Perception of Prosodic Boundaries by Naïve and Expert Listeners in French Modelling and Automatic Annotation

George Christodoulides¹, Anne Catherine Simon¹, Ivana Didirková¹

¹Centre Valibé, IL&C, Université catholique de Louvain, Belgium

george@mycontent.gr, anne-catherine.simon@uclouvain.be, ivana.didirkova@uclouvain.be

Abstract

We present the results of a series of experiments in which naïve listeners and expert annotators were tasked to indicate the presence of a prosodic boundary in real-time, by tapping on a computer keyboard. These taps are attributed to potential boundary syllables using an algorithm. Both groups listened to 48 samples from a corpus of French speech in multiple speaking styles, which also contains an annotation of prosodic boundaries produced off-line by experts. Several information sources (acoustic features, lexical and shallow syntactic features) are correlated with the responses of naïve listeners and expert annotators. We compare the results for the two groups and describe an automated system for annotating prosodic boundaries.

Index Terms: automatic prosodic labelling, speech perception modelling, prosodic segmentation, disfluencies, prominence

1. Introduction

Prosodic information plays a central role in the comprehension of spoken language. It facilitates the recognition of spoken words, aids in the reconstruction of syntactic structure by resolving local and global ambiguities, and plays a role in discourse comprehension: in the processing of discourse structure, by indicating the information status (e.g. focus and accentuation) and discourse relations between elements, and by contributing to the integration of concepts with an existing discourse model (for a review, see [1]).

Syntax does not fully predict the way that spoken utterances are organized [2], and there is evidence that prosodic phrasing can facilitate or disrupt the processing of an utterance, especially in the case of structural ambiguities [3]. It is therefore interesting to study the placement of prosodic boundaries in relation to the syntactical structure of an utterance, in relation to the information status of elements before and after the boundary, and in relation to the discourse relations between utterances. According to the “informative prosodic boundary hypothesis” [4], listeners interpret prosodic boundaries by comparing and processing them relative to other prosodic boundaries in the same utterance, and treat a prosodic boundary as more informative when less expected (e.g. when separating a short rather than a long constituent) [5].

While most phonological models of French (e.g. [6], [7]) admit at least three degrees of prosodic boundaries, large-scale corpora are usually limited to one or two degrees of prosodic boundaries. Similarly to the perception of prosodic prominence, there is evidence that listeners perceive prosodic boundaries as a gradual phenomenon and in relative terms.

Furthermore, the prosodic annotation of spoken language corpora is typically performed by annotators who have the possibility to listen to speech segments multiple times, have access to additional information (such as a visual representation of the pitch track), and can revise their annotation decisions. However, speech perception in ecological situations is different, in that the incoming speech is only accessed once and in that comprehension happens in real time.

In this contribution we compare the results of a series of experiments in the real-time annotation of prosodic boundaries in French, by which we have tried to address the aforementioned concerns. First, the experimental design allowed us to attribute a continuous score to perceived prosodic boundaries. Second, these scores are correlated with a series of acoustic and prosodic features of the speech signal, as well as with POS and syntactic information (in order to capture both bottom-up interpretations and top-down expectations). Thirdly, the annotation was performed in real time. Finally, the same speech samples were submitted for annotation using the same experimental protocol to a group of naïve listeners and to a group of experts in linguistics and prosody; we compare the annotations produced by subject group and condition; such annotations are used to train automatic annotation models.

2. Related Work

Many spoken language corpora include some sort of prosodic segmentation into “intonation units” or annotation of “prosodic boundaries”. Such annotations include the “tone units” in the London-Lund Corpus of Spoken English [8]; “intonation units” in the Santa Barbara Corpus of American English [9]; “accentuation units” and “intonation units” in the Aix-Marsec Corpus of Spoken British English [10]; “accentuation units” and “intonation units” based on the Autosegmental-Metrical framework in the Rhapsodie Corpus [11] and in the C-PROM-PFC Corpus [12]; and two levels of prosodic boundaries marked in the LOCAS-F Corpus [13]. All of these annotations are presumably off-line expert annotations.

Previous research on the acoustic correlates of prosodic boundaries has focused on comparing the features of units bearing a prosodic boundary (as perceived and annotated by experts) to the features of units not bearing a perceived prosodic boundary. There are methodological differences and language-specific conclusions, but some general trends emerge. The main acoustic correlates for the perception of a prosodic boundary have been shown to include silent pauses following the boundary syllable (e.g. [14], [15]) and syllable duration or lengthening (e.g. [16]). In French, where prosodic boundaries and syllabic prominence are strongly related, unlike lexical-stress languages such as English or Dutch, [17] observes that weak boundaries are associated with rising contours and strong boundaries with falling contours. Experiments have also shown that top-down linguistic expectations influence the perception of prosodic boundaries and prominence [18], while hesitation-related disfluencies also affect the perception of prosodic events ([19], [20]).

