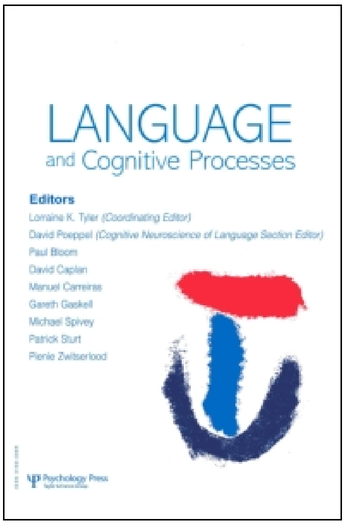

Language and Cognitive Processes
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713683153

The role of syntactic structure in guiding prosody perception with ordinary listeners and everyday speech

Jennifer Cole[a]; Yoonsook Mo[a]; Soondo Baek[a]
[a] Department of Linguistics, University of Illinois at Urbana-Champaign, Urbana, IL, USA

First published on: 11 June 2010

To cite this Article Cole, Jennifer , Mo, Yoonsook and Baek, Soondo(2010) 'The role of syntactic structure in guiding prosody perception with ordinary listeners and everyday speech', Language and Cognitive Processes, 25: 7, 1141 — 1177, First published on: 11 June 2010 (iFirst)

To link to this Article: DOI: 10.1080/01690960903525507

URL: http://dx.doi.org/10.1080/01690960903525507

LANGUAGE AND COGNITIVE PROCESSES
2010, 25 (7/8/9), 1141–1177

Psychology Press
Taylor & Francis Group

The role of syntactic structure in guiding prosody perception with ordinary listeners and everyday speech

Jennifer Cole, Yoonsook Mo, and Soondo Baek
Department of Linguistics, University of Illinois at Urbana-Champaign, Urbana, IL, USA

The relationship between syntactic and prosodic phrase structures is investigated in the production and perception of spontaneous speech. Three hypotheses are tested: (1) syntax influences prosody production; (2) listeners' perception of prosodic boundaries is sensitive to acoustic duration; and (3) syntax directly influences boundary perception, (partly) independent of the acoustic evidence for boundaries. Data are from the Buckeye corpus of conversational speech, and the real-time prosodic transcription of those data by 97 untrained listeners. Inter-transcriber agreement codes boundary strength at word junctures, and Boundary scores are shown to be correlated with both the syntactic context and vowel duration of a word. Vowel duration is also correlated with syntactic context, but the effect of syntactic context on boundary perception is not fully explained by vowel duration. Regression analyses show that syntactic clause boundaries and vowel duration are the first and second strongest predictors of boundary perception in spontaneous speech.

Keywords: Prosody perception; Prosody production; Syntax; Spontaneous speech; Prosody transcription.

Correspondence should be addressed to Jennifer Cole, Department of Linguistics, 4080 Foreign Languages Building, University of Illinois at Urbana-Champaign, 707 South Mathews, Urbana, IL 61801, USA. E-mail: jscole@illinois.edu

This work is supported by NSF award IIS-0414117 to the first author and Mark Hasegawa-Johnson. For their varied contributions to the work presented here, we thank three anonymous reviewers, Mark Hasegawa-Johnson, Chilin Shih, Tae-Jin Yoon, Xiaodan Zhuang, and other members of the Prosody-ASR research group, and to José Ignacio Hualde. Special thanks to Zachary Hulstrom and Eun-Kyung Lee for their help with the prosody transcription experiments. Statements in this paper reflect the opinions and conclusions of the authors, and are not endorsed by the NSF or the University of Illinois.

http://www.psypress.com/lcp DOI: 10.1080/01690960903525507

INTRODUCTION

Through prosodic phrasing, languages encode the grouping of words into constituents that cohere semantically (Frazier, Clifton, & Carlson, 2004; Selkirk, 1984), and which can express the preferred rhythmic and intonational patterns of the language (Nespor & Vogel, 1986; Schafer, Speer, Warren, & White, 2000; Watson & Gibson, 2004). The prosodic structure of an utterance, including both phrasing and the marking of prosodic prominence, influences its phonetic expression in many ways, with effects realised at the level of segmental properties (e.g., vowel formant patterns, consonant voicing), and suprasegmental properties (pitch, loudness and duration).[1] Prosody also influences speech comprehension (Cutler, Dahan, & van Donselaar, 1997; Frazier, Carlson, & Clifton, 2006) in that listeners are guided in the interpretation of the syntactic and semantic contents of an utterance by the prosodic structures encoded in its phonetic form.

A simple and direct model of the effect of prosody on sentence comprehension can be characterised as follows. The speaker produces a prosodic structure for an utterance that reflects the grouping of words into syntactic or semantic units. The prosodic structure, part of phonological form, is interpreted in the phonetic implementation, shaping the articulation and resulting in acoustic patterns that encode the prosodic elements marking prominence and phrasing. The listener perceives these acoustic patterns as cues to the prosodic structure produced by the speaker, and interprets the syntactic and semantic properties of the utterance in conformance with the perceived prosodic structure.

The direct syntax–prosody processing model as sketched above is an idealisation, and though it may serve as a useful model of speech processing in laboratory conditions, the facts of speech production and perception in situations of natural speech communication are somewhat more challenging. First, there is the fact that there are many factors that contribute to the assignment of prosodic structure, some of which are related to the linguistic form of the utterance (e.g., syntactic or semantic factors), and others which reflect the speaker's affect, communicative intent, or speaker-selected production factors such as rate or clarity. And, like any other aspect of

[1] Although phonetic implementation of prosody can be seen in evidence from both articulation and acoustics, our project is focused on acoustic correlates and their relation to prosody perception. There are numerous works reporting on a wide range of acoustic parameters as correlates of prosody, of which we cite a few here: Beckman (1986), Beckman and Pierrehumbert (1986), Ladd (1996/2008) for F0; Turk and Sawusch (1997) and Wightman et al. (1992) for duration; Kochanski, Grabe, Coleman, and Rosner (2005) for overall intensity; Heldner (2003) and Sluijter and van Heuven (1996) for spectral emphasis and balance (intensities in sub-bands); van Bergem (1993) for formant structures; Choi, Hasegawa-Johnson, and Cole (2005) for various harmonic and voice source parameters.

speech production, prosody production is also subject to disfluency (DISF) or error. Variability in prosody production that is not due to linguistic form is shown in the Schafer et al.'s (2000) study of English spontaneous speech produced in laboratory conditions, and in the Yoon's (2007) corpus study of radio news announcers' read productions of the same news script. The possibility of variable prosody for a given sentence, even in the absence of structural ambiguity, means that the listener can not have rigid expectations about the prosodic form of an utterance based solely on the word string and prior syntactic context.

A second challenge comes from the fact of inter- and intra-speaker variations in the phonetic encoding of prosody. The acoustic correlates of prosody are variable in all speech styles, but especially so in spontaneous speech. For example, while English speakers producing read speech are fairly consistent in encoding prosodic phrase boundaries through lengthening of the phrase-final syllable rhyme (Wightman, Shattuck-Hufnagel, Ostendorf, & Price, 1992; Yoon, Cole, & Hasegawa-Johnson, 2007), they vary in their use of intensity, F0, glottalisation of phrase-initial vowels (Dilley, Shattuck-Hufnagel, & Ostendorf, 1996), and creaky voicing (laryngealisation) in phrase-final position (Kim, Yoon, Cole, & Hasegawa-Johnson, 2006). Acoustic effects of prosody also vary according to the phonological content of the word or syllable, for example, as shown by Lee and Cole (2007) and Mo (2008), whose studies of radio news speech and spontaneous speech show variability in final lengthening effects on vowels as a function of vowel phoneme.

The third challenge to a model of direct syntax–prosody processing relates to the listener. Listener variability in the perception of prosody has not been discussed as widely in the prosody literature as has speaker variability, but can arise from a number of factors. First, just as there may be DISF or errors in prosody production, the same is true for perception. A listener may fail to detect acoustic cues to prosody that are present in the signal due to performance factors (attention and fatigue), or due to interference from environmental noise or activity. Variability in prosody perception among listeners can also arise due to differences in linguistic experience, e.g., from unfamiliarity with a speaker's voice or his/her phonetic expression of prosody. Evidence for perceptual variability can be seen in studies that report on inter-transcriber agreement rates on tasks of prosody transcription.

Much of the contemporary research on prosody evaluates the prosodic elements of an utterance by means of prosody transcription, using a transcription standard such as the Tones and Break Indices (ToBI) system (Beckman & Ayers, 1997), which is based on the autosegmental–metrical model of prosody (Beckman & Pierrehumbert, 1986; Ladd, 1996/2008). One method for evaluating the reliability of prosody transcription is to employ multiple trained transcribers in the annotation of a common set of materials, and then to assess the agreement rate among them. High agreement is taken to indicate that the transcribers were uniform and consistent in their interpretation of prosody according to the transcription guidelines. Furthermore, if multiple transcribers produce the same prosody annotation for a given utterance, it is assumed that there were sufficiently salient cues to that prosodic structure in the speech signal. Ostendorf, Price, and Shattuck-Hufnagel (1995) conducted a reliability study for the ToBI system using transcriptions of radio broadcast news speech, reporting inter-transcriber agreement rates as high as 91% for the location of prosodic phrase boundaries, and 60% for the location of pitch accent. Dilley, Breen, Bolivar, Kraemer, and Gibson (2006) conducted a similar test of reliability for ToBI transcriptions on a small-scale, combined corpus of radio broadcast news speech and telephone conversational speech, reporting agreement rates of 88% on the location of prosodic phrase boundaries, and 87% on the location of pitch accent. Using a simplified ToBI transcription system (marking only the location and not the tonal type of boundaries and prominent words), Yoon, Chavarría, Cole, and Hasegawa-Johnson (2004) compared transcriber agreement rates for telephone conversational speech, with agreement rates at 89% for boundaries and 86% for pitch accent.

The inter-transcriber agreement rates reported from ToBI reliability studies are well above chance levels, and are taken as validation that the transcription method is generally reliable. But at the same time, they also reveal listener variability, even under idealised listening conditions. The transcription task for each of these studies provides the transcriber with a wealth of information on which to judge prosody, including visual displays of the speech waveform, pitch and intensity tracks and spectrogram, along with an auditory signal, and the transcriber is allowed to listen to the signal as many times as needed, stopping and starting at any location to focus on difficult regions. Transcribers typically undergo rigorous training in preparation for their task, including group discussion and resolution of problem cases. These conditions are far from similar to the conditions under which ordinary listeners perceive and interpret prosody, but even so, these trained transcribers disagree on the prosodic structure of an utterance for 10–30% of its content (counted in words).

There are many differences between the task of prosody transcription by trained transcribers and ordinary prosody perception in the course of everyday speech communication. If we are to take research findings based on expert prosody transcription as representative of what takes place in ordinary speech processing, then we must first establish a parallel between expert transcribers and ordinary listeners in their perception of prosody. A question of particular interest, and the focus of this paper, is whether untrained listeners are guided in their perception of prosody (specifically, prosodic prominence and prosodic phrase boundaries) by information beyond the phonetic form of the utterance, specifically, by information from the syntactic context. Lacking visual information from a graphical speech display or expert linguistic knowledge, is the untrained listener influenced by the syntactic, semantic, and discourse context when making explicit judgements about the prosodic elements in a naturally produced speech utterance? For instance, does the listener have expectations about the prosodic structure of an utterance based on the discourse context, which establishes the information status of each word and phrase, or based on his/her prior experience with utterances containing similar discourse contexts, similar words, or similar syntactic content? And although expert transcribers are trained to focus narrowly on the phonetic evidence for prosody, is there also a possibility of extra-phonetic factors influencing them, too?

