels for all annotator groups, with only small differences among the RPT groups, but with larger effects on prosodic marking for ToBI annotators. The findings suggest methods for optimizing the efficiency of RPT annotations. Overall, crowd-sourced prosodic annotation is shown to be efficient, and to rely on established cues to prosody, supporting its use for prosody research across languages, dialects, speaker populations, and speech genres.

Keywords: Prosody; Annotation; Crowd-sourcing; Generalized mixed effects model; Inter-rater reliability; Speech transcription

1. Introduction

Investigations into the prosody of spoken languages—whether done in the service of describing languages, theorizing about language structure, or modelling spoken language processing—rely on the analysis of prosodic data, which comes very often in the form of prosodic annotation. Important early discoveries about prosody, such as the role of phrasal prominence in marking information structure in English (Bolinger, 1954; Halliday, 1967; Chafe,

[☆] This paper has been recommended for acceptance by Prof. R. K. Moore.

[*] Corresponding author at: Northwestern University, Department of Linguistics, 2016 Sheridan Road Evanston, Illinois 60208, USA.
E-mail address: jennifer.cole1@northwestern.edu (J. Cole), timmahrt@gmail.com (T. Mahrt), jroy042@illinois.edu (J. Roy).

http://dx.doi.org/10.1016/j.csl.2017.02.008
0885-2308/ 2017 Elsevier Ltd. All rights reserved.

1987), drew on data in the form of native speaker intuitions of idealized speech. Other work, including most current research, examines recordings of elicited or spontaneous speech for which prosodic annotations are produced by trained annotators based on auditory impression alone (e.g., Crystal, 1969; Bolinger, 1982), or augmented by visual evidence from pitch tracks, waveforms, and spectrogram displays (e.g., Bolinger, 1958; Ladd, 1980; Pierrehumbert, 1980; Gussenhoven, 1984; Wightman et al., 1992; Grabe and Post, 2002; Calhoun et al., 2010).

Prosodic annotations of recorded speech have many obvious advantages over the more purely subjective and impressionistic annotations of earlier work. For instance, the recordings can be submitted to multiple independent annotators, with inter-annotator agreement rates offering a measure of the reliability of the annotation (Pitrelli et al., 1994; Yoon et al., 2004; Breen et al., 2012). In addition, and most obviously, the presence of an audio recording means that acoustic correlates of the prosodic features labelled by annotators can be measured and identified. The distribution of these acoustic cues can then be examined to assess the contrastive status of the annotated prosodic features, and to identify systematic patterns of contextual variation in the phonetic expression of those features. Furthermore, prosodically annotated recorded utterances are also useful as stimuli for research on the perception of prosodic features and their influence on sentence and discourse comprehension.

Unfortunately, the advantages of working with prosodic annotations of recorded speech are available only for languages and dialects for which there exists a prosodic annotation standard, an available supply of trained annotators, and the necessary resources of money and time to perform the annotation and its validation by means of reliability analysis. In practice, these requirements have restricted prosody research primarily to the “big” languages of the world, i.e., those that have the support of a large community of researchers with access to research funding, and to the standard varieties for which annotation systems have been developed. Thus, in comparison to the growing body of prosody research on e.g., standard and regional varieties of Dutch, English, French, German, Italian, Japanese, Portuguese, and Spanish, there remains scant research on the vast majority of languages, notably, for most of the smaller, “under-resourced” languages and for non-standard and L2 varieties, but also for some languages with large speaker populations and a body of linguistic scholarship, such as Arabic and Russian.[1]

Here we present *rapid prosody transcription* (RPT) as an alternative methodology for prosodic annotation, one that sidesteps the limitations of traditional annotation methods by using untrained annotators in place of trained experts, and coarse-grained prosodic features in place of a larger and more detailed feature inventory (Mo et al., 2008). The simplicity of an RPT annotation, deriving from its use of only two binary features, one for prominence and one for prosodic phrase boundaries, is offset by more nuanced distinctions that are revealed when annotations of the same speech materials are aggregated over a group of annotators. Differences among annotators in their rating of words as prominent or as preceding a prosodic boundary reveal complex patterns of association between prosodic features and the cues to these features that are present in the speech signal and in the broader linguistic context of the utterance (Cole et al., 2010a, 2010b)

As described in more detail below, RPT requires no training or special knowledge of prosodic theory, and RPT annotation tasks can be performed without supervision. RPT is not the first annotation method to rely on “nav̈e” annotators; similar methods have been used to obtain prosodic ratings/judgements for a variety of research interests in prior work (de Pijper and Sanderman, 1994; Swerts, 1997; Streefkerk et al., 1997, 1998; Buhmann et al., 2002; Wagner, 2005). The use of untrained annotators confers an advantage in that RPT can be performed outside the research laboratory, with speech materials presented via audio files that are accessed online, and annotations entered digitally and relayed to the researcher through the internet. Consequently, RPT annotators can be recruited from any location with internet access. These properties of RPT enable its use with any language variety and any genre of speech for which an orthographic transcript can be produced. Annotators can be recruited online through crowd-sourcing platforms or other internet resources, allowing researchers to investigate prosody through the lens of annotation data from a much larger sample of the language community. (Hasegawa-Johnson et al., 2015).

This paper reports on three large-scale RPT studies, one using RPT in a lab setting, and two using RPT deployed over the internet with annotators recruited through a crowd-sourcing platform. Our goal in this paper is to evaluate

[1] This is not to imply that there are no prosodic analyses of languages and varieties outside the privileged group that includes standard, L1 English. Prosodic analyses of other languages and varieties are few but are also gaining in number in the literature, critically, as supported by the development of prosodic annotation standards for those languages (Gussenhoven, 2004; Jun, 2006, 2014) and by guidelines for prosodic field work (Jun and Fletcher, 2014; Arvaniti, 2016).

the merit of RPT annotations obtained through crowd-sourcing for scientific and technological research dealing with prosody. Here we formulate two broad questions:

(1) Are RPT annotations reliable, in the sense that annotations produced by one group of annotators replicate with a different group of annotators?
(2) Which acoustic cues and/or what contextual information do RPT annotators use in assigning prosodic features?

Prosodic annotations are a valuable source of research data only to the extent that the annotations are reliable. Annotated data are considered reliable to the degree that multiple transcribers, trained under the same annotation procedure but working independently, agree on the label assigned to each linguistic unit (e.g., word). Agreement among annotators suggests “that they have internalized a similar understanding of the annotation guidelines, and we can expect them to perform consistently under this understanding.” (Artstein and Poesio 2008: 557). Prosodic annotations schemes have been subject to reliability testing on a limited scale, primarily for annotations done by highly trained and linguistically informed annotators, in a laboratory setting (Pitrelli et al., 1994; Yoon et al., 2004; Breen et al., 2012). In this paper we assess the reliability of RPT annotations, as produced by individuals who have had no practical training or familiarization with phonological models of prosody. We ask if crowd-sourced RPT annotations are as reliable as lab-based RPT annotations, considering that the listening environment, electronics, and workspace are not under control of the researcher. A further question is how the annotator’s language background affects their annotation—and specifically, if annotators who speak a prosodically distinct dialect of the target language are less consistent in how they assign labels, which would reduce annotation reliability.

Guidelines for prosodic annotation specify acoustic and perceptual criteria for assigning prosodic labels (ToBI: Beckman and Ayers 1997; Veilleux et al., 2006; RaP: Dilley and Brown 2005), and discourage consideration of contextual factors not directly reflected in the acoustic signal, such as syntactic context. When there is high inter-annotator agreement over a speech dataset, we infer that the annotators are implementing the guidelines in a similar fashion, making associations between the prosodic cues present in the speech signal and the prosodic categories defined by the annotation schema. With the RPT method, annotators do not receive guidelines or explicit criteria for assigning prominence and boundary labels. We are interested if RPT annotators, working without explicit guidelines, assess the prosodic features of an utterance differently than do annotators working with explicit labelling criteria. For this we examine pairwise agreement between each untrained RPT annotators with a consensus ToBI annotation from two trained annotators.

Related to research question (2), we are interested in the factors that influence an annotator’s assignment of prominence and boundary labels. To explore such differences between RPT annotations from different cohorts, and between RPT and ToBI annotations in more detail, we compare correlations between the prosodic labels of an individual annotator and various acoustic and contextual properties of the labelled word. We test three hypotheses. The first concerns differences between lab-based and crowd-sourced RPT annotations. Lab-based annotation work is done in a quiet laboratory environment using quality headphones, while crowd-sourced annotators work in their preferred listening device and in a location of their own choosing. Given this difference, we reason that lab-based annotators may be better able to attend to the sometimes subtle acoustic cues to prosody, and therefore we hypothesize that lab-based annotations will be more strongly correlated with acoustic cues than crowd-sourced annotations, and conversely, that crowd-sourced annotations will be more strongly correlated with contextual cues. Our second hypothesis concerns differences between annotators with different language backgrounds in the associations they draw between acoustic cues and prosodic labels. We hypothesize that there will be differences between annotators whose dialect matches that of the speech they are annotating and annotators whose dialect differs, reflecting differences in their exposure and familiarity with the prosodic patterns of the target dialect. Our third hypothesis concerns the difference between ToBI and RPT annotations. Given that the ToBI guidelines make explicit reference to acoustic cues, we hypothesize that a ToBI annotation will be more strongly correlated with acoustic properties of the annotated words than are RPT annotations. The converse of this hypothesis is that the lack of explicit guidelines may lead RPT annotators to give stronger weight to contextual factors (part of speech, word frequency, boundary status) in assigning prosodic labels.

Finally, as a practical matter for researchers interested in obtaining prosodic data using RPT, we ask how many annotators are needed to optimize annotation reliability, as measured by inter-annotator agreement. Issues related to the reliability and utility of crowd-sourced annotations have been explored for phone-and word-level speech processing tasks with promising results (Eskanazi et al., 2013), and the argument is made that the use of factored problems, of which RPT is shown to be an example, is key to scaling up the size and complexity of the transcribed data, mitigating the problems of diminished accuracy and precision that come with crowd-sourcing transcription (Hasegawa-Johnson et al., 2015).

The paper continues as follows. Section 2 provides some background on RPT, motivating its use in addressing the challenge that prosody annotators face from uncertain or ambiguous cues to prosodic features. RPT is discussed in relation to other approaches that resolve annotation uncertainty by comparing annotations from multiple annotators. Section 3 introduces the methods used in the studies reported here. First, the RPT protocol is described. This is followed by a description of LMEDS (*Language Markup and Experimental Design*), an open-access web application we use to automate data collection with RPT and to crowd-source RPT tasks over the internet, and the LMEDS implementation. We then introduce the statistical methods used to investigate annotation reliability (Fleiss's kappa statistic) and the factors that cue prosodic features (generalized additive mixed models). At the end of Section 3 we introduce three large-scale RPT studies of conversational American English using crowd-sourced and lab-based annotation. Results from these studies, presented in Section 4, inform our research questions and are presented in two parts. First is a comparison of inter-annotator agreement for RPT data collected from the lab-based and crowd-sourced annotators, addressing question of annotation reliability. This is followed by the results of the GAMMs, addressing our second set of research questions concerning the factors that influence prosodic feature assignment under the annotation procedures examined here. Section 5 discusses how the findings inform our research questions, and identify ways in which the RPT method can be optimized. Concluding remarks are presented in Section 6.

2. *Background: addressing the prosody annotator's challenge with multi-annotator solutions*

RPT was developed in response to the challenge that annotators face in trying to determine the appropriate prosodic label to assign to a given word. With a traditional approach to prosodic annotation performed by trained experts, e.g., annotation using the ToBI system (Beckman et al., 2005), annotators use a feature inventory consisting of a language-specific set of pitch accents and boundary tones developed through a linguistic analysis of the prosodic system. The inventory of tonal pitch accent and boundary features for American English is shown in (3) for the purpose of illustration. Pitch accents are tonal features (Pierrehumbert, 1980) assigned to words that meet a criterion of phonological prominence—a notion that is grounded in the metrical (strong/weak) patterning of words at the phrase level (Ladd, 2008: ch. 8; Büring, 2016). Boundary tones are assigned to words that are final (or less often, initial) in a prosodic phrase, with some frameworks differentiating the tone features assigned to lower- versus higher-level prosodic phrases (*intermediate* versus *intonational* phrases in ToBI systems). The task for the annotator is quite complex. The decision as to which pitch accent to assign to a word is intrinsically linked to the decision about the prominence level of the word, and similarly, the decision of which boundary tone to assign is linked to the decision about the degree of phrasal juncture following (or preceding) the word.

(3) Intonation features for American English (ToBI system)

Pitc accents:	H*, !H*, L*, L+H*, L+!H*, L*+H, L*+!H, H+!H*
Phrase accents:	H-, L-
Boundary tones:	H%, L%

From a production perspective, these decisions have an asymmetric dependency. For example, in English, a pitch accent can be assigned only to words that are structurally prominent, though not every prominent word is necessarily assigned a pitch accent. Thus, pitch accent assignment is logically dependent on structural prominence. Similarly, an edge-marking tone (a boundary tone or phrase-accent in the ToBI system) can be assigned only to a word that is final in a prosodic phrase, which makes the assignment of edge tones dependent on the parsing of words in prosodic phrase structure.