There are differences between the conditions under which
experts annotate prosodic phenomena on a speech recording and the real-time perception of these phenomena by naïve listeners. In real-time perception experiments, non-expert listeners tend to identify fewer prosodic boundaries than those annotated by experts: this is not only due to the task constraints, but also to the top-down influence of syntax and semantics (e.g. [21], [14], [20], [22]).

Finally, previous work on the automatic annotation of prosodic events has mainly focused on using supervised machine learning models trained on hand-labelled corpora. Working on English speech, [23] used decision trees and HMMs to detect and classify prosodic event sequences. [24] proposed an annotator for pitch accents (presence or absence) and prosodic boundaries, based on ANN and GMM models trained on the Boston Radio Speech corpus; they report a 85% and 93% accuracy respectively, when using both acoustic and syntactical features. [25] show that the accuracy of the automatic labelling for pitch accents and boundary tones is improved by jointly modelling simple lexical, syntactic and acoustic features, using a maximum entropy model (accuracy for accent and boundary tone detection: 86.0% and 93.1% on the Boston University Radio News corpus, and, 79.8% and 90.3% on the Boston Directions corpus, respectively). [26] developed an automatic system that associates ToBI labels to syllables, initially based on the expert annotation of the Boston Directions Corpus, while the AuToToBI system currently includes several models. Given that producing large training corpora is costly, [27] proposed a bootstrap process: using a large unlabelled corpus to perform unsupervised adaptation of an acoustic-prosodic model trained on a small labelled corpus.

3. Method

Potential samples were extracted from the LOCAS-F corpus [13], a multi-genre corpus of spoken French. The selected samples were monological units, 20-60 seconds long and between two pauses of at least 250 ms. They were clustered based on 4 criteria (articulation rate, silent pause ratio, melodicity, filled pauses to number of syllables ratio). We performed a stratified selection of stimuli in two groups (fluent vs. disfluent); 4 groups of 12 stimuli were selected, with an average duration of 29.9 seconds (min: 5.1, max 39.9). For each sample, a corresponding delexicalised version was obtained using speech synthesis and replacing each phoneme with another from the same class (and shaping the duration and pitch of the resynthesised stimulus to match the original). Four randomised sets of stimuli were created (12 natural, 12 manipulated speech) and each sample was presented only once to each participant of a perceptual experiment.

Participants were instructed to press the space-bar as soon as they heard the end of a “group of words” (the concept of prosodic boundaries was included in the instructions for experts). Naïve annotators could only listen to each sample once; experts had the option to repeat the annotation. The collection of responses was done in real time, in order to be as close as possible to natural conditions of speech perception and comprehension. Baseline reaction times to simple tones were measured before and after the perception task. The perception task was preceded by a working memory test, and a test of tonal acuity. The entire experimental sequence was programmed both using OpenSesame [28] and as stand-alone program running on Windows and Mac (using C++ and the Qt framework). The experiment software, called Tapping Annotator, records all keystrokes (press and release times) for further analysis. The software along with a Praaline [29] plug-in to perform the analysis of the data collected is available at https://github.com/praaline/tapping-annotator.

3.1. Data Analysis

For each subject, we calculated a mean RT from their responses to the pure tones. These values were subtracted from their responses in order to centre them with respect to a potential location of a prosodic boundary and to reduce variability induced by individual motor skill differences. A moving average (window size 250 ms) of the number of responses was calculated and the local maxima of this value were considered as the perceived prosodic boundary (PPB) sites. In order to group subject responses correlated with a PPB, we applied the following algorithm: starting from the centre and within a window of 500 ms on either side, a response is attributed to the PPB if its distance from the previous response is less than 300 ms; we are thus attempting to detect clusters of responses triggered by the same cues. These responses are subsequently treated as a group: each group gives rise to a PPB. These PPBs were correlated with the nearest final syllable of a token (PPB sites falling within a silent pause were attributed to the previous final syllable). For each PPB we calculate three measures (boundary force, delay and dispersion). The boundary force is the percentage of participants who registered a response at this PPB site. This procedure has been presented in [22].

3.2. Participants

Two sets of experiments were conducted: one set with naïve listeners, i.e. native speakers of French without specialised knowledge or training in phonology or corpus annotation, and one set with expert listeners, i.e. researchers having worked extensively in the field of French phonology, prosody and spoken corpus linguistics.

In total, 133 university students took part in the perceptual experiment for naïve annotators (86 in the first session and 47 in the second session). They were all studying at the faculties of Psychology and Modern Languages at the University of Louvain in Belgium. Eight participants were excluded from the final analysis. We sent out 83 personal invitations to colleagues to participate in the perceptual experiment for expert annotators; 30 experts completed the experiment and sent in their annotations (one participant had to be excluded from the final analysis due to a technical issue). The results are based on the analysis of the responses of 125 naïve annotators (115 female, 10 male) and 29 expert annotators (20 female, 9 male). The median age of naïve annotators was 23 years old; the median age of expert annotators was 45 years old (min: 27, max: 73).