The present paper focuses on the relationship between the perception of prosodic phrase boundaries and the syntactic form of an utterance, and is part of a larger study investigating how untrained listeners perceive prosody in spontaneous speech. The goal of this study is to determine the relative strength of (acoustic) phonetic and extra-phonetic factors as correlates of perceived prosody. The approach relies on prosody transcription using multiple, untrained transcribers (15–22 per utterance), and a set of probabilistic prosody features that encode the strength of the prosodic feature (prominence or boundary) based on the proportion of transcribers who perceive that feature. The component of this study presented below addresses the following questions:

- *Variability by listener and speaker.* Do untrained listeners agree with one another in their perception of prosody for a given utterance (listener variability)? Are there differences across speakers in the patterns of prosody as perceived by listeners (speaker variability)?
- *Syntax–prosody association.* What is the relationship between the syntactic properties of an utterance and perceived prosody? How closely do prosodic boundaries perceived by untrained listeners reflect syntactic boundaries, and are there differences among syntactic boundary types in their correlation with perceived prosody? Are there asymmetries in the syntax–prosody correlation for boundaries at the right vs. left edge of syntactic and prosodic domains?
- *Acoustic duration as a cue to syntactic or prosodic phrases.* If syntactic structure is correlated with prosodic phrase structure as perceived by ordinary listeners, is that correlation mediated through the phonetic form?

We turn in the following section to a brief review of prior research on the link between prosody and syntax, both in production and perception. These works paint a complex picture of prosody processing, where a range of linguistic factors, including syntactic factors, interact to determine the prosodic form of an utterance.

LINGUISTIC FACTORS THAT INFLUENCE PROSODY

Prosodic phrases group words together in units that reflect syntactic structure, but syntactic structure does not fully predict prosodic phrasing. This claim follows directly from the observation that speakers vary in the prosody they assign to syntactically similar forms (Schafer et al., 2000; Yoon, 2007), and is reflected in a number of production models that accommodate misalignment of prosodic and syntactic units (Bachenko & Fitzpatrick, 1990; Gee & Grosjean, 1983; Watson & Gibson, 2004, among others). Linguistic theories of prosodic phrase structure recognise patterns of alignment between the edges of prosodic and syntactic constituents, but these constituents are frequently not co-extensive (Selkirk, 1984, 1986, 2000; Shattuck-Hufnagel & Turk, 1996). Yet despite the frequent mismatch between syntactic and prosodic phrases, the syntactic properties of an utterance do play a significant role in determining prosodic structure in speech production. For example, the part-of-speech of a word and the type of phrase it belongs to (e.g., main clause, subordinate clause, adjunct, parenthetical) influence the likelihood that a prosodic phrase boundary will appear before or after the word (Price, Ostendorf, Shattuck-Hufnagel, & Fong, 1991; Schafer et al., 2000; Selkirk, 1984, 1986; Watson & Gibson, 2004).

Syntax is not the only linguistic factor that influences the prosodic structure associated with an utterance. Semantic factors on prosodic phrasing are invoked in Selkirk's (1984) Sense Unit Condition (*elements in an intonational phrase must cohere semantically*) and Frazier et al.'s (2004) Semantic Coherence Constraint (*don't put semantically unrelated elements in the same prosodic phrase*). Other factors influencing prosodic phrasing relate to phonological structure, and include the length of the syntactic phrase, the distance from the previous prosodic boundary, and the location of the nuclear pitch accent (Nespor & Vogel, 1986; Watson & Gibson, 2004). The syntactic, semantic and phonological factors, along with factors related to speaking rate and style, collectively influence the prosodic structure assigned to an utterance, as demonstrated in Calhoun's (2006) work on the automatic prediction of prosody.

Given the complexity of the system that underlies the assignment of prosody, we might expect that the empirical evidence for the role of any single factor will be blunted by the effects of interacting factors. From the perspective of the listener, we may ask if and how acoustic cues to prosody are interpreted in terms of any single conditioning factor. How does the listener determine if the acoustic evidence for a prosodic phrase boundary reflects the syntactic, semantic, or phonological context? Evidence from speech comprehension studies establishes that prosody does contribute to syntactic interpretation (Price et al., 1991; Schafer et al., 2000; Weber, Grice, & Crocker, 2006), but the listener's evaluation of syntactic structure at any location in the utterance seems to depend on the integration of prosodic cues over a larger domain (Frazier et al., 2006). Frazier et al. (2004) state that syntactic structure on its own doesn't force or block any pattern of prosodic phrasing, and they as well as Pynte (2006) claim that a prosodic break may occur in almost any location in a sentence, if motivated by phonological and/or semantic factors.

The production studies cited above find evidence that syntactic structure is (at least partially) encoded in prosodic structure, while the perception and comprehension studies find corresponding evidence that perceived prosody influences the interpretation of syntactic structures. These findings are compatible with the idealised direct syntax–prosody processing model, as described in the Introduction section. The speaker produces a prosodic structure that reflects the syntactic phrasing of the utterance, encoding the prosodic structure in the phonetic implementation of the utterance. The listener perceives the phonetic signal, decodes the prosodic information and uses it to guide an analysis of the syntactic structure for the utterance. The listener's task is complicated by the fact that syntax is not the sole determinant of prosodic structure, so inferences about syntactic structure based on perceived prosody must be tempered by consideration of other factors that shape prosodic structure. For example, the presence of a prosodic phrase boundary preceding sentence-final PP might signal high attachment of the PP, but it may also reflect purely phonological factors, such as the length of the preceding NP, suggesting that the impact of perceived prosody on syntactic judgements must be weighted by taking into account the phonological context (and possibly other factors). Nonetheless, to the extent that syntactic judgements are significantly correlated with perceived prosody, we find support for the model of direct prosody–syntax processing.

TESTING THE PROSODY–SYNTAX RELATIONSHIP IN SPONTANEOUS SPEECH

Our study seeks new empirical evidence for the prosody–syntax relationship, by investigating the perception of prosody in spontaneous, conversational speech. Two hypotheses of the direct prosody–syntax processing model are tested in the experiment presented below: (1) syntactic phrasing influences the prosodic phrase structure a speaker assigns to an utterance, and (2) listeners respond to acoustic cues to prosody in their judgement of the prosodic structure of the utterance. (An additional hypothesis from this model, that the listener's judgement of the prosodic structure influences syntactic interpretation, is not addressed here.) These hypotheses are tested using speech data from a corpus of spontaneous, conversational speech. Perception data are in the form of prosody judgements on these materials, obtained from untrained listeners performing real-time prosodic transcription. Acoustic measures of vowel duration provide production data, and a manual annotation of syntactic phrase structure provides data on the syntactic context of each word in the corpus.

The first hypothesis is tested in two ways: an indirect measure of the influence of syntax on prosody production is from the listeners' perspective, in the relationship between syntactic structure and perceived prosody, and a more direct measure uses acoustic evidence of prosody production, in the correlation between syntactic phrase structure and the acoustic correlates of prosodic phrasing. If syntax influences prosody production, then we expect to observe acoustic effects of prosodic phrase boundaries in locations predicted by the syntax.

We focus our analysis on acoustic effects of prosody in the lengthening of the phrase-final syllable rhyme, which is one of the most studied and robust effects of prosodic phrasing in English (Lehiste, 1972; Wightman et al., 1992; Yoon et al., 2007). Acoustic evidence of final lengthening is measured in the duration of word-final stressed vowels.[2] Unstressed vowels are excluded to avoid durational effects of unstressed vowel reduction, but since over three-quarters of the words in the corpus under study are monosyllabic, the majority of word-final vowels are also stressed.[3]

[2] In addition to durational effects of prosody, we also find overall intensity (Root Mean Square) and acoustic measures of creaky voicing (H1*–H2* and H2*–H4*) to be significantly correlated with the perception of prosodic phrase boundaries for at least some vowel phonemes. We report only the duration findings here, as duration was not only the strongest correlate (based on Pearson's *r* values), but is also the only acoustic measure that is significantly correlated with boundary perception across most of the vowel phonemes (Mo, 2008). Pause duration is also expected to cue prosodic boundaries, but we have not yet examined pause duration in our materials. For the data reported here, speech excerpts were selected to minimise the occurrence of disfluency within the excerpt, where silent and filled pauses were one of the factors used to identify disfluency. We expect that this selection criterion has skewed the distribution of pause duration at prosodic juncture in these materials. In our ongoing work we are investigating the influence of pause duration on prosody perception with longer excerpts for which pause duration was not a selection criterion.

[3] Prosodic prominence also conditions lengthening of a stressed vowel (e.g., Turk & Sawusch, 1997), so the duration measure examined here may in some cases exhibit combined effects of prominence and boundary lengthening. Prominence is coded in our data with a probabilistic *P*-score assigned to each word, parallel to the assignment of B-scores, which means that we can not simply separate prominent (pitch-accented) words from non-prominent words (unaccented), as has been done in prior studies that are based on ToBI-style prosody transcription. Instead, we

(*continued on next page*)

The second hypothesis is tested by measuring the correlation between the listeners' judgements of prosodic phrase boundaries and stressed vowel duration as the acoustic correlate of prosodic phrase boundary. If listeners are sensitive to the acoustic encoding of prosodic structure, their judgements of prosodic phrase boundary location should coincide with the acoustic evidence of final lengthening. We expect to find longer stressed vowels in words that are perceived as final in the prosodic phrase.