Viewed from the perspective of the listener, the directionality of the link between prosodic structure and tone features is less obvious. For instance, a listener may identify prominence on the basis of pitch evidence, but it is also possible that a listener identifies pitch accent on the basis of perceived prominence. The difficulty arises from the fact that the acoustic cues to prominence (spectral and durational measures) and f0 cues to pitch accent can be ambiguous, especially in spontaneous speech (Cole and Shattuck-Hufnagel, 2016). A listener who is uncertain about the status of a word as prominent may draw on pitch evidence to determine the presence of a pitch accent but it's equally possible that a decision to label an ambiguous pitch pattern as a pitch accent may be informed by the perceived prominence of the word. Similar ambiguities arise with cues for boundary tones and prosodic phrase juncture. Ambiguities of this sort make the annotator's task difficult, slow down the annotation process, and are a likely source of inter-annotator disagreement, as annotators may weigh the available cues differently in deciding on the prosodic feature assignment for a given word.

One of the ways that annotators cope with uncertainty in the assignment of prosodic features is to consider the features assigned by another annotator. This can be done through consensus labelling, where two or more annotators work together to agree on an annotation, usually after discussing the cues and labelling criteria for each word where there is disagreement. Another approach is to rely on majority rule over an odd number of independent annotators (typically, three). Yet another method is to resolve the uncertainty consistently in favour of one of the independent annotators, while reporting on inter-annotator agreement as a measure of the degree to which the deciding annotation is representative of other annotators. What all of these methods have in common is the reliance on multiple annotators while producing an output annotation that assigns each word receiving a unique accent and boundary label (including "unmarked" as one of the labels), concealing any underlying disagreement among annotators in the choice of label.

RPT builds on the multi-annotator approach, but with several critical differences from other approaches. First, annotators are not trained to assign features based on a specific annotation framework, or with reference to specific cues as criteria for feature assignment. Rather, RPT annotators are given minimal instruction to assign very coarse-grained features marking words as prominent/not-prominent, or as preceding/not-preceding a boundary, on the basis of their immediate auditory impression alone. Second, RPT annotation is based only on auditory impression, unlike annotation in the ToBI framework, where annotators are trained to interpret visual cues from the pitch track, waveform and spectrogram to augment their auditory impression of prosody. Third, RPT annotators work independently, so there is no opportunity to resolve difficult feature assignments through interaction with other annotators. The uncertainty that arises due to cue ambiguity is not resolved, rather it is captured in a measure of the agreement among annotators in the prosodic marking of each individual word in the speech sample. This brings us to a fourth critical difference between RPT and a traditional prosodic annotation—RPT output is analysed over the entire group of annotators, and measures the perceptual salience of the prominence or boundary feature for a given word on a quasi-continuous scale, from zero to one, where this value represents the proportion of annotators who judged the word as prominent (the "p-score"), or as preceding a boundary (the "b-score"). In comparison, traditional annotation methods do not convey information about the perceptual salience of prosodic features for words in the annotated speech sample, although such distinctions definitely exist, as reflected measures of inter-annotator agreement.

3. Methods

3.1. The RPT protocol

RPT was developed by the first author and members of her lab as a means to obtain prosodic annotation for large, multi-talker databases of conversational speech.[2] RPT involves annotation performed in real time while the annotator listens to recorded speech. Audio files are presented with an accompanying transcript, with punctuation and capitalization removed. Transcripts using the normal orthographic conventions of the language are recommended to facilitate the annotator's task. Laughter, disfluencies, or other non-lexical vocalizations can be transcribed or not, depending on the goals of the research, and considering the impact the detailed transcription would have on the annotation task. Annotation consists of marking a word as *prominent,* or as preceding a prosodic *boundary* at the end of a prosodic phrase. Every word that appears in the transcript can potentially be marked. Prominence and boundary can be defined for the annotator in different ways, depending on the purpose of the annotation. Our prior RPT studies have used different definitions that vary in specificity, with the least specific definitions shown in (4). In other studies, not reported here, we have used more explicit definitions that identify specific cues, in order to focus the annotator's attention on specific auditory criteria related to perceived pitch, loudness and tempo, or (in different experiments) on the functions of prosody in segmenting speech into coherent chunks and signalling discourse meaning.

[2] Findings from our linguistic, phonetic, and computational modelling work on RPT are reported in a series of articles and conference papers: Mo et al., 2008, 2009; Cole et al., 2010a,b; Mahrt et al., 2011, 2012a, 2012b; Cole et al., 2014a,b; Jyothi et al. 2014a,b; Hualde et al., 2016.

(4) RPT instructions with non-specific definitions of prominence and boundary

Boundaries: "Speakers break up utterances into chunks that group words in a way that helps the listener interpret the utterance. You will mark locations where you hear a boundary between two chunks of speech. Note that chunks can vary in size, and boundaries do not necessarily correspond to locations where you would place a comma, period, or other punctuation mark, so you must really listen and mark the boundary where you hear a juncture between chunks."

Prominence: "In normal speech, speakers pronounce some word or words in a sentence with more prominence than others. The prominent words are in a sense highlighted for the listener, and stand out from other non-prominent words. Your task is to mark words that you hear as prominent in this way."

Prominence and boundary marking are performed in separate annotations tasks done in succession over the same speech sample, and can be done in either order—prominences first, or boundaries first. In our work with American English we have found no discernible effect of task order on the outcome when considered over an entire group of annotators (as described below).

The output of RPT as produced by a single annotator is a binary coding of each word in the speech sample for prominence (1 for prominence, 0 for not-prominent), and a similar binary coding for a boundary following the word (1 for boundary, 0 for no-boundary). Pooling annotations from a group of annotators who are assigned the same speech materials, we calculate for each word a quasi-continuous measure of the proportion of annotators who marked the word as prominent (the *p-score*) or as preceding a boundary (the *b-score*). With a large enough group of annotators, these scores can be interpreted as representing the likelihood that the word will be heard as prominent or as followed by a prosodic boundary for a new annotator from the same population. The b-scores and p-scores may also be understood as representing the perceptual salience of the boundary or prominence status of a word. A word that many annotators mark for prominence or boundary is one that presents strong cues that signal the prosodic feature, allowing for the possibility that the cue set may include not only information from the acoustic signal, but also non-acoustic, contextual factors that elicit an expectation of the presence of a prosodic feature (Cole et al., 2010a, b). Conversely, words that few or no annotators mark can be understood as signalling the absence of prominence or boundary cues. Words with ambiguous or conflicting cues are expected to have p-scores or b-scores in the intermediate range. Statistical modelling, as discussed below, can be used to explore the relationship between prosodic marking and the acoustic cue values (e.g., related to f0, intensity, duration, spectral balance) and contextual factors (e.g., syntactic phrase boundary, part of speech, n-gram word frequency, information status) associated with a word.

RPT annotation, as performed in our work and in prior work of the first author, is done in real time. The annotator marks the location of boundaries and prominences on a printed transcript or on a transcript displayed on a computer monitor, while listening to the audio recording. Annotators do not control the audio playback, so annotation of the entire audio file proceeds in pace with the audio presentation of the file. This method is intended to capture the annotator's immediate, subjective judgment of prominence and prosodic phrasing. Through pilot studies we have settled on a procedure where annotators get two passes through the audio file for each task (prominence and boundary marking), with the option in the second pass to change the feature assignment of any word (selecting or de-selecting the word for feature assignment).[3] This procedure allows an individual annotator to annotate an entire speech sample in time equal to 4 times the duration of the audio file, making RPT one or two orders of magnitude faster than a ToBI annotation, in our experience. The actual number of annotation hours required for a complete RPT task depends on the number of annotators assigned to the task, but the number can be roughly computed as $4dn$, where d is the total duration of the audio files to be annotated, and n is the number of annotators assigned to the task.

Because the individual RPT annotator receives no training or feedback during annotation, and because of the time constraints on annotation, RPT annotators are likely to be less consistent and less accurate in labelling all of the prosodic events for which there are detectable acoustic cues. This loss of precision is compensated for, in part, by having annotations of the same materials from multiple annotators, as discussed above. Thus, an individual RPT annotator may fail to mark prominence on a word that a trained expert annotator marks as prominent (e.g., a prominence expressed with salient acoustic cues, on a word that is in an appropriate discourse context), but it is unlikely that all

[3] RPT can, of course, be customized for different research needs. For instance, the inventory of features can be expanded, e.g., to allow two degrees of prominence, or two levels of prosodic boundary. Annotation of prominences or boundaries can be done alone, ignoring the other dimension. The real-time constraint on RPT can be removed, the number of listening passes can be reduced or increased, and annotators can be allowed to control audio playback, moving back and forth in the audio file. Each of these decisions is likely to have an effect on the resulting annotation, so differences in annotation protocol should be considered when comparing annotations across studies—as is always the case.

annotators in the group will fail to mark prominence on such a word. We emphasize that with RPT it is the annotations over the *entire group of annotators*, as captured by p-scores and b-scores or through a statistical model as described in Section 3.3, that convey important information about the prosodic patterns in the speech materials, as perceived by listeners. The number of annotators that should be included in an RPT study is the number that optimizes reliability, as measured through inter-annotator agreement (Section 3.3). Results presented in Section 4.1 suggest that 10−12 annotators are sufficient to guarantee replicability of the annotations at the group level, at least for the kind of American English speech used in the studies reported here.

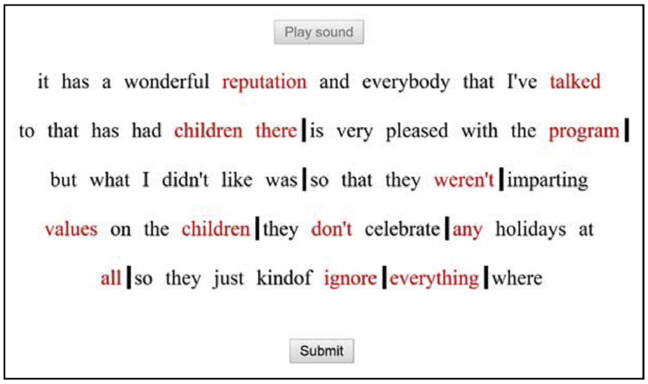
3.2. LMEDS and its implementation for crowd-sourcing

LMEDS, the Language Markup and Experimental Design Software, is an open source web-based platform for running experiments on speech processing over the internet (Mahrt, 2016). It was developed specifically as a computer interface for Rapid Prosody Transcription (Cole et al., 2014a,b), however, it can be used for any experiment that tests participants' identification or discrimination response to an audio stimulus. As used for RPT, LMEDS provides a digital interface through which an annotator marks words on a printed transcript for a corresponding speech audio file, selecting words heard as prominent or as preceding a prosodic boundary. These two tasks are performed in separate listening passes through the audio file, in either order, as chosen by the researcher. During the boundary-marking task, selecting a word causes a vertical bar to appear after the word—providing visual feedback of the perceived boundary. During the prominence-marking task, selecting a word causes the word to change from black (indicating not prominent) to red (indicating prominence). Prosodic marking operates as a toggle in LMEDS. An annotator can change a word's status between prominent and not prominent, or she can add and remove a boundary following a word, with successive clicks on the word. The visible boundary and prominence markings change with each click. A screen shot of LMEDS is shown in Fig. 1.

After an annotator completes annotation of an audio file, the presence or absence of each prosodic marking on a word gets digitized as 1 or 0, respectively, resulting in the assignment of the binary-coded prominence and boundary features for each word. LMEDS also automatically calculates the average over all annotators of the binary prominence values, and similarly for the boundary values, to produce the p-scores and b-scores for all the words in the annotated speech sample.

Beyond its use for RPT, LMEDS has been used for data collection in other prosody research with experiments testing memory for prosody (Kimball et al., 2015) and identification and discrimination of prosodically encoded focus (Mahrt, *in progress*). LMEDS facilitates the administration of annotation tasks and other experiments with tools for collecting informed consent and demographic information. It allows the researcher to modulate how users interact with the audio stimulus and the response screen, for example, by specifying the minimum and maximum number of times an audio file can be played or whether the user is automatically transitioned to a new task once the current task is finished. It also keeps track of user interactions with the page including the number of times that audio files are listened to and how long users spend on each task, though its functions do not currently include reaction time measurement.

As a computer-based platform for collecting listener response data, LMEDS offers the advantage that the participant's responses are automatically digitized, thereby avoiding errors that can otherwise arise in manually driven



it has a wonderful reputation and everybody that I've talked
to that has had children there | is very pleased with the program |
but what I didn't like was | so that they weren't | imparting
values on the children | they don't celebrate | any holidays at
all | so they just kindof ignore | everything | where

Fig. 1. Screen image of the LMEDS interface for RPT annotation.

Please cite this article as: J. Cole et al., Crowd-sourcing prosodic annotation, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.02.008

data transfer. LMEDS offers further convenience for the researcher by providing an experiment report that aggregates data across an entire annotation study or experiment in a format that is easily viewed as a spreadsheet, where each item response (coded as 0 or 1 for the prosodic feature) is matched with the coded attributes of the item in the stimulus (e.g., part of speech, unigram frequency, information status label, maximum f0). This data structure can be directly input for analysis to a statistical package such as *R*.