4. Results and Discussion

4.1. Comparison between Naïve and Expert Annotators

Figure 1 shows the histograms of the Boundary Force variable for each subject group (naïve, expert) and condition (natural speech, synthesised speech). There are 3307 syllables that may potentially bear a boundary. In the case of naïve annotators, 718 syllables were perceived as bearing a prosodic boundary: 658 were perceived as boundaries in both conditions; 147 were perceived as boundaries only in natural speech and 102 only in manipulated speech. In the case of expert annotators, 1212 syllables were perceived as bearing a prosodic boundary: 658 were perceived as boundaries in both conditions; 356 were perceived...
as boundaries only in natural speech and 198 only in manipulated speech. Experts annotated more prosodic boundaries than naïve annotators.

For each stimulus, we calculated Fleiss’ kappa as a measure of inter-annotator agreement: each syllable at the end of a multi-word unit was considered a potential boundary site, and each participant that listened to the stimulus was considered an annotator who chose or did not choose to indicate a prosodic boundary on that syllable. Figure 2 shows the distribution of the Fleiss’ kappa values for each subject group and condition. The inter-annotator agreement of experts was significantly higher than that of naïve annotators, across conditions. The inter-annotator agreement for synthesised speech was significantly higher than that for natural speech, across groups.

4.2. Comparison with the Off-line Annotation

The original corpus also contains an off-line annotation of perceived prosodic boundaries by two experts. Each word was marked as being followed by a strong PB (/\), an intermediate PB (\), as not followed by any boundary (0), or as a hesitation (hesi). A function was also attributed to each PB, based on the shape of the corresponding intonation contour: C (continuation), T (final prosody), S (suspen) and F (focus). This annotation was primarily based on the annotators’ perception; however they did have visual access to the pitch contour as displayed in Praat [30]. In cases of disagreement, the annotators listened to the relevant section once again and agreed on the final PB and contour label.

Figure 3 shows the relation between this off-line annota-

4.3. Predictor Relative Importance

In order to evaluate the relative importance of each syllable feature in determining whether it will be perceived as a prosodic boundary, we used linear regression predicting the boundary force value on the basis of features. The predictors were the duration of the subsequent silent pause (if any), the relative duration of the syllable compared to the previous 2 syllables, the relative mean pitch compared to the previous 2 syllables, and the type of syntactic boundary at the end of the syllable (e.g. end of multi-word unit, end of clause, discourse marker etc. [31]). The resulting regression trees, for each subject group and condition, are shown in Figures 5, 6, 7 and 8.

We observe that the most important acoustic correlate for the perception of a prosodic boundary (across subject groups

![Figure 1: Histogram of perceived boundary force per subject group and condition.](image1)

![Figure 2: Fleiss’ kappa per subject group and condition.](image2)

![Figure 3: Box-plot of the perceived boundary force, compared to the off-line annotation, per subject group and condition.](image3)

![Figure 4: Scatter plot of the perceived boundary force per condition. Each point represents a PB, the x-axis shows the force attributed by the naïve listeners and the y-axis by the expert listeners. Colour coding indicates the corresponding off-line annotation.](image4)
and conditions) is the silent pause. This is followed by the co-occurring syntactical boundary in the case of natural speech (where listeners have access to this information; this is not the case in the synthesised speech condition). The next predictor is the relative syllable duration, indicating final lengthening of boundary syllables. Predictors related to pitch were deemed less important by the linear regression model (they do not appear in the 3-level trees). Of these, pitch trajectory (inter-syllabic and intra-syllabic pitch movement) was included in the model (with p < 0.001).

4.4. Automatic Annotation

We have also developed an automatic annotation system for prosodic boundaries in French, using the Promise annotator [32], with 3 different models trained on the off-line annotation, the naive annotation, and the expert annotation. The automatic annotation problem is approached as a sequence labelling problem, and modelled using Conditional Random Fields, as described in [33]. With respect to the prosodic boundary strength annotation, the model achieved a 92% precision in identifying strong prosodic boundaries and a 65% precision in identifying intermediate prosodic boundaries (based on the results of a 5-fold cross-validation, on the LOCAS-F corpus). The evaluation of the automatic annotation of boundary strength is shown in Table 1.

5. Conclusion

We presented the results of a series of experiments on the perception of prosodic boundaries in French. Using a real-time annotation experiment, we collected responses from naive listeners and experts, annotating natural speech and delexicalised speech samples. The results show that the annotation by experts on-line was close to the annotation by experts off-line; that expert annotators indicated more prosodic boundaries than naive annotators, but that there was a high level of agreement between the two groups. Linear regression models indicate that the most important acoustic correlate for the perception of a prosodic boundary is the subsequent silent pause, followed by the co-occurrence of a major syntactical boundary and by final lengthening. Finally, we used the data collected to develop an automatic annotation system released under an open source license (www.praaline.org/promise).

6. Acknowledgements

This research is supported by the Belgian F.R.S./FNRS (MIS Project PhraDiCo: Prosodic phrasing for discourse comprehension. Perception of prosodic boundaries and their contribution to the construal of discourse coherence, convention No. F 4520.16, PI: Anne Catherine Simon).
7. References