Our study takes the direct prosody–syntax processing model one step further, and tests a third hypothesis about the role of syntax in prosody perception: (3) syntactic phrasing is a direct influence on listeners' perception of prosodic phrase structure. This hypothesis is motivated by considering the effect of the listener's prior linguistic experience on the perception of a new utterance. If the listener's prior experience suggests a strong relationship between syntactic phrase structure of a certain type and the occurrence of a prosodic phrase boundary, then this association could in principle bias the listener to hear a prosodic boundary in the presence of the triggering syntactic form, even if the acoustic evidence for a prosodic boundary was weak or absent. Hypothesis 3 is tested first through the correlation between syntactic phrase structure and listeners' judgements of prosodic phrase boundaries. We expect to find a greater incidence of perceived prosodic phrase boundaries at locations where there is a high-level syntactic phrase boundary (e.g., a clause boundary). A significant correlation could arise due to a direct influence of syntax on prosody perception, but it could also arise if the acoustic evidence supports a judgement of prosodic phrase boundary at locations that coincide with syntactic phrase boundaries. To test for the independence of syntactic phrase structure on prosody perception, we conduct regression analyses with syntactic phrase boundaries and acoustic duration measures as predictors of perceived prosodic phrase boundaries. As shown below, the evidence from regression analysis supports the claim

use correlation and regression analysis to look at the relationship between duration, B-scores and *P*-scores. Comparison of correlation coefficients between duration and B-scores (Kendall's tau = .369) vs. duration and *P*-scores (Kendall's tau = .243) shows that duration is more strongly correlated with B-scores (all duration measures are normalised via *z*-transform). The correlation between *P*-scores and B-scores is even weaker (Kendall's tau = .204). Furthermore, regression analysis shows that *P*-scores only very weakly predict B-scores ($r^2 = .027$). Stepwise regression analysis shows that duration is the primary predictor of B-scores ($r^2 = .239$; shown in Table 6, Model B) and *P*-scores as a second factor contribute only marginally as a predictor ($r^2 = .008$). Looking at it from the perspective of duration modelling, we also find that B-scores are stronger predictors of vowel duration ($r^2 = .278$) with *P*-scores again as weak predictors ($r^2 = .039$). These findings demonstrate that boundary effects on duration outweigh prominence effects, and thus that boundary lengthening effects on words marked as prominent cannot be solely attributed to prominence-based lengthening.

that syntactic structure makes an independent contribution to the perception of prosodic phrase structure, beyond that of acoustic vowel duration.[4]

METHODS AND MATERIALS

The speech materials used for the prosody transcription experiment were drawn from the Buckeye corpus of spontaneous, conversation-style speech collected from interviews with adult speakers from Columbus, OH (Pitt et al., 2007). Two excerpts of between 11 and 22 seconds long were extracted from the interviews of 36 speakers, for a total of 72 excerpts, and two similar excerpts taken from one other speaker were used for an initial practice transcription. The excerpts were extracted at junctures between talker turns, often at topic junctures, and always at a natural prosodic break, typically marked by pause. Four stimuli sets were created from these materials. Two sets of test excerpts were constructed each containing one of the two excerpts from 18 speakers, with a third and fourth set containing one of the two excerpts from a second set of 18 speakers. Each test excerpt appeared in only one set, and each speaker is represented by no more than a single excerpt within a set. Each set also contained a practice excerpt from the extra speaker. The test excerpts were randomly sequenced in each set. Orthographic transcripts were created for each excerpt, removing all punctuation and capitalisation, and including

[4] A reviewer asks about the possibility of directly testing the independence of acoustic and lexico-syntactic cues to prosody by testing prosody perception with delexicalised speech—using filters or transformations of the acoustic signal to remove segmental information that reveals the lexical content of the speech. This approach is illustrated in the work of de Pijper and Sanderman (1994) who tested prosodic boundary perception by untrained listeners with delexicalised speech materials which were created by resynthesising speech after replacing the first eight spectral peaks with peaks of fixed frequency and bandwidth, and also manipulating pitch and Linear Predictive Coefficient (LPC) gain. This manipulation has the effect of rendering every vowel as schwa-like in its spectral features, and eliminating consonantal distinctions. The resulting materials were judged by de Pijper and Sanderman to preserve prosodic cues while rendering the utterances otherwise unintelligible; and the procedure was considered more successful than simpler alternative methods involving only low-pass filtering or spectral inversion. We have also considered methods for delexicalisation in our work on prosody perception, but like de Pijper and Sanderman, we have been dissatisfied with the filtering methods we have tested thus far, which were either unsuccessful in removing segmental cues to lexical content or successful in delexicalisation but with distorted or very unnatural sounding prosody. We did not attempt the complex method of spectral peak substitution used by de Pijper and Sanderman, which is unsuitable for our purposes given that we are interested in both segmental and suprasegmental effects of prosody. A related suggestion from this reviewer was to ask transcribers to mark prosody on the text without listening to the associated speech file, in which case lexico-syntactic features alone would guide the annotation. We did not collect such data in the initial phase of this project, whose findings are presented here, but are currently doing so for the second phase of data collection, and expect to report on the findings in our future work.

transcription of DISF and filled pauses. The transcriptions were sequenced to match the auditory presentation of the recorded excerpts, and printed on multiple sheets of paper using 14 pt, Times New Roman font. Listeners provided a coarse prosody annotation for each excerpt by making marks on the printed transcription sheet. The prosody annotation involved marking the location of prosodic phrase boundaries and prominent words in separate tasks (the data on prominence perception are not discussed further here).

Between 15 and 22 listeners transcribed each excerpt for its prosodic content, and the transcriptions were pooled from the entire group to obtain a population-wise, probabilistic measure of the prosodic status of each word. This method is adapted from similar methods used by Buhmann et al. (2002), Streefkerk, Pols, and ten Bosch (1997, 1998), and Swerts (1997).

A total of 97 listeners took part in the transcription experiment, all undergraduate students recruited from introductory level linguistics courses at the authors' home institution. Listeners had no prior training in prosody transcription or prosodic phonology and were monolingual, native speakers of American English. The majority were residents of Illinois from childhood. Data from six listeners were excluded due to failure to follow the transcription guidelines or because they were found not to be monolingual. Listeners were randomly assigned to one of four test groups, corresponding to the four sets of speech materials. Listeners were seated individually at computers in a classroom and told that they would listen through headphones to excerpts from recorded interviews with speakers of American English, and use a pencil to mark the transcript indicating the grouping of words into "chunks" of speech. The experimenter defined chunking by reading the short script in Appendix 1, but no example sound file was played to demonstrate what chunk boundaries might sound like. Subjects were also told that there is no single "right" transcription for an excerpt, and they should not be concerned with how one person's transcription compares to anyone else's.

The participants listened to each excerpt twice in succession, marking the location of "chunk" boundaries with vertical lines between words on the printed transcript. The transcription was performed in real time, and was solely based on auditory impression; listeners were not aided by any visual, graphic display of the speech signal. Listeners also were neither able to start or stop the auditory presentation, nor could they repeat any excerpt after the second presentation. Listeners were allowed to mark changes to their prosody annotation, by striking through a mark to "delete" it, and by circling a stricken mark to recall it. No further changes could be recorded.

Data coding

For each word a probabilistic Boundary score (B-score) is calculated based on the proportion of transcribers from the total group (15–22 per excerpt set), who perceived a prosodic boundary (a juncture between “chunks”) following that word. B-score values are between 0 and 1, with extreme high or low scores indicating greater agreement among transcribers, presumably reflecting lesser ambiguity in the prosodic organisation of the utterance, and/or the presence of more salient cues to the presence or absence of a prosodic boundary. Figure 1 shows an example of a partial excerpt, plotting the B-score for each word. This example illustrates a typical finding, namely, that there are many words that transcribers agree are *not* before a boundary, and many fewer words where transcribers reach the same rate of agreement on the positive assignment of a boundary.

In addition to the coding of B-scores, each word is annotated for the highest level syntactic boundary that coincides with its right and left edges, separately, based on a manual syntactic parse based on Penn Treebank annotation guidelines (Marcus, Marcinkiewicz, & Santorini, 1993). The syntactic categories specified in the parse are listed in Table 1.

A sample coding of a partial excerpt is shown in Table 2. The full excerpt from which that sample was taken is shown in (1), with bracketing of the final portion reflecting the location of perceived boundaries labelled by two or more transcribers. This example illustrates the fact that the juncture between any two successive words is coded twice, as a left edge and as a right edge. Consequently, the left and right edge B-scores in the coding are matched: every left edge B-score corresponds to the right edge B-score of the preceding word.

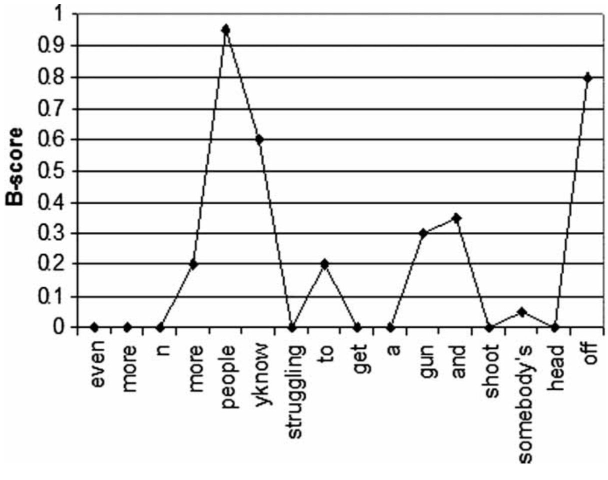

Figure 1. Graph of probabilistic boundary (B) score for each word in a sample utterance from the test corpus. B-scores represent the proportion of transcribers (out of 20 for this excerpt) who perceived a prosodic boundary following the given word.

TABLE 1
Syntactic categories used in the manual annotation in which each word was assigned a "left-edge" label for the highest level syntactic category whose left edge coincides with the left edge of the word

S	matrix S L: uh | *it's just you ...* R: still *live in the city of Columbus* | uh...
S-BAR	subordinate or relative clause L: the fact | *that it was something...* R: *when our kids grew up* | then there was so...
S2	S preceded by (non-coordinating) conjunction or relative pronoun L: the fact that | *it was something...* R: *not defined*
CC-S	coordinating conjunction following or preceding a sentence L: we had standards | *and* there were certain things... R: we had standards *and* | there were certain things...
CC-XP	coordinating conjunction preceding or following an XP L: we did | *or* didn't... R: metal detectors *and* | police
Phrase (XP)	any XP that is not a clause (examples with VP) L: people yknow | *struggling to get...* R: I *like Columbus* | also
Within phrase (W/P)	any word boundary that does not align with a coded syntactic boundary (a non-initial (L) or non-final (R) word) <rarely occurs with boundary label>
Disfluency	filled pause, repetition & repair disfluencies L: in 1981 | *uh* lived... R: and then the protestant | *the uh...*
Discourse marker (DM)	*yknow, like, so, I mean, ...* L: the uh | *sorry* the uh... R: organ playing and *yknow* | praise the Lord...

Note: A "right-edge" label was similarly assigned to the right edge of each word. Other categories with counts less than 10 in this dataset are not shown.

(1)

<start of excerpt> they don't have the money to throw away and like catholic schools they don't have the money to put on those programs so some of those people kind of get pushed aside and so yknow as you can see I'm standing on m soapbox here I do like Columbus public schools but yknow i i i hate when people say yknow look at these test scores [because it really doesn't reflect what's there] [because] [it's like saying] [all kids carry guns to school <end of excerpt>

RESULTS I: LISTENER AND SPEAKER VARIABILITY

To assess the reliability of the prosody transcriptions obtained from untrained listeners, the agreement rate between all transcribers assigned to

TABLE 2
A partial excerpt showing the word sequence (from top to bottom), the left- and right-edge syntactic category for each word, and the B-score associated with that word edge

Left-edge B-score	*Left-edge syntactic category*	*Word*	*Right-edge syntactic category*	*Right-edge B-score*
0.25	**SBAR**	because	SubConj	0
0	S2	it	NP	0
0	VP	really	ADVP	0
0	VP2	doesn't	W/P	0
0	VP2	reflect	W/P	0
0	SBAR	what's	REL	0
0	ADVP	there	**SBAR**	0.45
0.45	**SBAR**	because	**SC**	0.75
0.75	**S2**	it's	W/P	0
0	SBAR	like	SC	0
0	S-ING	saying	**W/P**	0.1
0.1	**SBAR**	all	W/P	0
0	W/P	kids	NP	0
0	VP	carry	W/P	0
0	NP	guns	VP	0
0	PP	to	W/P	0
0	NP	school	SBAR	.