3.2.1. Technical specification

LMEDS has two halves: server-side code that runs on a remote server and prepares webpages for viewing and processing users' responses; and client-side code that runs in the user's web browser and controls how users can interact with the webpage (such as by limiting how many times the user can listen to an audio file or transitioning the page to a new task). The server-side code is written in Python and the client-side code is written in JavaScript. There are no external dependencies, so LMEDS can be uploaded on any server that has Python enabled, with minimal configuration. For experiments that are carried out in a lab setting on a dedicated computer, LMEDS can be run directly on computer with Python installed, in which case internet access is not required.

Although it was written in Python and JavaScript, LMEDS was designed for use by experimenters who might not have programming experience. Programming experience is necessary to extend LMEDS with new functionality but not to create and run experiments that do not require functions beyond those already provided.[4] LMEDS experiments require two specification files—one specifying the sequence of instructions and tasks and one specifying all of the text a user might experience in the experiment. By splitting these into two parts, the same experiment can be easily run in multiple languages. The sequence file and dictionary file are plain text and can be edited in any text editor, such as Notepad on Windows or TextEdit on Mac.

3.3. Statistical modelling

We use two statistical models to test the reliability of RPT annotations and to test the relationship between the annotated prominence or boundary features of a word and its associated acoustic and non-acoustic (contextual) cues.

3.3.1. Fleiss' kappa

The first statistical tool is Fleiss's kappa (Fleiss, 1971) as produced by the R system with the irr package (Gamer et al., 2012). This is a standard measure of agreement between multiple raters performing a categorical labelling (classification) task, where agreement is measured against the chance likelihood of agreement given the frequency of occurrence of each label in the full dataset. Fleiss' kappa takes values between 0 and 1, with 1 indicating perfect agreement and 0 indicating agreement at chance level. Higher values of Fleiss' kappa indicates that annotators are behaving more alike one another in assigning prominence and boundary features based on the available cues in the speech sample, validating the annotation scheme as reliable over the speech sample, and predicting similar reliability for other similar speech samples.

The observed rate of inter-annotator agreement for a given corpus depends on the difficulty of the annotation over the corpus (i.e., the degree of annotator uncertainty in the assignment of prosodic features), and the degree to which annotators respond similarly to the available cues. If the cues are salient and unambiguous, and if annotators are appropriately sensitized to the relevant cues, they are more likely to agree in their annotations. With ambiguous or conflicting cues agreement rates will be more variable, as annotators may vary in how they resolve the available cues.[5] In addition, agreement rates are expected to be more variable across small annotator cohorts than across large annotator cohorts, for the simple reason that as annotators are added to the cohort, there is a greater likelihood of having annotators with similar response patterns within the cohort. An important question for the researcher using RPT is whether there is an optimal cohort size at which agreement rates converge such that increasing cohort size

[4] A user manual is included with LMEDS which includes detailed instructions for how to create a new experiment, including the specification of tasks in an experiment, the creation of a language dictionary, running experiments, and data aggregation. LMEDS is also distributed with a complete, non-trivial example experiment and a local server that can be used for testing purposes and for running experiments locally.

[5] It's also possible that annotators may vary in how they respond to salient and unambiguous cues, but that is expected to be rare given that the prosodic features in question encode grammatical information about, e.g., focus or syntactic constituency, that are important for the successful communication of the linguistic message.

through the addition of annotators does not add further benefit in lowering the variability of agreement rates. The optimal cohort size would then produce the annotation with the greatest reliability in the sense that the annotations will be minimally sensitive to the specific annotators performing the task.

We use a bootstrapping method to determine the optimal cohort size for each of the three RPT experiments reported below. From each group of N annotators we examine agreement within smaller cohorts ranging in size from 2 to $N-1$ annotators. We use a bootstrapping method to randomly sample (with replacement) 10,000 cohorts at each cohort size, and we calculate Fleiss' kappa as a measure of agreement across the 10,000 cohorts of the same size. The average Fleiss' kappa value over 10,000 cohorts at each cohort size will converge on the Fleiss' kappa value for the full set of annotators, while the standard deviation of Fleiss' kappa is predicted to decrease as cohort size increases, up to the size where variability in Fleiss' kappa stabilizes, which then represents the optimal reliability of the annotation scheme for the speech materials. The average for each bootstrapping will just recover the average kappa in the data. Thus, if the agreement for all data is 0.5 and 10,000 samples with replacement are drawn on all pairwise agreement, the average will converge to 0.5. While the bootstrapping procedure cannot be used to determine how sensitive the number of raters is to the true population inter-rater agreement, it can be used to discuss at what point, for the observed set of data, does adding more raters truly add more information.

3.3.2. Generalized additive mixed models

The second statistical tool is logistic regression modelling with generalized additive mixed models (GAMM) (Wood, 2006, 2011), which we use in this paper to examine whether or not there is consistency between different annotator cohorts in their use of the acoustic and contextual predictors of prosodic prominence and boundaries. By fitting smooth functions to the predictors, GAMMs take into account the possibility of a non-linear relationship between the predictor and the dependent variable—here, binary prosodic features (see Wood, 2006 for a full overview and a description of the mgcv package used to compute these models in R). There are two types of statistical tests used in the GAMM: one is parametric and used for the predictors that are linear or categorical and the interactions that only include linear or categorical predictors. The non-parametric statistical tests are used for the non-linear predictors and any interactions that include the non-linear predictors.

In this study we run GAMMs to test and visualize the main effects of acoustic and contextual factors on prosodic marking, and differences between annotator groups in these effects. Separate GAMMs are used to test prominence marking and boundary marking, using the same set of predictors, and including Subject (annotator) and Word as random factors. The null hypothesis for each predictor is the same: the predictor or interaction term including the predictor has a zero effect on raters' assignment of prosodic prominence and boundaries. The effects of individual predictors on the likelihood of prosodic marking are assessed qualitatively using visualizations produced from the GAMM estimates via the visreg package (Breheny and Burchett, 2013).

3.4. RPT experiments

Three RPT experiments with American English were conducted to address our research questions (1) and (2) above, with two conducted over the internet with crowdsourced annotators, and one conducted in our lab with annotators recruited from the university student community. Data from these experiments are compared to a consensus ToBI annotation from two trained annotators (including the first author) done as part of another study (Cole et al., 2014b, 2016; Hualde et al., 2016).[6]

Materials: The same set of speech materials were administered for the three RPT experiments and the ToBI annotation. Materials consisted of excerpts from the Buckeye corpus of spontaneous narrative speech produced in the context of an ethnographic interview (Pitt et al., 2007). 16 excerpts of 13−24 s (average 18 s) were taken from

[6] The purpose of the ToBI annotation is to enable comparison between RPT and an annotation scheme that is based on a detailed phonological model, explicit acoustic criteria for each label (i.e., intonational feature), and for which annotators undergo extensive practical training. We take the consensus ToBI annotation as representative of such a scheme, while leaving open the possibility that different ToBI annotators might disagree on some of the labels. To the extent that the guidelines are clear, and annotators are implementing them with similar fidelity, these differences should be minimal. It would be of great interest to compare inter-annotator differences for RPT and ToBI (or other annotation schemes), but obtaining the ToBI annotations would require vastly more time, money and human effort than are available to us at this time. We leave this as a topic for future work, while noting that is precisely the difficulty in validating a ToBI annotation that motivates the development of alternative annotation methods, such as RPT.

interviews with 16 speakers of the Buckeye corpus, comprising 931 words total. Each excerpt was extracted in a separate audio file (WAV format) with accompanying word- and phone-aligned transcripts extracted from those published with the corpus.

Participants: RPT experiments were conducted with annotators from three cohorts:

- [Lab-US] native monolingual speakers of American English who were recruited at the University of Illinois
- [MT-US] native speakers of American English who were recruited on Amazon Mechanical Turk, with IP addresses in the US
- [MT-India] people presumed to be speakers of Indian English who were recruited on Amazon Mechanical Turk, with IP addresses in India

32 participants were recruited for each cohort. All RPT cohorts annotated using the LMEDS system. The cohorts recruited online worked unobserved and without supervision in a location of their choosing, while the cohort recruited at the university made their annotations in a computer lab with an experimenter present.

Method: The participants in this study used RPT to annotate prosodic boundaries and prominence. Using the LMEDS web application, annotators listened to an audio file of recorded speech and marked a transcript displayed in a browser window on a computer monitor, clicking on words perceived as preceding a boundary, or in a separate task, clicking on words perceived as prominent.

The RPT experiment began with a request for informed consent and a brief language background survey, both administered through LMEDS. For these RPT experiments, the LMEDS interface running in a browser window presented the annotator with a button that plays an audio file, the interactive transcript of the speech in the audio file, and a button to progress to the next audio file in the experiment. Annotators were given a practice annotation task on a short audio file. The practice segment was excluded from all analyses presented below. No feedback was given on the practice annotation, or on any other annotation performed during the experiment.

For each audio file, annotators first performed the boundary-annotation task, with two listening passes during which the audio file played from start to finish without interruption. Annotators could change boundary marks at will during both listening passes, though to a limited extent given that the annotation is done in real time with the presentation of the audio file. After the second listening pass for boundary annotation, the annotator clicked on a button to proceed in the experiment, moving to prominence annotation for the same audio file. During prominence annotation, boundaries that were previously marked on the same transcript (by the same annotator) remained on the screen, although the annotator was unable to interact with them. Prominence marking was also done in two obligatory listening passes, again with the allowance that prominence marks can be changed during both listening passes. In total, each participant listened to an audio file four times in direct succession, before proceeding to the next audio file. Although boundaries could be freely marked and unmarked across two repetitions of the audio file, and similarly with prominences, only the final marked or unmarked status of each word was recorded. In future studies, it would be interesting to examine words that had their status changed between the first and second audio playback—behaviour that might reveal insights into the perceptual salience of certain prominences, or prominences in certain contexts.

Data coding: The RPT annotations are coded in a binary format for each transcribed word, as described above (Section 3.1). Table 1 shows the set of four acoustic measures and three contextual features for each word that were selected as potential cues for prominence and boundary marking, based on findings from previous studies showing RPT p-scores and b-scores to be correlated with acoustic measures of phone and pause duration, intensity, and f0, and also with contextual factors such as the word's part of speech and unigram frequency (Cole et al., 2010a, 2010b).[7] These cues are entered as predictors in the statistical model reported in Section 4.2. Acoustic measures were extracted using the phone and word level alignments provided by the Buckeye Corpus, to which we added segmentation of the lexically (primary) stressed vowel for each word, based on dictionary specification. From each stressed vowel, we used Praat (Boersma and Weenick, 2016) to measure log maximum f0 and RMS intensity. We assessed durational cues to prosody in two ways. The duration of the silent interval following each word, if any, was

[7] Other non-acoustic, contextual features found to be correlated with perceived prosody in our prior RPT studies include the right- and left- edge syntactic boundaries of a word and lexical givenness, neither of which is investigated here. We have added pause duration as an acoustic cue (especially for boundaries), which was not considered in our earlier work, but which we expect co-varies with syntactic context. In other ongoing work, we are expanding our focus on information status (including both lexical and referential givenness) as a predictor of perceived prominence in RPT tasks where listeners are presented with an entire, intact discourse, rather than the discourse fragments used here.

taken as a measure of pause duration.[8] We also assessed the word phone rate following Pfitzinger (1998), which takes speech rate into account in measuring local changes in tempo such as those due to prosodic lengthening or shortening. Word phone rate was calculated using a 500 ms window with a timestep of 10 ms. Local phonerate was only calculated for windows that did not contain any silences, according to the phone-aligned transcripts.

The log frequency of each word was calculated based on the Switchboard corpus of telephone conversations (Godfrey et al., 1992). We utilized the Switchboard corpus for frequency counts instead of the Buckeye corpus due to its greater length (38 h versus 240 h) and because it is similar in style and genre to the spontaneous monologues in the Buckeye corpus. Each word was also manually labelled for part of speech. The additional factor of a word's status for boundary marking was considered as a potential cue for prominence marking, since in English it is the final content word of a sentence that is assigned the default nuclear prominence, and that word has a high probability of being marked for boundary. Boundary marking was coded as 1 for words that are marked as preceding a Boundary by the same annotator, and as 0 for words that are not marked for Boundary by the same annotator.

Intensity and f0 were centred and scaled within speaker while the other continuous measures were untransformed.

4. Results

4.1. Inter-annotator agreement

We tested inter-annotator agreement for each RPT experiment using Fleiss' kappa to measure the agreement over 32 annotators in marking prominence and boundaries on the full corpus of 931 words. Fleiss' kappa scores, shown in Table 2, are higher for boundary marking than for prominence marking for the two US cohorts (Lab-US and MT-US), consistent with findings from our prior studies. The MT-India annotators do not show the same asymmetry, and have substantially lower agreement rates for both features compared to the US cohorts. Also of interest, we observe no difference in agreement rates for prominence marking for Lab-US versus MT-US annotators, but the Lab-US annotators show higher agreement on boundary marking.

Fleiss' kappa statistic measures agreement within each RPT annotator cohort, but does not tell us to what degree the RPT annotations agree with a traditional annotation using the ToBI system. For that comparison we have calculated Cohen's pairwise kappa for each RPT annotator's prosodic marking paired with the ToBI annotation of the same materials. Because the ToBI annotation for English is much more detailed than an RPT annotation (see inventory of ToBI features in (1)), we have collapsed all ToBI pitch accents into a single category designating "prominence", and all ToBI phrase accents and boundary tones together are collapsed into a single "boundary" category.[9] Fleiss' and Cohen's kappa values use the same scale, with values ranging between 0 (low) and 1 (high).