Note: Bold category labels mark locations where two or more listeners marked a right or left prosodic ("chunk") boundary. See Table 1 for definitions of the primary syntactic labels; labels appearing here but not in Table 1 are low-frequency labels in our dataset and are not included in the analysis (e.g., SC "subordinating conjunction").

the same excerpt set was calculated using two measures. Cohen's kappa coefficient (Cohen, 1960) measures the agreement between a pair of transcribers for each word, comparing the boundary labels (boundary and no-boundary). Cohen's kappa is considered a better measure of agreement than the simple percent agreement rate, because it factors in the chance probability of agreement based on the most frequently occurring label. The suggested interpretation of the kappa statistic is that values between 0.41 and 0.6 indicate moderate agreement, and higher values indicate substantial to perfect agreement (Landis & Koch, 1977). We also use Fleiss' kappa statistic (Fleiss, 1981; see also Artstein & Poesio, 2008), which calculates agreement between multiple transcribers over the level expected by chance, and its z-transform for significance testing.

The full results of the reliability analysis, including prominence and B-scores, are presented in Mo, Cole, and Lee (2008). The results for B-scores are summarised here. Cohen's kappa coefficients over the transcriber pairs in each set ($N = 1{,}322$ pairs) show a normal distribution with values ranging from .240 to .850, and a mean kappa of .582, indicating a mean level of moderate agreement above chance. Fleiss' kappa coefficient was calculated separately for the entire group of transcribers assigned to each of the four excerpt sets and Fleiss' kappa coefficients indicate moderate or better agreement above chance levels. The z-normalised kappa scores were tested for significance at a 99% confidence level ($z = 2.32$), and all scores were highly significant ($p < .001$; Set 1: $\kappa = .61$, $z = 19.43$; Set 2: $\kappa = .54$, $z = 21.87$; Set 3: $\kappa = .62$, $z = 25.05$; Set 4: $\kappa = .58$, $z = 26.22$).

From the reliability analysis we conclude that untrained listeners are reliably consistent and systematic in their perception of prosodic boundaries in spontaneous, conversational speech. The distribution of B-scores over all the words in the dataset reveals that agreement is highest for words that listeners judge to be medial in a prosodic phrase, i.e., words with a B-score of zero, and this is the B-score for about 70% of the words in the dataset. The remaining 30% of words are perceived as being final in a prosodic phrase by one or more transcribers (B-score > 0), and the count of words with B-scores at a given level decreases as the B-score increases (i.e., there are few words, which many or all transcribers agree are final in a prosodic phrase).

Looking at the distribution of B-scores across speakers provides an indirect measure of speaker variability in prosody production, because although listeners vary in their perception of prosody, the same group of listeners have transcribed a given set of excerpts, so any differences in the patterns of perceived prosody between speakers can be taken to reflect the individual speaker's implementation of prosody. Figure 2 illustrates speaker variability in a plot of the mean interval between boundary marks and the mean interval between prominent words based on each listener's transcription of each speaker, individually. This plot indicates roughly how frequently listeners hear prominent words and prosodic phrase boundaries for a given speaker, and reveals variation across speakers. For some speakers, listeners perceive on average short intervals between prosodic phrase boundaries, while for other speakers the interval is longer. This variability could be due to differences between speakers in the length of prosodic phrases they assign to their utterances, and/or to the salience of the phonetic cues to prosodic boundary.

An interesting observation from Figure 2 is that speakers vary in the relative intervals between prominent words and boundaries in their speech, as perceived by listeners. Speakers in the group at the left end of the plot have a shorter mean interval between perceived boundaries than between prominent words, indicating that at least some of the prosodic phrases perceived by listeners contain no prominent word, contrary to the prediction from linguistic models of prosody. Speakers in the middle group have roughly equal mean intervals between prominent words and boundaries, which is consistent with a pattern in which listeners identify one prominent word (maybe the nuclear prominence) in each prosodic phrase. The group of speakers at the right of the plot, which is the largest group, have a longer

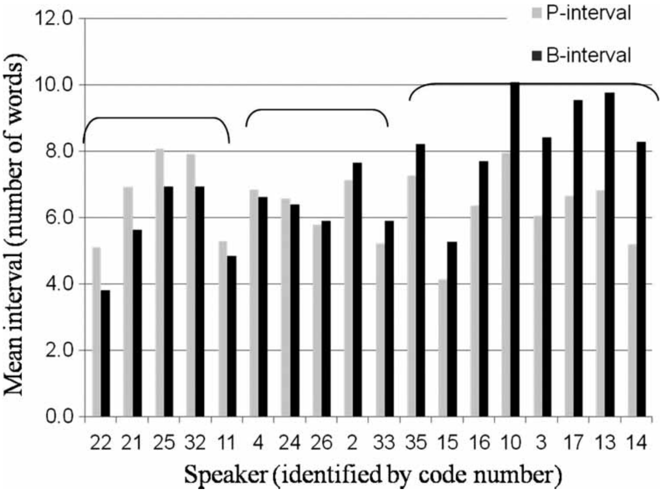

Figure 2. Plot of the mean interval between perceived prominences (P-interval) and boundaries (B-interval) for 18 speakers is shown individually. Intervals are measured in number of words between successive boundaries and prominent labels on an individual transcription, and mean intervals are calculated based on transcriptions from all the transcribers assigned to this excerpt set (16 transcribers for this set), for each speaker. Speakers divide roughly into three groups, as shown, based on the relative length of P-intervals and B-intervals, as described in text.

mean interval between boundaries than between prominent words, indicating that at least some prosodic phrases contain multiple prominent words.

The results presented in this section show that, despite the variability across listeners in the perception of prosodic phrase boundaries, listeners agree on the location of boundaries at levels well above chance. Boundary perception is at least partly systematic, and the next two sections present results showing the contribution of syntactic and acoustic information in predicting listeners' perception of prosodic phrasing. The distribution of B-scores across speakers suggests significant inter-speaker variability as well, which underscores the challenge for the listener in utilising prosodic information to interpret syntactic structures. If speakers vary in the length of prosodic phrases they employ, this might result in a different pattern of association between syntactic and prosodic phrases for different speakers, and listeners must be sensitive to these speaker-based differences to make appropriate use of prosodic cues.

RESULTS II: BOUNDARY PERCEPTION BY SYNTACTIC CATEGORY

Hypothesis 1 of this study states that syntactic phrasing influences the production of prosody. The prediction from this hypothesis is that prosodic phrases will be most likely to occur at the edges of syntactic constituents, with greater probability for higher-level constituents such as clauses than for lower-level constituents such as NPs or VPs. This hypothesis is confirmed indirectly (using perceptual data) in counts of the frequency of boundary marking at the edges of syntactic categories of different types.

Table 3 displays the frequency of boundary marking at the left edge of each syntactic category for about half of the data in this study (excerpts from Sets 1 and 2, the data that were annotated for syntactic structure), under two criteria of boundary marking. The left part of the table (columns 3–5) shows the frequency of boundaries that are labelled by at least half of the transcribers who listened to this subset of the data (18 transcriber for each set), and the right part (columns 6–8) adopts a weaker criterion for boundary marking, counting all boundaries labelled by one or more transcribers. There are two trends to notice in these data. First, the shaded cells mark those syntactic categories with the highest rates of boundary marking (over 25%) at their left edge. Three categories stand out under both criteria for boundary: discourse marker (DM), DISF, and coordinating conjunctions that conjoin clauses (CC-S). Listeners are very likely to perceive the beginning of a prosodic phrase at these locations. Under the weaker criterion of boundary, clauses (S and SBAR) and coordinating conjunctions joining other categories (e.g., CC-NP, CC-VP) also have high rates of boundary marking. Other syntactic categories are less often or even rarely perceived as locations where prosodic phrases begin. Notably, the left edge of a phrase-medial word (“within phrase”, W/P) is the least likely location for a prosodic phrase under the stricter criterion of boundary, or is among the three least likely locations under the weaker criterion. Given the large number of W/P syntactic labels (410, or 23% of the total number of syntactic edges), the rarity of perceived prosodic boundaries internal to a syntactic constituent is a significant, though unsurprising pattern in these data.

The bolded percentages in columns 5 and 8 highlight those categories whose left edges contribute more than 10% to prosodic boundary marking. Under both criteria of boundary, the left edge of matrix sentences, conjunctions joining sentences, DISFs, and DMs are the top contributors to perceived boundaries. Under the weaker criterion, the left edges of subordinate or relative clauses (SBAR), noun phrases, and phrase-medial words (W/P) also contribute to perceived boundaries. The finding for phrase-medial words is somewhat surprising. If we think that speakers do not, in fluent speech, actually construct prosodic phrases that begin in the middle of a syntactic constituent, then the occurrence of boundary marks at such locations may be viewed as the error rate for this transcription task with this type of speech. We will see below that the agreement rate (reflected in the B-score) for boundaries in phrase-medial locations is very, very low, lending further support to the view that such boundaries are marked (or perceived)

TABLE 3
The frequency of boundary marking at the *left edge* of each syntactic category

1. Category type	*2. Total number of left edges*	*3. Number of edges with B-score ≥0.5*	*4. Percentage of B-marked edges in syntactic category (%)*	*5. Percentage of B-marked out of total number of B-marked edges (%)*	*6. Number of edges with B-score >0*	*7. Percentage of B-marked edges in syntactic category (%)*	*8. Percentage of B-marked out of total number of B-marked edges (%)*
VP	234	12	5.1	5.19	22	9.4	5.70
ADVP	65	2	3.1	0.87	8	12.3	2.07
ADJP	22	1	4.5	0.43	1	4.5	0.26
W/P	410	9	2.2	3.90	42	10.2	**10.88**
PP	147	6	4.1	2.60	32	21.8	8.29
DM	60	27	45.0	**11.69**	56	93.3	**14.51**
NP	327	22	6.7	9.52	62	19.0	**13.99**
VP	234	12	5.1	5.19	32	13.7	8.29
DISF	78	32	41.0	**13.85**	69	88.5	**17.88**
S	237	52	22.5	**22.51**	117	49.4	**18.13**
CC-NP	19	2	10.5	0.87	14	73.7	2.59
CC-S	76	35	46.1	**15.15**	75	98.7	**13.86**
CC-VP	15	3	20.0	1.30	11	73.3	2.03
SBAR	112	16	14.3	6.93	55	49.1	**10.17**
Total	2036	234		100.00	596		100.00

Note: Column 2 displays the total number of left edges for each syntactic category in a subset of the data representing 18 speakers (excerpt Sets 1 and 2). Column 3 gives a count of those syntactic category left edges that coincide with a prosodic phrase boundary with a B-score of 0.5 or greater, and column 4 represents the same number as a percentage of the total for that category (=(col. 3/col. 2) × 100). Column 5 represents the number in the third column as a percentage of the total number of left-edges boundaries (=234 in this example). Columns 6–8 are the same as columns 3–5, but with the threshold for establishing boundary marking set lower, at one or more transcribers. Bold and shaded cells are discussed in the text.

in error. Finally, the totals at the bottoms of columns 3 and 6 show that boundaries are perceived at the left edges of about 11% of words under the strict criterion, and at about 29% of the words under the weaker criterion.