Table 1.
Acoustic and contextual factors tested as cues for prominence and boundary marking. Abbreviations used in this paper shown in parentheses.

Acoustic factors:	RMS intensity of stressed vowel (intensity) log max f0 of stressed vowel (f0) pause duration following word (pause) word phone rate (phone-rate)
Contextual factors:	log word frequency (frequency) part of speech (POS) prosodic boundary (boundary)—considered only for prominence marking

[8] Short silent intervals may represent stop closure rather than an intended Pause, but no attempt was made to discriminate the status of a measured silent interval as stop- or pause-related. All measurable silent intervals were included in the Pause measure, and the possible inclusion of "pauses" originating in stop closure adds noise to the statistical model presented below, especially in the lower range of Pause duration.

[9] Collapsing ToBI pitch accent categories, and similarly collapsing phrase accents and boundary tones, essentially defines the prosodic category labels in terms their underlying prosodic structure. Grouping pitch accents and boundary marking tones into two broad categories is something that is done in all published studies of prosodic annotation reliability (Pitrelli et al., 1994; Yoon et al., 2004; Breen et al., 2012). Indeed, such as grouping is the basis of the "ToBI-lite" annotation scheme used by Yoon et al. (2004), and is also embodied in the RTP annotation scheme in the Breen et al. (2012) study, where annotating "beat" prominence precedes annotation of intonational melodies associated with those prominences. As noted by a reviewer, it is possible that with a different operationalization of the prominent/non-prominent distinction ToBI, we would arrive at different results in comparing RPT with ToBI. We are pursuing a detailed comparison of RPT prominence labels across each ToBI pitch accent category in separate work (for preliminary results see Cole et al., 2014b).

Please cite this article as: J. Cole et al., Crowd-sourcing prosodic annotation, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.02.008

Table 2.
Fleiss' kappa (κ) over for prominence and boundary marking, calculated over 931 words and 32 annotators, for each RPT experiment.

Experiment	Prominence (κ)	Boundary (κ)
Lab-US	0.31	0.51
MT-US	0.31	0.43
MT-India	0.24	0.23

Table 3 shows the range, mean and s.d. of Cohen's kappa over the 32 annotators in each RPT cohort. The mean of the pairwise Cohen's kappa is substantially lower than Fleiss' kappa for each RPT cohort from Table 2, indicating a higher agreement among RPT annotators than between RPT and ToBI annotators. The most striking observation here is the very large range of Cohen's kappa values in each cohort. For the two US cohorts, the highest of the pairwise kappa scores are in the range of "moderate" to "substantial" agreement (Landis and Koch, 1977), while the lowest kappas are in the range of "unreliable" or in the case of negative kappa values, below chance levels. Of interest, the highest kappa scores for the US cohorts come close to kappa scores for ToBI annotations, as reported by Yoon et al., (2004) and Breen et al., (2012), and discussed further in Section 5. The MT-India cohort shows much lower pairwise kappa scores, indicating that these annotators differ substantially in their prosodic markings from a trained ToBI annotator.

We turn next to examine the relationship between annotation reliability and the size of the annotator cohort. For this we look at the variation in Fleiss' kappa calculated over cohorts ranging in size from 2 to 31 annotators, for each of 10,000 randomly sampled cohorts at each cohort size. The analysis was repeated for each of the three RPT experiments. Fig. 2 plots the standard deviation of the 10,000 Fleiss' kappa values for each cohort size and for each RPT experiment. A clear pattern emerges for prominence and boundary data alike, which confirms the predicted effect of lower variability in agreement scores for larger cohorts. The pattern is remarkably similar for all three RPT annotator groups, and reveals a sharp decrease in the s.d. of kappa scores as cohort size increases from 2 to roughly 5 annotators. This decrease in variability continues to reduce for cohorts of greater than 5 annotators, but much more gradually in an asymptotic pattern, with very little appreciable difference in variability among cohorts with more than 20 annotators.

For the purpose of comparing the pattern across the RPT experiments, we set an arbitrary threshold for the s.d. of $\kappa = 0.05$ and examine the number of annotators needed for agreement rates to stabilize at this level. We feel that this threshold, shown in the horizontal black line in Fig. 2, yields a conservative estimate of the minimum number of annotators needed to obtain reliable annotations using RPT. Comparing first the two cohorts from the US (Lab-US and MT-US), we find a surprising difference: the crowdsourced annotators (MT-US) cross the threshold at a cohort size of 7 for both prominence and boundary annotations, while 11 or 12 lab-based annotators are required to cross the same threshold. Less surprising are the differences between the US cohorts and the Indian cohort. The MT-India annotators show a similar pattern as the Lab-US annotators for boundary marking, but there is more variability among Indian annotators in prominence marking than there is among US annotators, for all cohort sizes.

The analysis of kappa score variability allows us to compare the reliability of annotations produced by annotator cohorts differing in size, in language background, and in their status as lab-based or crowd-sourced annotators. These analyses tell us about the level of within-cohort agreement, but they do not tell us whether the annotations

Table 3.
The range (mean, s.d.) of Cohen's pairwise kappa (κ_C) over the prominence and boundary markings from each RPT annotator paired with the ToBI annotation.

Experiment	Prominence (κ_C)	Boundary (κ_C)
Lab-US	−0.13−0.46 (0.22, 0.14)	0.16−0.66 (0.44, 0.14)
MT-US	−0.06−0.46 (0.19, 0.13)	0.45−0.59 (0.32, 0.15)
MT-India	−0.28−0.35 (0.12, 0.15)	−0.07−0.52 (0.25, 0.18)

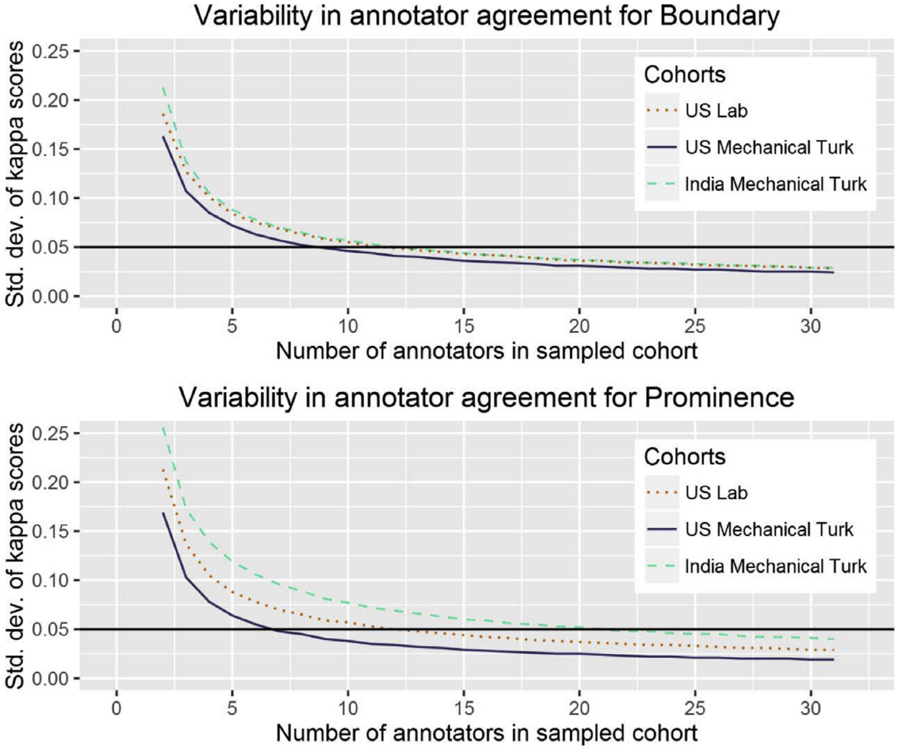


Fig. 2. The standard deviation of 10,000 measures of Fleiss' kappa over annotators' prominence and boundary marking from three RPT experiments (coded by colour). Each kappa score was calculated for a subset of annotators randomly sampled from the full set of 32 annotators for each experiment. The size of the sampled cohort, ranging from 2 to 31, is shown on the x−axis. The black line marks an arbitrarily set threshold of s.d. for kappa scores at 0.05 (see text). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

themselves are similar across cohorts, nor do they indicate whether annotator cohorts are similar in how the available cues influence prosodic marking. Two groups may have comparable agreement rates (indicated by kappa scores), or comparable agreement variability (indicated by kappa score s.d) and yet they may still differ in how the annotators in each cohort carry out the annotation tasks. Annotators may agree on the assignment of a prosodic feature based on different criteria, using different annotation strategies or possibly relying on different cues. The possibility of between-group and within-group variability in how the annotation task is performed is explored in greater detail in the next section.

One thing that we cannot assess with the statistical method employed here is the likelihood of a problematic annotator in a set of data. The assumption made in most experiments, and in annotation studies, is that participants (annotators) are acting in good faith. Quantifying the likelihood of someone is participating in an experiment either in bad faith or in some way independent from instruction would require either a pre-processing diagnostic to identify such individuals, or inclusion of factors that predict such behaviour in the statistical model. For example, in the annotation task described for this study, if a participant was simply marking every fourth item as prominent, this could be identified in the data cleaning phase and the participant eliminated before the statistical analysis. There are, however, an infinite amount of ways to mark words as prominent that do not involve the actual assessment of prominence (e.g. if the word starts with an "s"; if the word rhymes with "bus"; mark every 3rd word as prominent, etc.). To the extent that such strategies can be enumerated and are considered likely for a given set of raters, they be operationalized and built into the statistical model as a predictor. The assumption in the results presented above is that all raters are acting in good faith (even if they are not using the same cues or strategies to assess prosodic marking).

4.2. *Factors that predict prosodic feature assignment (GAMMs)*

This section presents results from generalized additive mixed models that test the effects of the acoustic and contextual factors shown in Table 1 on individual annotators' marking of prominence and boundaries.

4.2.1. *Assessing the full model*

Tables 4 and 5 display results for the GAMM that includes the main predictors from Table 1, plus their interactions with group (US Lab, US Mechanical Turk, Indian Mechanical Turk and ToBI) and a random intercept for both annotator and word. Continuous predictors are fit with a smooth function, indicated by s() while categorical predictors have the same standard regression interpretation. The model allows us to compare RPT annotations from the three RPT experiments, and to compare these with the annotations produced by trained ToBI annotators. More specifically, we are interested in the following comparisons: Lab-based versus Mechanical Turk annotator cohorts; expert versus non-expert cohorts; US versus Indian Mechanical Turk cohorts.

We begin by considering the deviance explained—a percentage that, intuitively, expresses the amount of information in the data captured by the model. For boundary marking, we see that the full model, with all of the main predictors and their interactions with group, captures 51% of the information in the data, while for prominence marking, the model only captures 31% of the information in the data. These results mirror our analysis of the Lab-US cohort data alone when individual differences are accounted for (Roy et al., 2017). The only predictor that fails to obtain statistical significance in boundary marking (at $\alpha = .05$), as a main effect or in interaction with group, is phone-rate as a measure of prosodically induced lengthening or shortening. All predictors obtain statistical significance as main effects for the marking of prosodic prominence.

Table 4.
Non-parametric test for smooth terms (fixed and random effects). Significant effects are marked with shaded cells.

Term	Boundary (Deviance explained = 51%)				Prominence (Deviance explained = 31%)			
	Effective DF[a]	Residual effect DF	Chi. sq	*p*-value	Effective DF	Residual effect DF	Chi. sq	*p*-value
s(intensity)	7.79	8.45	97.51	0.0000	7.89	8.56	72.09	0.0000
s(f0)	7.60	8.37	43.06	0.0000	8.32	8.78	112.96	<.0001
s(pause)	7.77	8.30	69.25	<.0001	6.49	7.24	34.03	0.0000
s(phonerate)	1.00	1.00	1.82	0.1769	7.64	8.41	66.99	<.0001
s(frequency)	1.00	1.00	8.57	0.0034	2.23	2.28	12.86	0.0020
s(word)	212.90	270.00	3947.13	<.0001	245.90	270.00	4383.45	<.0001
s(subject)	88.74	93.00	3289.37	<.0001	91.01	93.00	5316.04	<.0001
s(intensity):groupLab_US	1.00	1.00	7.04	0.0080	1.00	1.00	12.77	0.0004
s(intensity):groupMT_Indian	1.00	1.00	6.88	0.0087	1.00	1.00	30.71	0.0000
s(intensity):groupMT_US	6.09	7.13	28.36	0.0003	2.27	2.86	15.74	0.0011
s(intensity):groupToBI	1.00	1.00	5.68	0.0172	0.00	0.00	0.00	1.0000
s(f0):groupLab_US	2.55	3.19	5.87	0.1497	1.00	1.00	3.66	0.0559
s(f0):groupMT_Indian	1.34	1.60	5.94	0.0588	4.67	5.72	15.53	0.0117
s(f0):groupMT_US	0.00	0.00	0.00	0.9926	1.00	1.00	3.37	0.0665
s(f0):groupToBI	1.00	1.00	0.11	0.7447	0.62	1.04	0.03	0.7841
s(pause):groupLab_US	3.53	4.42	90.07	<.0001	1.00	1.00	1.53	0.2164
s(pause):groupMT_Indian	1.00	1.00	9.31	0.0023	1.00	1.00	1.16	0.2826
s(pause):groupMT_US	6.77	7.78	44.72	0.0000	1.00	1.00	2.01	0.1560
s(pause):group ToBI	4.14	5.01	16.82	0.0047	0.00	0.00	0.00	0.9985
s(phonerate):groupLab_US	1.00	1.00	0.69	0.4087	1.00	1.00	3.98	0.0462
s(phonerate):groupMT_Indian	2.91	3.80	8.38	0.0511	4.98	6.09	29.04	0.0001
s(phonerate):groupMT_US	1.00	1.00	0.15	0.7023	2.55	3.28	11.07	0.0158
s(phonerate):groupToBI	1.19	1.36	1.81	0.3162	0.00	0.00	0.00	0.9977
s(frequency):groupLab_US	3.33	4.09	10.80	0.0318	1.00	1.00	0.81	0.3672
s(frequency):groupMT_Indian	1.00	1.00	0.34	0.5605	7.31	8.28	69.44	<.0001
s(frequency):groupMT_US	1.49	2.08	6.97	0.0448	1.00	1.00	1.00	0.3173
s(frequency):groupToBI	1.86	2.33	2.53	0.3719	0.00	0.00	0.00	0.9948

[a] Effective Degrees of Freedom are the model df that are used for the smoothing functions. They represent a function of non-linearity as 1.0 indicates a linear relationship between the DV and s(IV) and further than 1.0 indicates non-linearity. This measure is described in Wood (2006: 170−171). The residual effective DFs are those degrees of freedom not used by the smooth function.