Table 4 presents the same frequency data for boundary marking at the right edge of syntactic categories.[5] The percentage of boundary marked edges within each syntactic category under the stronger criterion of boundary (column 4) is similar to the pattern observed for left edges in Table 3, with two differences: the right edges of PPs and SBAR are frequently perceived as the end of a prosodic phrase, while the right edges of conjunctions joining clauses (CC-S) are not. Strikingly, under the weaker criterion of boundary (column 7), nearly every syntactic category right edge is frequently perceived as a prosodic phrase boundary, with only phrase-medial words (W/P) and noun phrases displaying a lower rate of boundary perception. The overall high rate of boundary marking (under the weak criterion) at most syntactic right edges (excepting NP edges) suggests that almost any syntactic edge is a potential location for a prosodic phrase boundary, as noted by Frazier et al. (2004) and Pynte (2006). In their contribution to the total number of perceived boundaries (columns 5 and 8), matrix sentences again stand out, this time for their right-edge location, and DMs and DISF also contribute substantially. Unlike with the left-edge boundaries, right edges of noun phrases do not contribute much to boundary perception, and phrase-medial locations are even more likely to be heard as prosodic boundaries.

Overall, the frequency data in Tables 3 and 4 establish that there is a higher rate of boundary marking at the edges of higher-level constituents (matrix sentences) compared to lower-level constituents, and boundaries marked at higher-level syntactic edges contribute more to the total number of perceived prosodic boundaries.

To examine these patterns of boundary marking more closely, we consider next the distribution of B-scores by syntactic category. Figure 3 displays mean B-scores over all the words in the dataset, grouped according to the left- and right-edge syntactic category label of each word. Recall that every word is coded for two B-scores, a left-edge score codes the proportion of transcribers who marked a boundary preceding the word, and a right-edge score similarly coding boundary marking following the word. In the same

[5] Although there are an equal number of left and right syntactic edges coded in this dataset (each word contributes one left and one right edge), the total number of left and right syntactic edges are not equal in Tables 3 and 4 due to categories that are omitted because they have fewer than 10 instances in the dataset, or because they are not coded for both left and right edges. Examples of the latter are the right edges of subordinating conjunctions whose left edge would typically be coded as SBAR, or left edges of gerundive or subject-less infinitival clauses whose right edge would typically be coded as S or SBAR in the guidelines adopted here.

TABLE 4
The frequency of boundary marking at the *right edge* of each syntactic category

1. Category type	*2. Total number of right edges*	*3. number of edges with B-score ≥0.5*	*4. Percentage of B-marked edges in syntactic category (%)*	*5. Percentage of B-marked out of total number of B-marked edges (%)*	*6. Number of edges with B-score >0*	*7. Percentage of B-marked edges in syntactic category (%)*	*8. Percentage of B-marked out of total number of B-marked edges (%)*
VP	134	12	8.96	5.66	49	36.57	8.28
ADVP	49	7	14.29	3.30	14	28.57	2.36
ADJP	5	1	20.00	0.47	4	80.00	0.68
W/P	1016	21	2.07	9.91	125	12.30	**21.11**
PP	7	4	57.14	1.89	5	71.43	0.84
DM	64	26	40.63	**12.26**	57	89.06	9.63
NP	285	11	3.86	5.19	56	19.65	9.46
VP	134	12	8.96	5.66	49	36.57	8.28
DISF	78	37	47.44	**17.45**	76	97.44	**12.84**
S	159	81	50.94	**38.21**	157	98.74	**26.52**
CC-NP	16	2	12.50	0.84	6	37.50	0.92
CC-S	77	11	14.29	4.64	32	41.56	4.90
CC-VP	14	1	7.14	0.42	4	28.57	0.61
SBAR	22	11	50.00	4.64	19	86.36	2.91
Total	2060	237		100.00	653		100.00

Note: The data layout is the same as for Table 3.

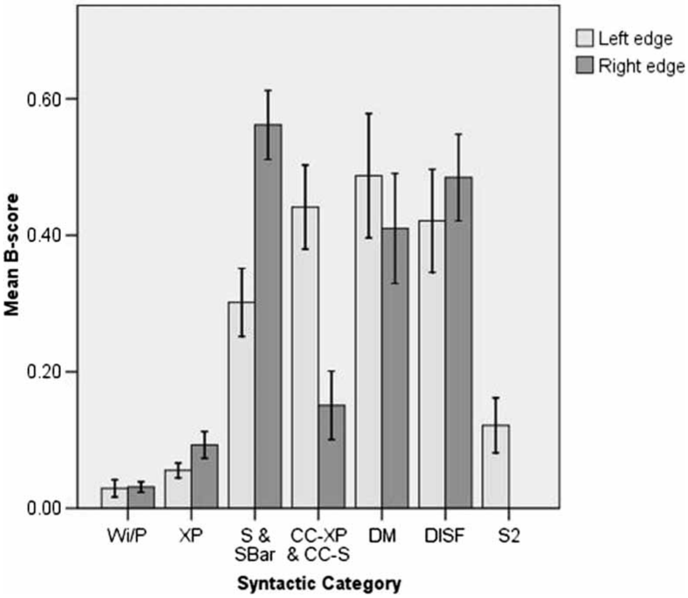

Figure 3. Mean B-scores and error bars (95% CI) by the syntactic category label of the highest syntactic constituent that begins (left edge) or ends (right edge) at the location of the perceived prosodic boundary. Data from all excerpts (Sets 1–4). The XP category (in the XP and CC-XP groups) combines the major phrase categories of NP, VP, ADJP, ADVP, and PP. The S2 category codes subordinate and relative clauses, and is only coded as a right-edge category (see text).

way, each word also contributes two syntactic category labels. Figure 3 plots these left- and right-edge B-scores separately.

These data reveal differences among categories in the overall level of B-scores, as well as asymmetries in the B-scores associated with left vs. right edges. The lowest B-scores are observed at the left and right edges of words that are phrase-medial (W/P) and at the edges of major phrases (XPs) that are not clauses, including NPs, VPs, ADJPs, PPs, and ADVPs. These locations may be marked by one or two transcribers, but rarely more, resulting in low B-scores. Notice that the W/P category was seen to contribute to more than 10% of the total number of marked boundaries in the entire set of transcripts (Tables 3 and 4), but the very low B-scores for phrase-medial words indicate that listeners rarely agree on the perception of a boundary at those locations. The low B-scores of XP and W/P categories contrast with the overall high B-scores of the other category labels, and notably with the higher-level syntactic categories marking clauses (S and SBAR). This difference was tested for significance using the non-parametric Kruskal–Wallis test of mean differences (because sample variances were not homogeneous between the groups), and is found to be highly significant, $\chi^2(2993, 1) = 100.6$, $p < .001$.[6]

A large-edge asymmetry is observed for left or right syntactic edges labelled as coordinating conjunctions (e.g., *but*, *and*) that conjoin sentences (CC-S) or other major categories (CC-XP). The difference in mean B-score between left and right edges for these categories is significant by the Kruskal–Wallis test, $\chi^2(1218) = 52.3$, $p < .001$. The high left-edge B-score indicates that listeners tend to hear prosodic boundaries preceding these conjunctions, rather than following them, and suggests that the conjunction is functioning as a proclitic in the prosodic phonology. There is also a large-edge asymmetry for the clausal categories (S and SBAR), but in the opposite direction from the asymmetry observed with coordinating conjunction categories. This difference is also significant under ANOVA [variances were homogeneous; $F(1, 372) = 52.2$, $p < .001$], and indicates that listeners tend to hear prosodic boundaries at the right edge of clause-level constituents.

A final observation from Figure 3 is that of relatively high B-scores associated with the left and right edges of DMs and DISFs. There are no significant differences between the means of left- and right-edge B-scores by ANOVA. The high B-scores and lack of edge effects indicate that listeners tend to hear DMs and DISFs as separate prosodic domains. We're hesitant to call these prosodic phrases, since it's not clear that they exhibit intonational patterns of a complete phrase, but at a minimum we can say that these elements are not perceived as integrated into the preceding or following prosodic phrases.

The results presented in this section provide indirect measures confirming hypothesis 1. There is a higher rate of boundary marking at the edges of clauses than at the edges of lower-level syntactic constituents or phrase-medial positions. Moreover, listeners agree more on the perception of prosodic boundaries at clause edges, especially right edges, than they do on boundaries in other locations, including the edges of non-clause constituents and medial positions in a syntactic domain. The clause edges have higher B-scores. These findings are predicted by hypothesis 1 if the acoustic evidence of prosodic phrase boundaries is also strongest at clause edges than other locations. We turn next to the analysis of the acoustic evidence.

RESULTS III: ACOUSTIC VOWEL DURATION BY SYNTACTIC CATEGORY

The distributions of B-scores according to syntactic category show differences in the patterns of prosodic boundary perception as a function

[6] All differences reported here as significant by non-parametric analyses of mean differences were also confirmed as significant under ANOVA.

of the syntactic category label of the word preceding and following the boundary. We consider next whether this effect of syntax on boundary perception is mediated through the acoustic signal. Hypothesis 2 of this study is that listeners respond to acoustic cues to prosodic boundaries in their judgement of the boundary status of each word. If the effect of syntactic category on prosodic boundary perception is mediated through the acoustic signal, we expect to find acoustic evidence of prosodic boundaries at those syntactic edges that are associated with higher B-scores. Prior studies show final lengthening is a very robust acoustic effect of prosodic phrase boundaries (Kim et al., 2006; Wightman et al., 1992; Yoon et al., 2007), with increased duration of the syllable-final vowel at increasingly higher levels of prosodic boundary (word < intermediate phrase (ip) < intonational phrase (IP)). Mo (2008) further shows that compared to overall vowel intensity, final vowel duration is a much more consistent correlate of a perceived prosodic phrase boundary across different vowel phonemes, based on the same dataset that is analysed here. Based on these prior findings, we focus on the duration of the final vowel of each word as an acoustic correlate of the perceived boundary following the word. We do not pursue acoustic correlates of the left edge of a perceived prosodic phrase in this paper, leaving that for future research.

Figure 4 displays the (normalised) duration of the word-final vowel and the B-score of each word. Words whose right edges are labelled as DM or DISF are excluded from this step of the analysis, as we have no clear prediction about final lengthening for words of this type. The plot shows a clear trend that words with higher B-scores have longer final vowels, as predicted by hypothesis 2.

Figure 5 shows the distribution of normalised final vowel duration values for five syntactic categories (using the right-edge syntactic label of each word). CC-XP (not shown) has a much skewed distribution and large variance, and is omitted from the analyses that follow. The variable duration patterns for CC-XP, which include conjunctions coordinating NPs, VPs, and other non-clausal phrases, may indicate highly variable patterns in prosodic phrasing, too. These conjunctions may sometimes group prosodically with the conjunct to the left and sometimes with the one to the right, in phrases like "*we did* ***or*** *didn't do* … ". Also excluded from the analysis of vowel duration are the categories DM and DISF, for which the linguistic model makes no strong predictions about vowel duration. Among the five categories shown in Figure 5, there is a trend in longer vowel durations at the end of clauses (S and SBAR), shorter durations following clause-joining conjunctions (CC-S), and the shortest durations at the end of other phrases (XP) and in phrase-medial positions (W/P). The differences in mean vowel duration between the five categories shown in Figure 5 are significant by the Kruskal–Wallis test, $\chi^2(4, 1062) = 118.49$, $p < .001$.