Please cite this article as: J. Cole et al., Crowd-sourcing prosodic annotation, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.02.008

Table 5.
Parametric tests for categorical predictors. Significant effects are marked with lightly shaded cells. Dark shading of cells for the boundary predictor (bottom rows) indicate the factor is excluded from the analysis of boundary marking.

	Boundary				Prominence			
Term	Estimate	Std. error	z	P-value	Estimate	Std. error	z	P-value
(Intercept)	−4.09	0.29	−14.00	<.0001	−0.70	0.21	−3.37	0.0008
Adjective	0.90	0.27	3.27	0.0011	−1.02	0.16	−6.49	<.0001
posAdverb	1.19	0.27	4.43	0.0000	−0.98	0.14	−7.08	<.0001
posVerb	0.03	0.30	0.10	0.9243	−0.95	0.16	−5.84	<.0001
Group MT_Indian vs. Lab US	−0.37	0.31	−1.22	0.2243	−0.32	0.25	−1.28	0.2008
Group MT_US vs. Lab US	−1.38	0.32	−4.28	<.0001	−0.23	0.24	−0.93	0.3543
Group ToBI vs. Lab US	1.31	1.15	1.13	0.2574	1.06	1.00	1.06	0.2891
posAdjective:group MT_Indian	0.33	0.23	1.42	0.1551	0.11	0.13	0.89	0.3728
posAdverb:group MT_Indian	0.61	0.19	3.22	0.0013	0.39	0.10	3.90	0.0001
posVerb:group MT_Indian	0.33	0.21	1.55	0.1207	0.03	0.10	0.31	0.7575
posAdjective:group MT_US	0.52	0.25	2.05	0.0402	0.12	0.12	0.95	0.3424
posAdverb:group MT_US	0.68	0.21	3.24	0.0012	−0.02	0.10	−0.19	0.8520
posVerb:group MT_US	0.46	0.24	1.90	0.0580	0.02	0.10	0.21	0.8366
posAdjective:groupToBI	−0.88	0.78	−1.12	0.2634	0.61	0.49	1.24	0.2138
posAdverb:groupToBI	0.38	0.63	0.61	0.5448	0.52	0.40	1.30	0.1948
posVerb:groupToBI	0.15	0.68	0.22	0.8292	0.18	0.40	0.45	0.6557
boundarym1					0.54	0.07	7.21	<.0001
Group MT_Indian:boundarym1					0.32	0.10	3.07	0.0022
Group MT_US:boundarym1					0.24	0.11	2.15	0.0313
Group toBI:boundarym1					−0.44	0.35	−1.25	0.2097

4.2.2. Overall effects of annotator group

For the planned comparisons among the three RPT groups and the ToBI annotation, Table 5 shows only one significant difference among the groups in their overall pattern of prosodic marking: The MT-US group produces fewer prosodic boundaries than the Lab-based group, with log-odds of boundary marking at −1.38 (SE = .32, $Z = -4.28$, $p < .0001$).

4.2.3. Assessing group differences in effects of cues on prosodic marking

We move next to our primary interest in whether there are differences among the RPT annotator groups in how cues are used in prominence and boundary marking. For this we turn to examine the effects of the interaction terms for each individual cue paired with group. Parametric tests assess the interaction between group and a categorical predictor, using a reference level for each level of the comparison (coded as treatment contrasts). For example, for the categorical predictor POS, the main effects are tested with respect to Noun as the reference level. For interaction terms with group, the group reference level is the Lab-US cohort and, thus, the reference level of the POS and group interaction is with respect to Noun and Lab-US cohort. For interaction terms of group with the non-linear predictors (the acoustic measures and word frequency), there does not need to be a reference level selected and the statistical significance for each level of the group interaction can be computed separately and tested against the main effect (see Wood 2006 for further discussion of interaction effects in GAMMs).

For the interactions where at least one annotator group obtains a statistically significant difference, we present a visualization of the effect across the range of the predictor. In each of these visualizations of GAMM estimates shown below, the y-axis represents the model-estimated probability of prosodic marking (prominence or boundary) and the x-axis represents the range of values for the given predictor in our data. The black line is the actual model-estimated probability of prosodic marking, while the grey area represents the standard error around the estimate. Along the x-axis, there are short black bars that represent the distribution of our data across each predictor with the thickness representing the amount of data and white space where there is no data. The reader will note that the error for the ToBI annotation is much higher in the presented graphs than the other three groups—this is due to there being only one ToBI annotation versus 32 annotations from the other groups.

The following paragraphs examine the interaction of each cue with group, separately for boundary and prominence marking.

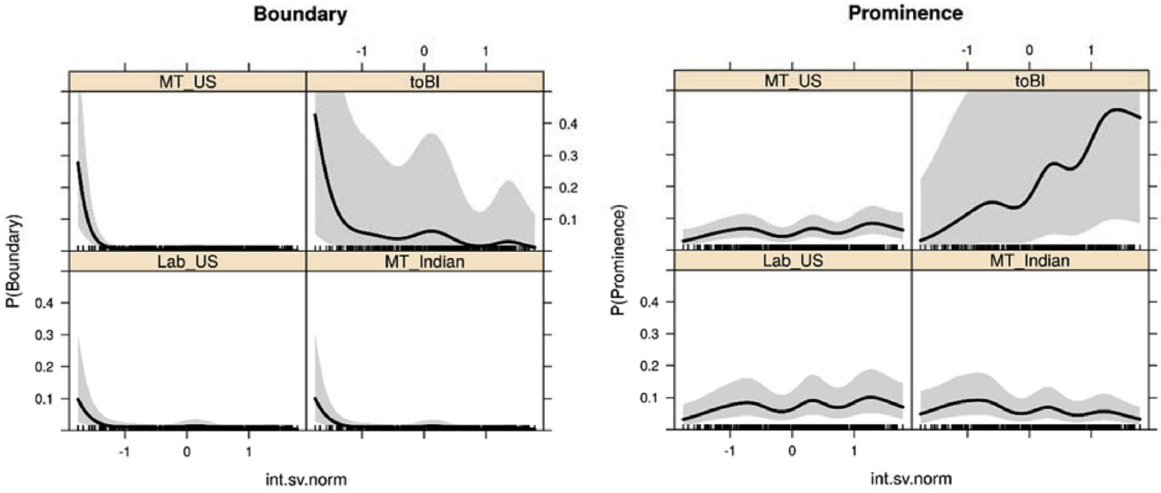


Fig. 3. Estimated probability of prosodic marking (y−axis) across the range of normalized intensity values (x−axis) by group, for boundary marking (left panel) and prominence marking (right panel). The black line shows the model estimated probability and the grey band shows the confidence interval around the estimate. The distribution of intensity values in the data is shown by the thickness of the short black bars along the x−axis, with white intervals at values where there are no data. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

RMS intensity: The GAMM results show a significant interaction between intensity and group for both boundary and prominence marking (see Table 4), which is illustrated by the visualizations in Fig. 3. Effects on boundary marking (left panel) show that the annotator groups use intensity in a consistent way, differing mainly in the magnitude of the effect intensity has on the probability of boundary marking, and the range of intensity values over which the effect is seen. The general pattern for the RPT annotator groups, as for the ToBI annotations, shows a sharp increase in the probability of boundary marking for words whose stressed vowels have the lowest intensity measures, with a rapid drop in the magnitude of the effect for across higher values of intensity. The effect is somewhat stronger for the MT-US cohort than for the Lab-US and MT-India cohorts. The expert ToBI annotation shows an exaggerated pattern, both in the magnitude of the intensity effect on boundary marking, and in terms of the range of intensity values that trigger the boost in boundary marking. The general direction of the effect is, however, the same for the

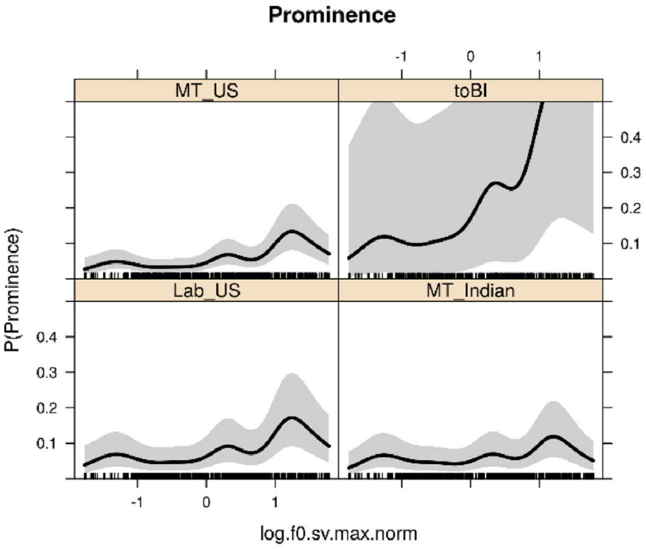


Fig. 4. Estimated probability of prominence marking (y−axis) across the range of normalized log max f0 values (x−axis) by group. Graph details as in Fig. 3.

expert annotation as for the three RPT groups. The effect of intensity on prominence marking is quite flat for the RPT annotators, with only small oscillations across the entire range of intensity. The ToBI annotations, however, show a very different effect, with a marked increase in the probability of prominence marking that grows across the entire range of intensity values.

Log Max f0: The maximum f0 in the stressed syllable is a significant predictor of both prominence and boundary marking in the full model, but it has a significant interaction with group only for model estimates of prominence marking (see Table 4). Fig. 4 shows the interaction effect. The RPT groups show a very similar pattern, with a small increase in the probability of prominence marking for words with very high normalized log max f0 values. The ToBI annotation shows the same overall growth pattern, but as we saw for the interaction of intensity and group, the magnitude of the effect is much greater, and the effect spans a larger range of f0 values.

Pause duration of following word: There is a significant interaction of pause duration (for pause following the target word) with group for boundary marking alone (Table 4). As shown in Fig. 5, the four annotation cohorts show somewhat different patterns of oscillation in the effect of pause on boundary marking. The three RPT cohorts show a common trend where the probability of boundary marking is lowest for words with the shortest duration of following pause, and increases with longer pauses. The RPT cohorts also show a common peak in the effect of pause on boundary marking at similar pause duration values (near 600 ms). The effect of pause on the ToBI boundary annotation shows two peaks of much greater magnitude, one at a smaller value of less than 200 ms, and a second peak at roughly 600 ms. All graphs show a sharp decline in the probability of boundary marking for words with the very longest following pauses, but there are actually very few data points at this end of the pause duration continuum, and the effect here could be driven by the idiosyncrasies (possibly including hesitation disfluency) of very few data points.

Word phone-rate: Word phone-rate, as a measure of local lengthening or shortening, interacts with group only in its effect on prominence marking (Table 4), illustrated in Fig. 6. Once again, we see that the RPT cohorts exhibit a similar pattern, differing in magnitude. The probability of prominence marking is boosted for words with the lowest phone-rates (longer phone-normalized word duration), and gradually decreases for words with higher phone-rates (shorter phone-normalized word duration). This pattern is consistent with a lengthening effect of prominence. For the MT-India cohort there is an apparent boost in prominence for the extreme high values of phone-rate, but again, we see that the data is extremely sparse in this region, which calls for caution in interpreting the observed pattern. As we have seen before, the ToBI annotation shows the same overall pattern as the RPT annotations, with a boost in the probability of prominence marking for words with the lowest phone-rate, but the effect is of much larger magnitude, and is in fact eclipsed by the cut-off on the upper end of the *y*-axis, which we have kept constant in all of the graphs to facilitate comparison of effect magnitude across cues.