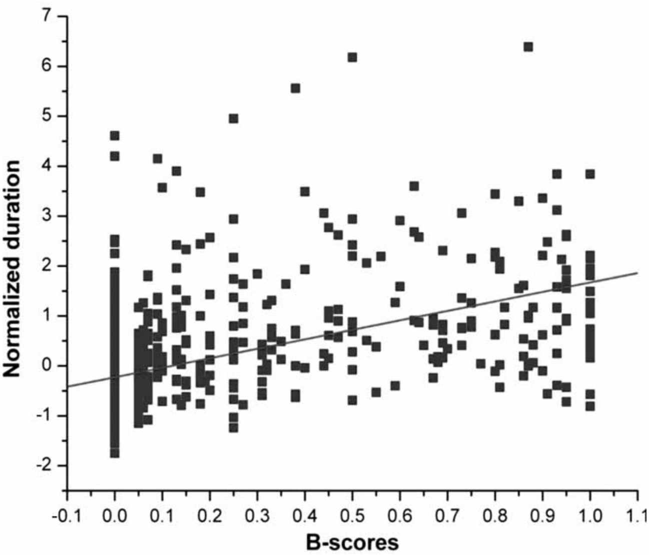

Figure 4. Scatterplot of B-scores and normalised duration of the final vowel of each word in the database, excluding words whose right edges are marked as discourse marker or disfluency ($N = 1422$).

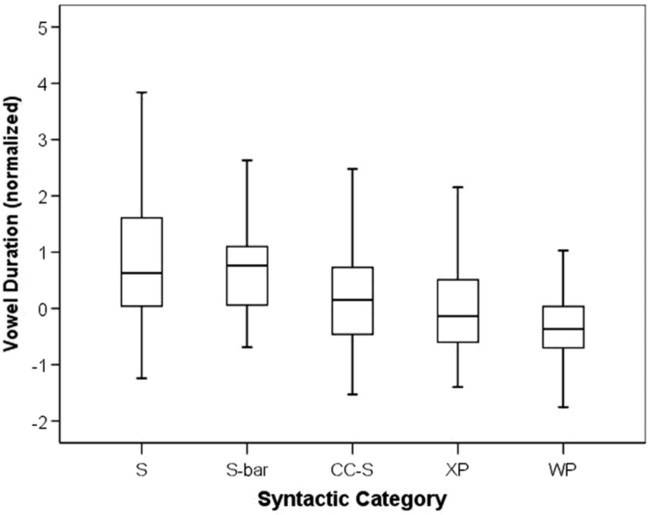

Figure 5. Normalised duration measures (z-transformed) of final vowel for each word grouped by the syntactic category label at its right edge. Excludes discourse markers and disfluency.

The vowel duration measure shows patterned variation that corresponds with B-scores, and with the right-edge syntactic category label of a word. These findings support hypothesis 2, but the question remains whether the syntactic effect on prosodic boundary perception can be fully attributed to the pattern of vowel duration at the right edge of syntactic constituents. To further explore the relationship between vowel duration, syntactic category and prosodic boundary perception, the next section presents correlation analyses of these variables.

RESULTS IV: CORRELATION AND REGRESSION ANALYSES OF BOUNDARY PERCEPTION, SYNTAX, AND ACOUSTIC DURATION

Non-parametric correlation analysis (Kendall's tau) and linear regression analyses are used to test the strength of the relationship between B-scores, right-edge syntactic category labels, and final vowel duration. Left-edge categories are not considered here, due to the expectation of weaker acoustic correlates to the prosodic boundary at the left edge. All regression models except for the final one presented (see Table 7, Model B) exclude words marked at their right edge with the "syntactic" labels marking DISF and DMs. Since there are no inherent numerical values for syntactic categories of different types, this categorical variable is recoded into a series of six binary dummy variables that code, for example, if a word-edge label is XP (yes = 1, no = 0) or if a word-edge label is S or SBAR.

There are significant correlations between each pairing of the three variables under discussion, as shown in Table 5. Separate correlations were tested for each syntactic category, except that the two clause-level categories

TABLE 5
Significant correlations (Kendall's tau) between B-scores, vowel duration, and syntactic category (right edge)

Variables		*Kendall's tau*
B-score	Vowel duration	.369**
B-score	S and SBAR	.541**
	XP	−.049*
	W/P	−.320**
Vowel duration	S and SBAR	.241**
	XP	.047 (p = .059)
	WP	−.208**

Note: Significant correlations marked by *p < .05, **p < .001.

S and SBAR were combined. There were no significant correlations involving the CC-S and CC-XP categories. Looking at the correlations between B-scores and syntactic features, we observe that clause endings (S and SBAR) and phrase-medial locations have a stronger correlation with B-scores than non-clause endings (XP), indicating that listeners are most consistent in hearing prosodic boundaries at the ends of clauses, and are pretty consistent in *not* hearing a boundary in phrase-medial locations, but are much more variable in their perception of boundaries at the ends of non-clausal phrases. The strongest correlation among these variables is between B-scores and the clausal syntactic categories of S and SBAR, and this correlation is even stronger than the correlation between B-scores and vowel duration, or between vowel duration and the S–SBAR syntactic variable. This finding indicates that the effect of syntactic category on B-scores can not be fully attributed to patterns of vowel duration encoding prosodic phrase boundaries at the end of clauses. This finding also supports hypothesis 3 that syntactic context plays a direct role on prosodic boundary perception. The independence of vowel duration and the syntactic categories on B-score prediction is further tested in a hierarchical regression model discussed next.

Table 6 shows results from hierarchical regression models with five syntactic categories combined as the independent variables in one step, and normalised word-final vowel duration as the independent variable in a second step. Model A first evaluates the syntactic category variables, and shows an r^2 value indicating that syntactic category accounts for nearly 30% of the variance in B-score values. Step 2 evaluates the predictive value of normalised duration, which accounts for about an additional 10% of the variance. Model B evaluates the predictor variables in the opposite order, and shows that when duration is evaluated first, it accounts for about 24% of the variance, with syntactic category variables picking up another 16% of the variance. These models show that the set of syntactic category variables and

TABLE 6

Results of two hierarchical linear regression models with syntactic categories (S, SBAR, CC-S, XP, and W/P) and normalised vowel duration as predictor variables and B-scores as dependent variable

		r	r^2	*Significant F* change
Model A	Step 1: syntactic categories	0.545	0.297	<0.001
	Step 2: normalised duration	0.635	0.403	<0.001
Model B	Step 1: normalised duration	0.488	0.239	<0.001
	Step 2: syntactic categories	0.635	0.403	<0.001

Note: Models A and B differ in the order in which the independent variables are evaluated.

vowel duration contribute significantly as predictors of B-score, together accounting for about 40% of the variance in B-scores, confirming our hypothesis that syntactic category contributes to boundary perception beyond what is predicted based on vowel duration patterns. The two models are very similar, with only a small difference in r^2 between them in the first step, where syntactic category variables are stronger predictors of B-score variation than normalised duration. This difference reflects the finding from the non-parametric correlation analysis (Table 5) that syntactic categories have a stronger correlation with B-scores than vowel duration does.

Adding all factors individually to a stepwise regression model indicates which variables contribute the most, and in which order, in predicting B-score variance. Table 7 presents the results from two more stepwise regression models. Model A uses the same dataset and the same predictor variables that went into the hierarchical regression model of Table 6, but does not force all the syntactic category labels to be evaluated in the same step. Here we see that the syntactic variable of clause categories (S and SBAR) emerge as the strongest predictors of B-scores in the first step, with vowel duration as the second step. The syntactic variables of phrase-medial (W/P) and other non-clausal phrase (XP) categories also contribute significantly in Steps 3 and 4. Model B presents the fullest regression model so far, adding back into the dataset the B-scores corresponding to words labelled as DMs and DISF (right edge only), and adding one new dummy variable to code these elements in a single, combined category. The resulting optimal model is like Model A in that clause and vowel duration are the first two strongest predictors, and also shows a strong relationship between the DM and DISF

TABLE 7
Stepwise regression models with full set of syntactic category variables and duration (Model), and with the added "syntactic" variables for discourse markers (DMs) and disfluencies (DISFs)

		r	r^2	*Significant F* change
Model A: excluding DM and DISF	Clause (S and SBAR)	0.596	0.355	<0.001
	Vowel duration	0.678	0.458	<0.001
	W/P	0.680	0.461	0.006
	XP	0.686	0.468	<0.001
Model B: including DM and DISF	Clause	0.537	0.288	<0.001
	Vowel duration	0.640	0.410	<0.001
	DM and DISF	0.686	0.471	<0.001
	W/P	0.688	0.473	0.021

Note: These are the optimal models from stepwise analysis where at each step the variable is selected that is the strongest significant predictor of B-score variance at that step. Predictor variables that are not significant for these models are not shown.

categories and B-scores by adding the variable for these items in the third step, before the variable for phrase-medial words. XP is no longer a significant predictor of B-score variance in Model B.

Summarising the findings from correlation and regression analysis, we find first of all that there are strong three-way correlations between syntactic category information, vowel duration, and perceived prosodic boundaries, confirming expectations based on prior works. Regression analyses show that syntactic category information makes an independent contribution to predicting prosodic boundary perception either before or after the contribution of vowel duration is partialed out. When all the syntactic variables are considered together with vowel duration, it is the clause-level categories and vowel duration that emerge as the two strongest predictors of B-scores, in that order. This result is maintained even when DMs and DISFs are added back into the dataset, with an additional variable.

DISCUSSION

Syntax is reflected in prosody production and perception

The distribution of B-scores at locations marking different kinds of syntactic edges establishes a systematic relationship between syntactic structure and prosody, as expected based on prior works showing a syntax–prosody dependency. The strongest relationship holds for the right edge of a clause-level constituent (S and S bar); listeners are more likely to hear a prosodic boundary in this location than at left edges, which generally supports Selkirk's (1986) analysis of English as a language where syntax and prosody are right-edge aligning. But for coordinating conjunctions linking XPs or clauses, listeners are more likely to hear a prosodic phrase boundary at the left edge of the conjunction than at the right edge, suggesting that conjunctions tend to behave prosodically like proclitics rather than enclitics. The same left-edge bias is observed for elements tagged as DMs.

The acoustic evidence from word-final vowel duration shows that boundary perception correlates with vowel duration, with higher B-scores associated with words that have longer final vowels. Vowel duration is also correlated with syntactic category in the expected direction: words at the end of higher syntactic domains such as clauses have longer duration than words at the end of lower syntactic domains, or words in a medial position of a syntactic domain. Collectively, these findings support hypothesis 1 that speakers' production of prosody reflects the syntactic structure of the utterance, and hypothesis 2 that listeners respond to acoustic cues to prosody in judging the location of prosodic phrase boundaries. The regression analyses also show that the contribution of syntax to boundary perception is a bit larger than the contribution of vowel duration, and that these two factors contribute independently, to some degree. The regression model that best predicts variation in B-scores at the right edges of words is the one that first evaluates the syntactic label at the right edge of the word, and then evaluates the vowel duration. This finding indicates that the relationship between vowel duration and B-scores is different depending on the syntactic category label of the word.

This study considers only one acoustic measure, word-final vowel duration, as the acoustic correlate of prosodic boundaries. As discussed earlier, the choice to use vowel duration as the sole cue is based on findings from prior studies that show duration to be the strongest cue, and the cue that is most consistent across phonological contexts. But we acknowledge here the possibility that other acoustic cues, such as F0, intensity, or glottalisation, may also contribute to listeners' perception of prosodic boundaries, and that regression analysis using a combined set of acoustic variables may show a larger predictive value for acoustic factors overall. In that case, the relative contributions of syntactic and acoustic factors would need to be re-evaluated as well. We are exploring this possibility in our ongoing work. The conclusion we draw from the present analyses is that relative to vowel duration, which is arguably the most robust acoustic cue to prosodic phrase boundaries in English, syntactic factors are even more strongly correlated with boundary perception, suggesting a partial independence between acoustic and syntactic cues to perceived prosody.

The best regression model utilising acoustic vowel duration and the full set of syntactic predictor variables accounts for 46.8% of the variation in B-scores excluding DMs and DISF, and 47.3% of the variation when they are included. Clearly, these factors play a major role in boundary perception, but equally, there must be other factors at play that are not yet included in the best model. We have neither accounted for factors related to focus or information structure, nor phonological factors related to the location of the nearest prominence or boundary to the left or right. We expect significant gains in the model once these factors are included.

B-scores in relation to discrete levels of prosodic boundary

The syntactic context that shows the highest rate of boundary marking, and also the highest rate of inter-listener agreement on boundary perception (i.e., the highest B-scores) is the right edge of a matrix sentence, relative clause, or subordinate clause. Lower-level syntactic categories are not reliably perceived as locations for prosodic phrase boundaries. The autosegmental–metrical model of English intonation (see Ladd, 1996/2008) specifies two levels of prosodic phrasing, the IP and the ip. The IP is higher in the prosodic hierarchy, and comprises one or more ips. These two levels are differentiated acoustically in that the higher-level prosodic boundary (IP) exhibits longer final lengthening, falling F0 contours ending in lower F0 values, lower intensity, and greater incidence of creaky voicing compared to the lower-level ip boundary (Kim et al., 2006). Differences between lower- and higher-level prosodic boundaries in terms of syntactic factors have not to our knowledge been explored, but we expect that the lower-level boundaries are more commonly associated with the edges of lower-level syntactic constituents, or are used in contexts of syntactic embedding.

We hypothesise that untrained listeners performing real-time prosody transcription are marking boundaries in locations that correspond to IPs in a careful ToBI transcription more consistently than in locations that would correspond to the lower-level ips, to the extent that these levels are effectively distinguished by speakers in our database. To test this hypothesis, we conducted a ToBI-style labelling (marking only locations of accent and boundary) with three trained, expert labellers (members of our research group) for a subset of the excerpts from the dataset reported here. The comparison between expert and untrained transcribers is reported in Mo et al. (2008), and the detail that is relevant here is that the words with the highest B-scores based on untrained transcribers correspond to words that are final in a higher-level prosodic phrase (IP) as transcribed by expert labellers. This finding, though based on a small sample from our data, supports our hypothesis that the untrained transcribers are marking higher-level boundaries.

There are several possible explanations for the less frequent perception of lower-level prosodic boundaries by untrained listeners. First, it is possible that the listeners do hear the lower-level boundaries, but simply don't have time to mark everything they hear, and opt to mark the (presumably) more salient higher-level boundaries. A second explanation is that speakers simply do not encode lower-level prosodic boundaries in the spontaneous, conversational speech style represented in our materials, unlike in read speech styles studied in prior works. A third possibility admits that speakers encode both lower- and higher-level boundaries, but are less consistent in the phonetic implementation of the lower-level boundaries, and/or implement these boundaries with subtler cues such that listeners do not reliably perceive them. A relevant observation is that both of the ip boundaries (L- and H-) are potentially ambiguous with the IP boundaries (L–L% and H–L%; among other ambiguities, see Beckman, 1996 for further examples). The lower-level boundaries are distinguished from the higher-level boundaries by the *degree* of F0 lowering, final lengthening or decreased intensity, and by the likelihood of creaky voicing. The inherent ambiguity of IP boundaries is lesser, since in at least some cases there are distinctive F0 contours that derive from the sequence of ip + IP tones, which do not arise with the single tones of the ip boundaries. The third explanation seems the most plausible to us, and is supported by the observation that even expert prosody transcribers struggle to annotate lower-level boundaries. Reliability studies of ToBI-style prosody transcriptions show lower agreement rates for ip compared to IP, and studies using statistical methods to predict boundaries similarly show lower rates of prediction for ip compared to IP (Yoon, 2007). These facts indicate that lower-level prosodic boundaries are not cued as effectively as higher-level boundaries, which may account for the rarity of boundaries perceived at locations marked by lower-level syntactic edges in our study.

The comparison between continuous-valued B-scores and discrete boundary labels such as (ip and IP) in the ToBI system is not intended as a claim that B-scores are an approximation of "true" discrete boundary labels. Our results are also compatible with a prosodic theory based on a continuous-valued boundary feature. Such a theory might posit a unique phonological boundary feature that is (usually) located to align with syntactic edges, but which is phonetically implemented with variation in the strength of the acoustic cues, such that "stronger" boundary cues are present at higher-level syntactic edges, or maybe as conditioned by non-syntactic factors. Prior work that assumes discrete levelled boundary features (e.g., IP and ip), such as our work on the Radio News and Switchboard corpora (Kim et al., 2006), finds significant differences in the acoustic correlates for two or more distinct levels, which suggests that the acoustic features are not drawn from a single uni-modal distribution, but it remains an open question whether there are two or more discrete categories of boundary level, or whether a better analysis is in terms of (one or more) continuous-valued boundary feature, where level distinctions such as IP/ip represent samples drawn from different regions along the continuum.

Sources of variability in prosody production and perception

The variability in the B-scores assigned to words in this study reflects the fact that among the 15–22 listeners transcribing each utterance, there is variability in their rate of inter-transcriber agreement in the perception of prosodic phrase boundaries. This variability in B-scores partly reflects variability among speakers in prosody production. As was shown in Figure 2, there is inter-speaker variation in the mean interval between transcribed boundaries, even with the same set of listeners for a given set of speech excerpts. This variability most likely reflects genuine differences between speakers in the mean length of their prosodic phrases (counted here in number of words), possibly reflecting differences in speech rate or affective factors. Inter-speaker differences in the interval between perceived prosodic phrases may also reflect differences in the salience of the acoustic cues to prosodic phrase boundaries; unless a given phrase boundary bears perceptually salient acoustic cues, it will not be detected and thus not counted in the B-score measure of perceived boundary strength.

Beyond these differences in speakers' production of prosodic phrase boundaries, the listener is another source of variability in observed B-scores, reflecting differences between transcribers in how they perform the task of transcription. Recall that the transcribers were given minimal instructions and no labelled training data or feedback that would have conditioned their transcription practice. It was expected that transcribers would differ in their sensitivity to the speech stimuli, and possibly also in their interpretation of the task and overall attentiveness when transcribing. These differences between transcribers are revealed in differences in the listener's rate of boundary marking. Over a subset of 36 transcribers, the mean interval between transcribed boundaries per listener ranges from 4.91 to 15.5 words (pooling all listeners' transcriptions the mean interval is 8.2 words, SD 2.17). But whereas the speaker-dependent variability in B-scores can be taken as evidence of real differences between speakers in prosody production, the listener-dependent variability in B-scores cannot so readily be taken as evidence of genuine differences between listeners in their perception of prosody. This study counts only those boundaries that listeners explicitly mark in a rapid and attention-demanding transcription task. It is possible that a listener's perceptual sensitivity might be different when gauged from a task of implicit boundary detection, for example, where the listener in some way responds to the perceived prosodic structure without explicitly identifying locations of prosodic boundaries.

We have not yet tested any implicit measures of boundary perception, so we do not offer any direct comparisons between our data and perceived prosodic boundaries based on implicit measures. But given the possibility of task-related factors influencing listeners, and resulting in inter-listener variation, the prosodic boundaries coding in our study cannot be interpreted on its own as direct evidence for the role of prosody in listeners' speech comprehension. We consider that those prosodic boundaries that are marked by listeners in rapid transcription may play a role in speech comprehension, but we do not claim that these are the only elements of prosodic structure that the listener perceives.

Boundary perception in relation to parsing and speech comprehension

Though this study did not test listeners' comprehension of syntax, the findings can be considered in light of models of syntactic parsing and speech comprehension. The coincidence of syntactic boundaries and perceived prosodic boundaries observed here, and their common expression in acoustic effects on vowel duration, supports models of spoken language processing in which prosody perception plays a role in speech comprehension for spontaneous speech, as has previously been shown for more controlled forms such as read speech. To the extent that speakers are consistent in producing longer syllable rhymes at the ends of major syntactic constituents like clauses, listeners may interpret syllable duration as a cue to syntactic structure. Furthermore, the co-occurrence of longer vowels with other acoustic prosodic features (not presented here), such as longer coda consonants, lower overall intensity, and spectral measures of creaky voicing or glottal tenseness (H1*–H2* and H2*–H4) serves to mark these regions in the speech stream as distinct from their surroundings, signalling a phonological and phonetic organisation that is aligned with the syntactic structure of the utterance, and by extension, with aspects of its semantic structure as well. The regular association of acoustic prosodic features with syntactic and semantic structure may condition listeners to perceive the phonological organisation of an utterance—its prosodic structure—where it is predicted by the syntactic context, even in contexts where the acoustic prosodic cues are weak or obscured. The perception of prosodic structure conditioned by comprehension of the syntactic context could explain the pattern of results obtained here, where syntactic factors appear to function at least partly independent of acoustic factors in influencing the perception of prosodic boundaries.

Prosody as a structural interface

Our findings support a model of language in which prosodic structure provides the means by which the phonological, syntactic, and semantic components of an utterance come together. As expressed by Arbisi-Kelm and Beckman (2009), prosodic structure can be considered as the "scaffolding" that marks certain locations in the speech stream as regions of convergence for critical and reliable information about the phonological, syntactic, and semantic content of an utterance. Viewed in this way, we would no sooner say that the perception of prosodic juncture cues syntactic structure than the converse that parsed syntactic structure cues phonological organisation. We further predict that semantic comprehension, e.g., the assignment of arguments identified from the speech stream to positions in lexical semantic structure, may function similarly as a cue to both phonological and syntactic organisation, and that prosody may likewise serve as cue to semantic structure. We leave it for future research to directly test these predictions by comparing the perception of prosody, as done here, with the listeners' comprehension of syntactic and semantic structures.

CONCLUSION

This study demonstrates that speakers encode syntactic structure in their production of prosody in spontaneous, conversational speech, and that the influence of syntax is strongest for clausal constituents at their right edge. In addition, ordinary listeners perceive prosodic boundaries in conversational speech in real-time, and their judgements also show a close relationship between perceived boundaries and syntactic structure. The relationship between syntax and prosodic boundary perception is mediated in part through the acoustic encoding of prosody. Acoustic cues to prosody are strongest at locations predicted by the syntactic structure. But the findings from regression analyses show that syntactic factors, specifically the context of a right-edge clause boundary, are a stronger predictor of boundary perception than the acoustic cue of word-final vowel duration, supporting the conclusion that syntactic factors make an independent contribution to boundary perception, and that relative to durational cues, the syntactic context is a stronger predictor or prosodic phrase boundaries. We conclude that listeners are guided in their perception of prosody by acoustic cues and syntactic context, and that the effect of syntactic context appears to be partly independent of the effect due to final vowel duration, the primary acoustic cue to prosodic phrase boundaries.