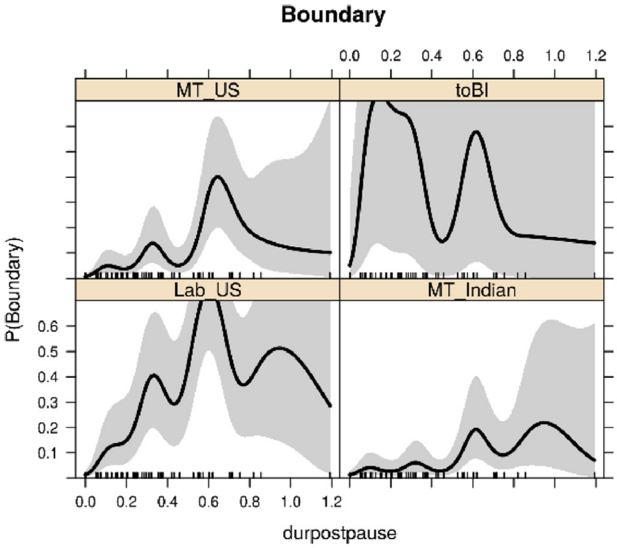


Fig. 5. Estimated probability of boundary marking (*y*−axis) across the range of duration values for the pause following the target word (*x*−axis) by group. Graph details as in Fig. 3.

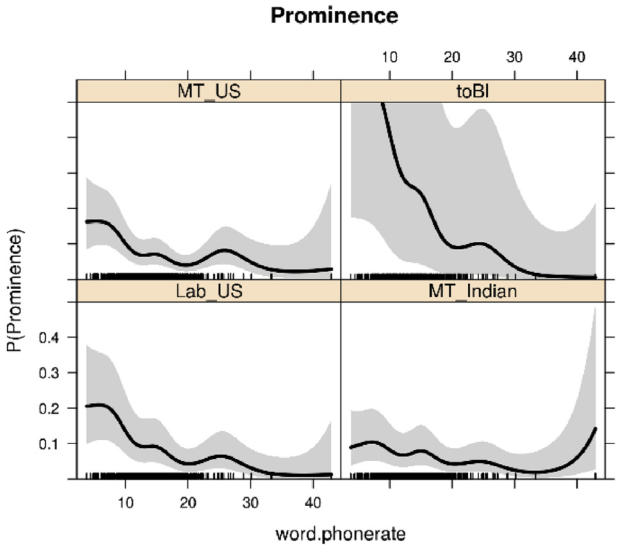


Fig. 6. Estimated probability of prominence marking (y−axis) across the range of duration values for word phone−rate (x−axis) by group. Graph details as in Fig. 3.

Word frequency: The log word frequency measure has a significant interaction with group for both boundary and prominence marking (Table 4), as illustrated in Fig. 7. For boundary marking (left panel), the US annotator cohorts are similar in having a negligible effect with at best a very tiny boost in the probability of boundary marking for words with the lowest frequency values. The effect is a little stronger for the MT-India cohort, and even stronger for the ToBI annotation. For prominence marking the effects are greater if not more differentiated. The two US cohorts show word frequency effects similar to that of the expert ToBI annotation, but with smaller magnitude that decays to near zero at mid-range frequency values. The MT-Indian cohort shows an effect that follows the same general trend, with decreasing probability of prominence marking in relation to increasing word frequency, but with more oscillation, thus marking a less consistent pattern than the US cohorts.

Part of speech: Although POS is a significant predictor of both boundary and prominence as a main effect when controlling for all other factors in the model, the interaction of POS with group is significant for only some comparisons (Table 5). As shown in Fig. 8, the effect of adverb increases the estimated probability of boundary marking by

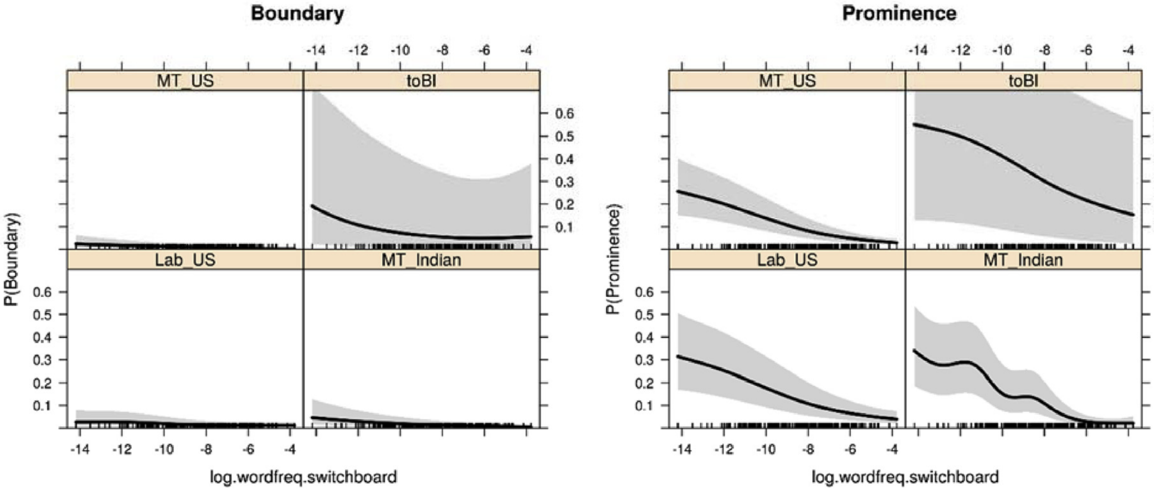


Fig. 7. Estimated probability of prosodic marking (y−axis) across the range of log word frequency values (x−axis) by group, for boundary marking (left panel) and prominence marking (right panel). Graph details as in Fig. 3.

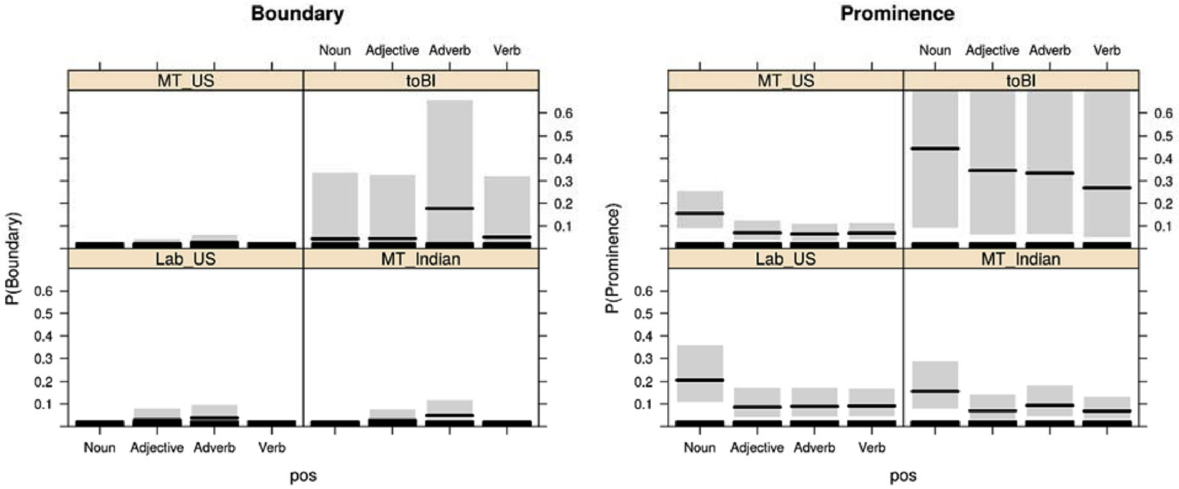


Fig. 8. Estimated probability of prosodic marking (y–axis) by part of speech category (x–axis) by group, for boundary marking (left panel) and prominence marking (right panel). Graph details as in Fig. 3.

at least a small amount in all groups, but there is variation in the magnitude of that effect with the RPT annotators having a much smaller effect than the ToBI annotators. In the marking of prominence, it is the noun category that increases the likelihood of prosodic marking, differing in magnitude across cohorts. We again see that the magnitude of the effects is greatest for the ToBI annotation.

Boundary marking: An overall effect of a word's boundary status on prominence marking was observed in the full model, when controlling for effects from all other factors (Table 5). In addition, there was a significant interaction of boundary marking with group, shown in Fig. 9, with the two MT cohort exhibiting a slightly stronger effect than the Lab-US cohort. It is very interesting to note that while the effect holds across all three RPT cohorts, there is hardly any apparent effect of boundary on prominence marking in the ToBI annotation.

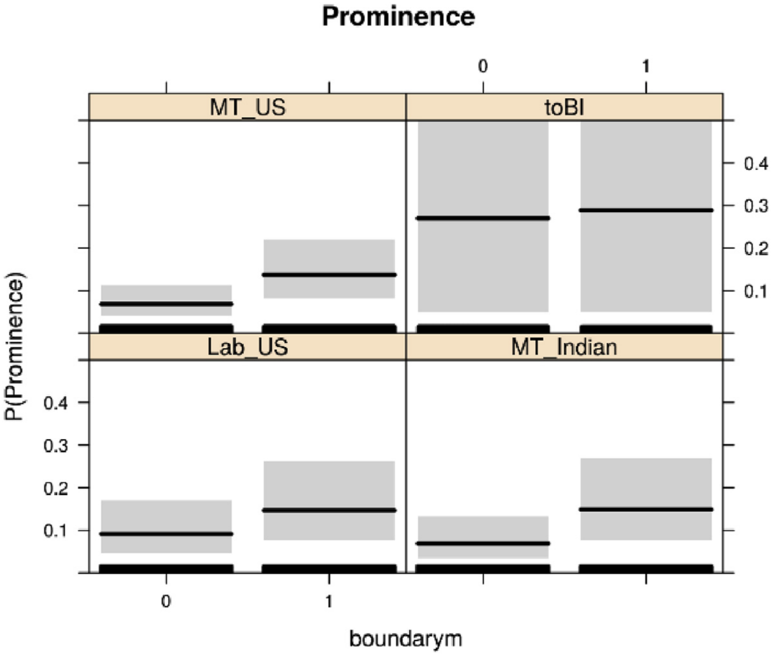


Fig. 9. Estimated probability of prominence marking (y–axis) in relation to the boundary status of the same word (x–axis) by group. Graph details as in Fig. 3.

5. Discussion

To summarize, RPT annotations of American English spontaneous speech were obtained from three annotator cohorts, including two cohorts of crowd-sourced annotators. These RPT annotations were compared with one another, and each was compared to a consensus ToBI annotation of the same materials performed by two trained annotators. Two research questions were posed. The first concerns the reliability of the RPT annotations, and in particular, reliability of RPT annotations produced by crowd-sourced annotators. The second question is whether the RPT annotators from each cohort are using the same acoustic and contextual information as cues in prosodic marking, and whether they use the same cues as are used by ToBI annotators. We review the findings here.

5.1. Annotation reliability

Annotation reliability was assessed through analysis of the kappa score for multi-rater (Fleiss' kappa) and pairwise agreement (Cohen's kappa). The kappa score measures the level of agreement among annotators, correcting for the chance likelihood of agreement.

Crowd-sourced versus lab-based RPT annotations: The lab-based RPT annotations are more reliable than the crowd-sourced RPT annotations for boundary marking, though for prominence marking the US crowd-sourced and lab-based cohorts are matched in reliability (Table 2).

US-based versus Indian RPT annotations: The RPT annotations from US cohorts (Lab-US and MT-US) are more reliable than the Indian cohort. The US cohorts show greater reliability for boundary marking than for prominence marking, while the Indian cohort has very low reliability for both tasks. The Indian RPT annotators do not appear to benefit from the greater salience of the acoustic or contextual cues for boundaries that inform prosodic marking for the US annotators. The overall lower reliability of the Indian annotations suggests an advantage for RPT annotators who have greater familiarity with the language variety represented in the speech materials, or who are themselves speakers of the same variety.

RPT versus ToBI annotations: The overall reliability of RPT annotations, as measured by Fleiss' multi-rater kappa statistic, show that while RPT annotators agree at a rate significantly above chance, the agreement levels are not high, falling in ranges that signal "fair" or "moderate" reliability (Landis and Koch, 1977). Notably, the RPT annotations are less reliable than what has been reported in prior studies for ToBI annotations for spontaneous American English. For example, Yoon et al., (2004) report kappa scores of 0.75 for the presence/absence of prominence marking and 0.67 for the presence/absence of boundary marking, while Breen et al., (2012) report kappa scores of 0.64 for prominence and 0.54 for boundaries. Clearly, the trained ToBI annotators obtain a higher level of chance-corrected agreement than we find with the RPT annotators. Some of the differences in kappa scores may reflect differences in the materials transcribed, or in annotator training procedures, but a large factor must be related to differences in the annotation procedure itself: RPT annotation is done in real time, and is based on auditory impressions alone, while ToBI annotation is done without time constraints, taking anywhere from 100 to 200 times real time (Syrdal et al., 2001), and is informed by visual information from the waveform, spectrogram and f0 track. In short, it would seem that compared to RPT annotators, the ToBI annotators can better utilize the information present in acoustic (and possibly, contextual) cues in deciding on the assignment of prosodic features, resulting in higher inter-annotator agreement, though also dramatically reducing annotation efficiency.