REFERENCES

Arbisi-Kelm, T., & Beckman, M. (2009). Prosodic structure and consonant development across languages. In M. Vigário, S. Frota, & M. J. Freitas (Eds.), *Phonetics & phonology: Interactions and interrelations* (pp. 109–136). Amsterdam: John Benjamins.

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics, 34*(4), 555–596. doi:10.1162/coli.07-034-R2

Bachenko, J., & Fitzpatrick, E. (1990). A computational grammar of discourse-neutral prosodic phrasing in English. *Computational Linguistics*, *16*, 155–170.

Beckman, M. (1986). *Stress and non-stress accent*. Dordrecht, the Netherlands: Foris.

Beckman, M., & Ayers, G. (1997). *Guidelines for ToBI labeling* (Version 3.0). Manuscript and accompanying speech materials. The Ohio State University. Retrieved September 10, 2008, from http://www.ling.ohio-state.edu/~tobi/ame_tobi/labelling_guide_v3.pdf

Beckman, M., & Pierrehumbert, J. (1986). Intonational structure in Japanese and English. *Phonology Yearbook*, *3*, 255–309. doi:10.1017/S095267570000066X

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 M. G. Rodriguez & C. P. S. Araujo (Eds.), *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC)* (pp. 779–785). Paris, France: Evaluations and Language Resources Distribution Agency.

Calhoun, S. (2006). *Information structure and the prosodic structure of English: A probabilistic relationship.* PhD dissertation, University of Edinburgh, UK.

Choi, J-Y., Hasegawa-Johnson, M., & Cole, J. (2005). Finding intonational boundaries using acoustic cues related to the voice source. *Journal of the Acoustical Society of America*, *118*(4), 2579–2588. doi:10.1121/1.2010288

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. doi:10.1177/001316446002000104

Cutler, A., Dahan, D., & van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech*, *40*(2), 141–201.

de Pijper, J. R., & Sanderman, A. A. (1994). On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. *Journal of the Acoustical Society of America*, *96*, 2037–2047. doi:10.1121/1.410145

Dilley, L., Breen, M., Bolivar, M., Kraemer, J., & Gibson, E. (2006). A comparison of inter-transcriber reliability for two systems of prosodic annotation: RaR (rhythm and pitch) and ToBI (tones and break indices). In *Proceedings of the International Conference on spoken language processing* (pp. 1619–1622). International Speech Communication Association. Available from http://www.isca-speech.org/archive/interspeech_2006

Dilley, L., Shattuck-Hufnagel, S., & Ostendorf, M. (1996). Glottalization of word-initial vowels as a function of prosodic context. *Journal of Phonetics*, *24*, 423–444. doi:10.1006/jpho.1996.0023

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (pp. 38–46). New York: John Wiley.

Frazier, L., Carlson, K., & Clifton, C., Jr. (2006). Prosodic phrasing is central to language comprehension. *Trends in Cognitive Sciences*, *10*, 244–249.

Frazier, L., Clifton, C., Jr., & Carlson, K. (2004). Don't break or do: Prosodic boundary preferences. *Lingua*, *114*, 3–27. doi:10.1016/S0024-3841(03)00044-5

Gee, J. P., & Grosjean, F. (1983). Performance structures: A psycholinguistic and linguistic appraisal. *Cognitive Psychology*, *15*, 411–458. doi:10.1016/0010-0285(83)90014-2

Heldner, M. (2003). On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in Swedish. *Journal of Phonetics*, *31*(1), 39–62. doi:10.1016/S0095-4470(02)00071-2

Kim, H., Yoon, T-J., Cole, J., & Hasegawa-Johnson, M. (2006). Acoustic differentiation of L- and L-L% in switchboard and radio news speech. In R. Hoffmann & H. Mixdorff (Eds.), *Proceedings of the Third International Conference on Speech Prosody 2006*, Dresden, Germany, May 2–5, 2006. Available from ISCA Archive: http://www.isca-speech.org/archive/sp2006.

Kochanski, G., Grabe, E., Coleman, J., & Rosner, B. (2005). Loudness predicts prominence: Fundamental frequency lends little. *Journal of the Acoustical Society of America*, *118*, 1038–1054. doi:10.1121/1.1923349

Ladd, D. R. (2008). *Intonational phonology* (2nd ed.). Cambridge: Cambridge University Press. (Original work published 1996)

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174. Available from http://www.jstor.org/stable/2529310.

Lee, E-K., & Cole, J. (2007). *Acoustic effects of prosodic boundary on vowels in American English*. Paper presented at the Proceedings of the 42nd meeting of the Chicago Linguistic Society.

Lehiste, I. (1972). The timing of utterances and linguistic boundaries. *Journal of the Acoustical Society of America*, *51*, 2018–2024. doi:10.1121/1.1913062

Marcus, M., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, *19*(2), 313–330.

Mo, Y. (2008). Duration and intensity as perceptual cues for naïve listeners' prominence and boundary perception. In P. A. Barbosa, S. Madureira, & C. Reis (Eds.), *Proceedings of the Fourth International Conference on Speech Prosody* (pp. 739–742). Campinas, Brazil, May 6–9, 2008. Available from ISCA Archive: http://www.isca-speech.org/archive/sp2008.

Mo, Y., Cole, J., & Lee, E-K. (2008). Naïve listeners' prominence and boundary perception. In P. A. Barbosa, S. Madureira, & C. Reis (Eds.), *Proceedings of the Fourth International Conference on Speech Prosody* (pp. 735–736). Campinas, Brazil, May 6–9, 2008. Available from ISCA Archive: http://www.isca-speech.org/archive/sp2008.

Nespor, M., & Vogel, I. (1986). *Prosodic phonology.* Dordrecht, the Netherlands: Foris.

Ostendorf, M., Price P., & Shattuck-Hufnagel, S. (1995). *The Boston University radio news corpus* (Technical Report ECS-95-001). Boston, MA: Boston University. Retrieved 12/2/2009 from The Linguistic Data Consortium, Philadelphia, PA, http://www.ldc.upenn.edu/Catalog/docs/LDC96S36/bur_crps.ps.

Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., et al. (2007). *Buckeye corpus of conversational speech* (2nd release). Columbus, OH: Department of Psychology, Ohio State University. Retrieved March 15, 2006, from www.buckeyecorpus.osu.edu

Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S., & Fong, C. (1991). The use of prosody in syntactic disambiguation. *Journal of the Acoustical Society of America*, *90*, 2956–2970. doi:10.1121/1.401770

Pynte, J. (2006). Phrasing effects in comprehending PP constructions. *Journal of Psycholinguist Research*, *35*, 245–265. doi:10.1007/s10936-006-9014-y

Schafer, A., Speer, S., Warren, P., & White, S. D. (2000). Intonational disambiguation in sentence production and comprehension. *Journal of Psycholinguistic Research*, *29*(2), 169–182. doi:10.1023/A:1005192911512

Selkirk, E. O. (1984). *The relation between sound and structure*. Cambridge: MIT Press.

Selkirk, E. O. (1986). On derived domains in sentence phonology. *Phonology Yearbook*, *3*, 371–405. doi:10.1017/S0952675700000695

Selkirk, E. O. (2000). The interaction of constraints on prosodic phrasing. In M. Horne (Ed.), *Prosody: Theory and experiment* (pp. 231–261). Dordrecht, the Netherlands: Kluwer.

Shattuck-Hufnagel, S., & Turk, A. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research*, *25*, 193–247. doi:10.1007/BF01708572

Sluijter, A., & van Heuven, V. J. (1996). Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America*, *100*, 2471–2485. doi:10.1121/1.417955

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. In R. J. J. H. van Son (Ed.), *Proceedings of the Institute of Phonetic Sciences* (Vol. 21, pp. 101–116). Amsterdam, The Netherlands: University of Amsterdam.

Streefkerk, B. M., Pols, L. C. W., & ten Bosch, L. F. M. (1998). Automatic detection of prominence (as defined by listeners' judgements) in read aloud Dutch sentences. In *Proceedings of the 5th International Conference on Spoken Language Processing* (Vol. 3, pp. 683–686). Sydney, Australia, 30th November–4th December 1998. Available from ISCA Archive, http://www.isca-speech.org/archive/icslp_1998.

Swerts, M. (1997). Prosodic features at discourse boundaries of different strength. *Journal of the Acoustical Society of America*, *101*, 514–521. doi:10.1121/1.418114

Turk, A., & Sawusch, J. (1997). The domain of accentual lengthening in American English. *Journal of Phonetics*, *25*, 25–41. doi: 10.1006/jpho.1996.0032

van Bergem, D. R. (1993). Acoustic vowel reduction as a function of sentence accent, word stress and vowel class. *Speech Communication*, *12*, 1–23.

Watson, D., & Gibson, E. (2004). The relationship between intonational phrasing and syntactic structure in language production. *Language and Cognitive Processes*, *19*(6), 713–755. doi:10.1080/01690960444000070

Weber, A., Grice, M., & Crocker, M. (2006). The role of prosody in the interpretation of structural ambiguities: A study of anticipatory eye movements. *Cognition*, *99*, B63–B72. doi:10.1016/j.cognition.2005.07.001

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America*, *91*, 1707–1717. doi:10.1121/1.402450

Yoon, T-J. (2007). *A predictive model of prosody through grammatical interface: A computational approach.* PhD dissertation, University of Illinois at Urbana-Champaign.

Yoon, T-J., Chavarría, S., Cole, J., & Hasegawa-Johnson, M. (2004). Intertranscriber reliability of prosodic labeling on telephone conversation using ToBI. In *Proceedings of the ISCA International Conference on spoken language processing (Interspeech 2004)* (pp. 2729–2732). Jeju Island, Korea, October 4–8, 2004. Available from ISCA Archive, http://www.isca-speech.org/archive/interspeech_2004.

Yoon, T-J., Cole, J., & Hasegawa-Johnson, M. (2007). *On the edge: Acoustic cues to layered prosodic domains.* Paper presented at the Proceedings of the International Congress of Phonetic Sciences, Saarbruecken, Germany.

APPENDIX 1

Scripted instructions read to participants in transcription experiment:

<the first part of the script pertains to prosodic prominence transcription>

. . . [A] feature of normal speech that we are interested in is the way speakers break up an utterance into **chunks.** These chunks group words in a way that helps the listener interpret the utterance, and are especially important when the speaker produces long stretches of continuous speech. An example of chunking that is familiar to everyone is the chunking that breaks digit sequences down into sub-groups.

For some of the excerpts you will hear, you will be asked to mark the chunks by inserting a vertical line between words that belong to different chunks. It is important for you to know that the boundary between two chunks does not necessarily correspond to the location where you would place a comma, period, or other punctuation mark, so you must really listen and mark the boundary where you here a juncture between two chunks. A chunk may be as small as a single word, or it may contain many words, and speakers can vary quite a bit in the size of the chunks they produce in a given utterance.