While the overall agreement among RPT annotators is lower than for ToBI annotators (as reported in prior work), the pairwise comparison of individual RPT annotators with the ToBI annotation shows a tremendous range over the 32 annotators in each cohort. Especially for the US cohorts, there are some individual RPT annotators whose annotations fairly closely match the ToBI annotations for the presence/absence of prominence and boundary labels. In both of the US cohorts, the top two or three RPT annotators show "moderate" to "substantial" levels of agreement with ToBI, with the top pairwise kappa scores at 0.46 for prominence, and 0.66 for boundary. The top pairwise kappa scores for boundary marking come close to what is reported by Yoon et al. and Breen et al. for agreement among trained ToBI annotators for spontaneous speech, as mentioned above. The finding that some RPT annotators produce annotations that closely resemble those from trained annotators suggests the use of a selection criterion for RPT annotators. Selecting only those RPT annotators whose prosodic marking matches ToBI on a small sample would be one way to reduce the overall number of RPT annotators assigned to a task, without comprising the reliability or quality (relative to ToBI) of the annotation.

Effect of cohort size on reliability: We examined the standard deviation of Fleiss' kappa statistics over randomly sampled cohorts of all sizes between 2−31 annotators, selected from our full cohort of 32 annotators form each RPT experiment. Higher standard deviation of Fleiss' kappa across annotator cohorts of the same size indicates that the annotations are more dependent on the behaviours of the specific annotators performing annotation than for cohorts that report a lower standard deviation of Fleiss' kappa. This is important information for the researcher planning on collecting prosodic annotations using RPT, who must decide how many RPT annotators to use for a given task. One would like to know the minimum number of annotators needed to optimize the reliability—and by extension, the replicability—of the annotation.

The analysis of cohort size on reliability reveals the same general pattern for all three RPT cohorts. For boundary marking, stable reliability is obtained for cohorts of between 10−12 annotators. For prominence marking, the optimal cohort size varies among the three RPT groups studied. The stability threshold of 0.05 on the kappa scale is achieved with as few as 5 annotators in the MT-US cohort, while the Lab-US cohort requires 12 annotators, and the MT-India cohort requires 20. Across tasks, cohorts less than 5 annotators are highly unstable, showing substantial variation in the reliability measure. For the US cohorts, this means that any particular RPT annotation from a cohort of 5 or fewer annotators is subject to the idiosyncrasies of those annotators, while annotations from cohorts of 10−12 annotators (or more), are less vulnerable to those individual differences, presumably because the larger cohort is likely to represent the full range of individual annotation behaviour present in the larger population. Very minimal advantages for reliability are gained with cohorts with more than 12 annotators, at least for annotation tasks with similar speech materials (i.e., spontaneous, interactive American English speech).

The MT-US cohort shows less variability in the reliability measure in relation to cohort size compared to Lab-US cohort. This is a somewhat surprising finding considering that the two cohorts are matched on the overall level of reliability for prominence, and further, that the Lab-US cohort shows higher reliability for boundary marking (see Table 2). The more stable reliability of the MT-US annotations indicates that annotators recruited through Mechanical Turk are more consistent in performing RPT. A possible explanation for this finding is that Mechanical Turk annotators were restricted to those individuals with a high record of past performance on Mechanical Turk, with their performance on our task also added to their record, while the lab-based annotators were mostly undergraduate students recruited from posted and email announcements, with no performance screening. It seems likely to us that the MT annotators brought better and more consistent work habits to the task.

5.2. Factors that cue prosodic marking

Apart from our interest in the reliability of prosodic RPT annotations, we want to know if annotators from different cohorts are attending in a similar way to the available cues for prosodic features.

Generalized additive mixed models were used to test the contribution of select acoustic and contextual factors (Table 1) on prominence and boundary marking. We find a main effect on both prominence marking on boundary marking for each predictor in the model, holding annotator group and other predictors constant. This finding tells us that the cues included in the model, selected on the basis of prior studies showing these factors to influence prosody, function in similar manner (i.e., without opposing effects) in influencing prosodic marking for crowd-sourced and lab-based annotators, and for US and Indian annotators.

Differences among the RPT annotator cohorts are confirmed on the basis of GAMM results showing significant interactions of the acoustic and contextual predictors with group, as reported in Tables 4 and 5. Interaction effects with group are manifest primarily in the magnitude of the effect that a predictor has on the probability of prosodic marking, and in a few cases, in whether the predictor has an effect or not. These interactions are assessed qualitatively with the aid of the visualizations of the model estimates for the effect of each predictor on the probability of prominence or boundary marking (Figs. 3−9). The visualizations show similar effect patterns across the RPT cohorts for all predictors, for both prominence and boundary marking. Looking at the visualizations for the RPT cohorts, we see large estimated effects of intensity and pause on the probability of boundary marking, more modest estimated effects of f0, phone-rate, frequency and POS on the probability of prominence marking, with very minimal estimated effects for the remaining predictors. What is interesting is the consistency of these effects across all three RPT cohorts, showing that despite differences among the cohorts in overall levels of inter-annotator agreement, the annotators in each cohort appear to be using the available cues in roughly a similar fashion for prosodic marking.

We turn now to consider how the results from the GAMM models relate to the specific hypotheses formulated in Section 1 about differences between annotator groups in the association between acoustic and contextual cues, and an annotator's assignment of prominence and boundary labels.

Crowd-sourced versus lab-based RPT annotations: We hypothesized that lab-based RPT annotators would exhibit a greater effect of acoustic factors on prosodic marking, and a lesser effect of contextual factors compared to crowd-sourced annotators. We found no such consistent pattern. We do observe group differences in the magnitude of effects on prosodic marking, as already noted, but the differences are not consistent in terms of which RPT cohort exhibits the greatest effect for acoustic versus contextual factors. This means that even though the listening environment for the crowd-sourced annotators is not controlled, they appear to have the same ability to access information about acoustic cues for prosody as the lab-based annotators have. Moreover, the crowd-sourced annotators do not seem to rely more on contextual factors in determining prosodic labels. In short, there is no evidence that RPT annotation carried out in a laboratory-style listening environment is different from RPT annotations in a less controlled environment.

US-based versus Indian RPT annotations: Our second hypothesis was that the Indian RPT annotators would differ from the US cohorts in how the available acoustic cues influence prosodic marking. This hypothesis rests on the observation that Indian and American varieties of English differ in the phonetic encoding of prosody, especially for prominence. We find weak support for this hypothesis in our data. The estimated effects of acoustic predictors for prosodic marking by the MT-India cohort are consistently weaker than for the US cohorts, though in the same direction. The estimated effect of word frequency, a contextual factor, on prominence marking wavers inconsistently for the MT-India cohort, unlike the smoother pattern for the US cohorts, while the effects of POS on prominence and boundary marking are very small for all cohorts, such that differences among the cohorts are difficult to identify from the graph.

RPT versus ToBI annotations: Our third hypothesis claims stronger effects on prosodic marking from acoustic factors for the ToBI annotators compared to the RPT cohorts, due to the ToBI annotators' access to both auditory cues and visual cues from the graphical speech display, their extensive training, and the specification of explicit annotation guidelines. This finding is robustly confirmed in the data. We find that with only two exceptions, the acoustic and contextual predictors have a much bigger estimated effect on prominence and boundary marking for the ToBI annotators than for they do for the RPT annotators. One exception to this pattern is in the effect of pause on boundary marking, where we see similarly large effects for the ToBI annotators as for the Lab-US cohort. We also hypothesized that contextual cues would have a stronger influence on prosodic marking for the RPT annotators, compensating for the smaller influence of acoustic cues. This finding is not confirmed. If anything, the ToBI annotations shown a slightly stronger influence form contextual factors than do RPT annotations. Overall, these findings suggest that the prosodic features assigned by the ToBI annotator are more strongly grounded in the available individual acoustic and contextual factors than are the prosodic features marked by the RPT annotators. At the same time, the way in which each predictor affects prosodic marking is the same for RPT and ToBI annotators alike: the model estimates show very similar patterns of linear and non-linear correspondences between predictors and prosodic marking.

A related observation about the comparison of RPT and ToBI annotations comes from our analysis of individual differences in annotators' use of cues in prosodic marking (Roy et al., 2017). We have noted in this paper that annotators differ widely in the extent to which their prosodic marking agree with the ToBI annotation, based on Cohen's kappa. In other work we have examine GAMM results for individual subjects (as random factors in the model) to see how individual annotators differ from one another, and from the ToBI annotators, in the association between individual predictors and prosodic marking. Summarizing from those results, we find that some RPT annotators show estimated effects of predictor variables that have the same large magnitude as for the ToBI annotators, indicating that not only are those RPT annotators using the same predictors as the ToBI annotators, but they are using them in the same way and with the same level of sensitivity. In the present paper we show GAMM visualizations only for the group interactions, not for individual annotators. The group interaction effects combine data from all annotators in the same cohort. The confidence intervals hint at the extent of individual variation, and at the presence of RPT "super-annotators", whose annotation behaviour is very similar to that of trained ToBI annotators.

A final observation on the GAMM results is that the confidence intervals around model estimates are similar in size across the RPT cohorts, indicating similar levels of within-group variation in how acoustic and contextual cues influence prosodic marking. This is interesting given that we find differences in within-cohort inter-annotator agreement for the same cohorts. Apparently, a cohort such as MT-India can produce annotations with (relatively) low reliability, and having low agreement with ToBI annotations, while also displaying consistency in the way that cues are related to prosodic marking, albeit weakly so. This finding points to the complexity of evaluating prosodic annotation, and the need to distinguish overall reliability of an annotation from the strength of the relationship between prosodic marking on one hand, and the acoustic and contextual factors that cue prosody on the other.

6. Conclusion

Findings from this study provide solid evidence for the viability of crowd-sourced prosodic annotation, using the simplified annotation method of RPT. The two primary findings are that crowd-sourced RPT annotations are at least as reliable as annotations from lab-based, student annotators, and that all annotators, including ToBI annotators, draw similarly on information from acoustic and contextual cues in assigning prosodic labels.

Two other findings are of special significance for the researcher planning to use RPT to obtain prosodic data. First is the result from the reliability analysis showing that the optimal number of annotators is between 10−12 for the type of speech materials tested here, or possibly fewer for the more consistently performing US annotators recruited through Mechanical Turk. The second finding relevant for designing an RPT study is that the most reliable annotations come from annotators who have strong familiarity with the language variety represented in the materials to be annotated. While annotators from a prosodically distinct dialect may rely on similar cues in assigning prosodic features, there is much less consistency among those annotators, resulting in annotations with low reliability.

RPT annotations agree only moderately with a ToBI annotation for the presence versus absence of prominence and boundary features, though within each annotator group there are a few “super-annotators” whose prosodic marking agrees with the ToBI annotations at much higher rates, even coming close to the level of agreement reported among trained ToBI annotators in other work. For researchers who seek annotations comparable to ToBI, these super-annotators could be identified in an initial test comparing a small sample of their annotations against a ToBI sample, if available. Other researchers may be less concerned about a match to ToBI, especially in the case of research on a language for which a ToBI-type annotation standard is unavailable.

Using RPT and LMEDS for prosodic annotation offers several advantages, first among which is efficiency and cost. Prosodic annotations using RPT and LMEDS were obtained for the experiments presented here in as little as a single day. Payments to annotators are minimized by not having a necessary training period, and by adopting time limits that enable complete annotation of materials in roughly four times the total duration of the audio files. A further advantage of using LMEDS is that it allows remote participation, making it possible to get annotations from a wider population that is typical of lab-based annotation.

A specific advantage of RPT, whether deployed in a lab or using remote, crowd-sourced annotators, is that it facilitates prosodic annotation for languages whose prosodic phonology has not yet been analysed, and which therefore lack a fully developed inventory of prosodic features and annotation guidelines. Studies of this sort that use the RPT method are reported in Jyothi et al., (2014), Luchkina et al. (2015) and Luchkina and Cole (2016). Moreover, RPT annotations offer something not offered in a typical prosodic annotation that is produced by one annotator, or in which annotator disagreements are resolved, and that is information about the perceptual salience of prosodic features. The disagreements among RPT annotators in the assignment of prosodic features have the potential to tell us something about the factors that cue listeners to perceive prosodic features in the course of comprehending spoken language. Research using RPT to study the perceptual processing of prosody is reported in Baumann (2014), Bishop (2013), Buxó-Lugo and Watson (2016), Erickson et al., (2015), Kim et al., (2016), and Turnbull et al., (2017). RPT has also been used to investigate how L2 speakers perceive prosody (Pintér et al., 2014; You, 2012). Finally, as researchers, including us, begin to investigate individual differences in our RPT data (Bishop, 2016; Roy et al., 2017), we see very interesting and systematic differences in how individual annotators respond to cues in assigning prosodic features, raising interesting questions for future research on whether and how such differences might impact the comprehension of grammatical and discourse meaning in spoken language.

Please cite this article as: J. Cole et al., Crowd-sourcing prosodic annotation, Computer Speech & Language (2017), http://dx.doi.org/10.1016/j.csl.2017.02.008

Acknowledgments

This work was supported by the grants to the first author from the National Science Foundation [BCS 12−51343 and SMA 14−16791]. Thanks to José Hualde, Chris Eager, and Suyeon Im for helpful discussion and for their contributions to data collection and design of RPT tasks used in this study. Thanks also to participants at the 3rd Conference on Experimental and Theoretical Approaches to Prosody for comments and questions on this work.

References

Artstein, R., Poesio, M., 2008. Inter-coder agreement for computational linguistics. Comput. Linguist. 34 (4), 555–596.

Arvaniti, A., 2016. Analytical decisions in intonation research and the role of representations: lessons from Romani. Lab. Phenol. J. Assoc. Lab. Phonol. 7, 1–43 http://dx.doi.org/10.5334/labphon.14.

Baumann, S., 2014. The importance of tonal cues for untrained listeners in judging prominence. In: Proceedings of the Tenth ISSP, pp. 21–24.

Beckman, M.E., Ayers, G.E., 1997. Guidelines for ToBI Labeling, Version 3. Ohio State University http://www.ling.ohio-state.edu//research/phonetics/E_ToBI/singer_tobi.html .

Beckman, M.E., Hirschberg, J., Shattuck-Hufnagel, S., 2005. The original ToBI system and the evolution of the ToBI framework. Prosodic Typology: The phonology of Intonation and Phrasing 10.1093/acprof:oso/9780199249633.003.0002.

Bishop, J., 2013. Information structural expectations in the perception of prosodic prominence. In: Elordieta, G., Prieto, P. (Eds.), Prosody and Meaning. Walter de Gruyter, Berlin.

Bishop, J. 2016. Individual differences in top-down and bottom-up prominence perception. *Proceedings of Speech Prosody*.

Boersma, P., Weenink, D., 2016. Praat: doing phonetics by computer. Retrieved from http//www.praat.org/.

Bolinger, D., 1954. English prosodic stress and Spanish sentence order. Hispania 37, 152–156. doi: 10.2307/335628.

Bolinger, D.L., 1958. A theory of pitch accent in english. Word 14, 109–149.

Bolinger, D., 1982. Intonation and its parts. Language 58, 505–533. doi: 10.1371/journal.pone.0005772.

Breen, M., Dilley, L.C., Kraemer, J., Gibson, E., 2012. Inter-transcriber reliability for two systems of prosodic annotation: ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch). Corpus Linguist. Linguist. Theory 8, 277–312. doi: 10.1515/cllt-2012-0011.

Breheny, P., Burchett, W., 2013. Visualizing regression models using visreg. http://myweb.uiowa.edu/pbreheny/publications/visreg.pdf.

Buhmann, J., Caspers, J., van Heuven, V.J., Hoekstra, H., Martens, J-P., Swerts, M., 2002. Annotation of prominent words, prosodic boundaries and segmental lengthening by non-expert transcribers in the Spoken Dutch Corpus. In: *Proceedings of LREC* 2002, pp. 779–785. Spain, Las Palmas.

Büring, D., 2016. *Intonation and Meaning*. Oxford University Press.

Buxó-Lugo, A., Watson, D.G., 2016. Evidence for the influence of syntax on prosodic parsing. J. Mem. Lang. 90, 1–13.

Calhoun, S., 2010. How does informativeness affect prosodic prominence? Lang. Cognit. Process. 25, 1099–1140. doi: 10.1080/01690965.2010.491682.

Chafe, W., 1987. Cognitive constraints on information flow. In: Tomlin, R.S. (Ed.), Coherence and Grounding in Discourse. John Benjamins Publishing, Amsterdam, pp. 21–51. doi: 10.1075/tsl.11.03cha. NL.

Cole, J., Mo, Y., Hasegawa-Johnson, M., 2010a. Signal-based and expectation-based factors in the perception of prosodic prominence. Lab. Phonol 1, 425–452.

Cole, J., Mo, Y., Baek, S., 2010b. The role of syntactic structure in guiding prosody perception with ordinary listeners and everyday speech. Lang. Cognit. Process 25 (7), 1141–1177.

Cole, J., Mahrt, T., Hualde, J.I., 2014a. Listening for sound, listening for meaning: task effects on prosodic effects on prosodic transcription. Speech Prosody, Dublin, Ireland.

Cole, J., Hualde, J.I., Mahrt, T., Eager, C., Im, S., 2014b. The perception of phrasal prominence in conversational speech. Poster presented at Laboratory Phonology. 14, Tokyo.

Cole, J., Shattuck-Hufnagel, S., 2016. New methods for prosodic transcription : capturing variability as a source of information. Lab. Phonol. J. Assoc. Lab. Phonol. 7 (1), 1–29 doi: http://doi.org/10.5334/labphon.29.

Crystal, D., 1969. Prosodic Systems and Intonation in English. Cambridge University Press, Cambridge, UK.

Dilley, L.C., Brown, M., 2005. The RaP (Rhythm and Pitch) Labeling System, Version 1.0. Available at: http://tedlab.mit.edu/tedlab_website/RaPHome.html (accessed 09.03.11).

Erickson, D., Kim, J., Kawahara, S., Wilson, I., Menezes, C., Suemitsu, A., Moore, J., 2015. Bridging articulation and perception: the C/D model and contrastive emphasis. Proc. Int. Congr. Phonetic Sci http://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0527.pdf.

Eskenazi, M., Levow, G., Meng, H., Parent, G., D., Suendermann, 2013. Crowdsourcing for Speech Processing: Applications to Data Collection, Transcription and Assessment. Wiley, New York.

Fleiss, J.L., 1971. Measuring nominal scale agreement among many raters. Psychol. Bull 76, 378–382. doi: 10.1037/h0031619.

Gamer, M., Lemon, J., Fellows, I., Singh, P., 2012. Various Coefficients of Interrater Reliability and Agreement [WWW Document]. http://cran.r-project.org/web/packages/irr/irr.pdf.

Godfrey, J.J., Holliman, E.C., McDaniel, J., 1992. SWITCHBOARD telephone speech corpus for research and development. In: Proceeding of 1992 IEEE International Conference Acoustic Speech and Signal Processing., 1, pp. 517–520. doi: 10.1109/ICASSP.1992.225858.

Grabe, E., Post, B., 2002. Intonational variation in English. In: Bel, B., Marlien, I. (Eds.), Proceedings of Speech Prosody, pp. 343–346.

Gussenhoven, C., 1984. On the Grammar and Semantics of Sentence Accents. Foris Publications, Dordrecht, NL.

Gussenhoven, C., 2004. The Phonology of Tone and Intonation. Cambridge University Press. doi: 10.1017/CBO9780511616983.

Hasegawa-Johnson, M., Cole, J., Jyothi, P., 2015. Models of dataset size, question design, and cross-language speech perception for speech crowd-sourcing applications. Lab. Phonol. 6, 381–431. doi: 10.1515/lp-2015-0012.

Halliday, M.A.K., 1967. Notes on transitivity and theme in English: Part 2. J. Linguist. 3, 199–244. doi: 10.1017/S0022226700016613.

Hualde, J.I., Cole, J., Smith, C.L., Eager, C.D., Mahrt, T., Souza, R.N., De Université, A., 2016. The perception of phrasal prominence in English, Spanish and French conversational speech. In: Proceeding of Speech Prosody, Boston, MA.

Jun, S.-A., 2006. Prosodic typology: The phonology of Intonation and Phrasing. Oxford University Press, Oxford, UK.

Jun, S.-A., 2014. Prosodic Typology II: The Phonology of Intonation and Phrasing. Oxford University Press, Oxford, UK.

Jun, S.-A., Fletcher, J., 2014. Methodology of studying intonation: from data collection to data analysis. In: Jun, S.-A. (Ed.), Prosodic Typology II: The Phonology of Intonation and Phrasing. Oxford University Press, Oxford, UK, pp. 493–519.

Jyothi, P., Cole, J., Hasegawa-Johnson, M., Puri, V., 2014. An investigation of prosody in Hindi narrative speech. In: Proceedings of Speech Prosody 7, Dublin.

Kim, J., Ramakrishna, A., Lee, S., Narayanan, S., 2016. Relations between prominence and articulatory-prosodic cues in emotional speech. Speech Prosody 2016, 893–896.

Kimball, A.E., Cole, J., Dell, G., Shattuck-Hufnagel, S., 2015. Categorical versus episodic memory for pitch accents in English. In: Proceedings of the International Congress of Phonetic Sciences, Glasgow, UK.

Ladd, B., 1980. The Structure of Intonational meaning: Evidence from English. Indiana University Press, Bloomington & London.

Ladd, D.R., 2008. Prosodic structure. Intonational Phonology. Cambridge University Press.

Landis, J.R., Kock, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159–174.

Luchkina, T., Cole, J., 2016. Structural and referent-based effects on prosodic expression in Russian. Phonetica 73, 279–313. doi: 10.1159/000449104.

Luchkina, T., Jyothi, P., Sharma, V., Cole, J., 2015. Prosodic and structural correlates of perceived prominence in Russian and Hindi. In: Proceedings of the International Congress on Phonetic Sciences, Glasgow.

Mahrt, T., 2016. LMEDS: Language Markup and Experimental Design Software. URL https://github.com/timmahrt/LMEDS.

Mahrt, T., *in press*. Acoustic Cues for the Perception of the Information Status of Words in Speech. Ph. D. thesis, University of Illinois Urbana-Champaign.

Mahrt, T., Cole, J., Fleck, M., Hasegawa-Johnson, M., 2012a. F0 and the perception of prominence. In: Proceeding of Interspeech 2012, Portland, Oregon.

Mahrt, T., Cole, J., Fleck, M., Hasegawa-Johnson, M., 2012b. Modeling speaker variation in cues to prominence using the Bayesian information criterion. In: Proceedings of Speech Prosody 6, Shanghai.

Mahrt, T., Huang, J-T., Mo, Y., Fleck, M., Hasegawa-Johnson, M., Cole, J., 2011. Optimal models of prosodic prominence using the Bayesian information Criterion. In: Proceeding of Interspeech, pp. 969–972.

Mo, Y., Cole, J., Lee, E-K., 2008. Näve listeners' prominence and boundary perception. In: Proceedings of Speech Prosody 2008, pp. 735–738. Campinas, Brazil.

Mo, Y., Cole, J., Hasegawa-Johnson, 2009. Prosodic effects on vowel production: Evidence from formant structure. In: Proceedings of Interspeech 2009, Brighton, UK.

Pfitzinger, H.R., 1998. Local speech rate as a combination of syllable and phone rate. In: Proceedings of ICSLP, pp. 1087–1090.

Pierrehumbert, J.B., 1980. The Phonology and Phonetics of English Intonation. Massachusetts Institute of Technology. doi: 10.1177/003368828401500113.

de Pijper, J.R., Sanderman, A.A., 1994. On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. J. Acoust. Soc. Am. 96, 2037–2047.

Pintér, G., Mizuguchi, S., Tateishi, K., 2014. Perception of prosodic prominence and boundaries by L1 and L2 speakers of English. In: Proceeding of Interspeech, pp. 544–547.

Pitrelli, J.F., Beckman, M.E., Hirschberg, J., 1994. Evaluation of prosodic transcription labeling reliability in the ToBI framework.. Third International Conference on Spoken Language Processing, pp. 123–126. Yokohama, Japan.

Pitt, M., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., Fosler-Lussier, E., 2007. Buckeye Corpus of Conversational Speech (2nd Release). Department of Psychology Ohio State University, Columbus, OH. (Distributor). URL www.buckeyecorpus.osu.edu .

Roy, J., Cole, J., Mahrt, T., (in review, 2017). Individual differences and patterns of convergence in prosody perception. Lab. Phonol. J. Assoc. Lab. Phonol.

Streefkerk, B.M., Pols, L.C.W., ten Bosch, L.F.M., 1997. Prominence in read aloud sentences, as marked by listeners and classified automatically. Proc. Inst. Phonetic Sci. 21, 101–116.

Swerts, M., 1997. Prosodic features at discourse boundaries of different strength. J. Acoust. Soc. Am, 101, 514–521.

Syrdal, A.K., Hirschberg, J., McGory, J., Beckman, M., 2001. Automatic ToBI prediction and alignment to speed manual labeling of prosody. Speech Commun. 33 (1), 135–151.

Turnbull, R., Royer, A.J., Ito, K., Speer, S.R., 2017. Prominence perception is dependent on phonology, semantics, and awareness of discourse. Lang. Cognit. Neurosci. doi: 10.1080/23273798.2017.

Veilleux, N., Shattuck-Hufnagel, S., Brugos, A., 2006. 6.911 Transcribing Prosodic Structure of Spoken Utterances with ToBI. January IAP 2006. Massachusetts Institute of Technology: MIT Open Course Ware, https://ocw.mit.edu. License: Creative Commons BY-NC-SA.

Wagner, P., 2005. Great expectations-Introspective vs. perceptual prominence ratings and their acoustic correlates. In: Proceedings of Interspeech 2005.

Wightman, C.W., Shattuck-Hufnagel, S., Ostendorf, M., Price, P.J., 1992. Segmental durations in the vicinity of prosodic phrase boundaries. J. Acoust. Soc. Am. 91, 1707–1717. doi: 10.1121/1.402450.

Wood, S.N., 2006. Generalized additive models : an introduction with R. Texts Stat. Sci. xvii, 392. doi: 10.1111/j.1541-0420.2007.00905_3.x.

Wood, S.N., 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J. R. Stat. Soc. Ser. B Stat. Methodol. 73, 3–36. doi: 10.1111/j.1467-9868.2010.00749.x.

Yoon, T.-J., Chavarria, S., Cole, J., Hasegawa-Johnson, M., 2004. Intertranscriber reliability of prosodic labeling on telephone conversation using ToBI. In: Proceedings of Interspeech, pp. 2729–2732. Jeju, Korea.

You, H.J., 2012. Determining prominence and prosodic boundaries in Korean by non-expert rapid prosody transcription. In: Proceedings of Sixth Speech Prosody.