Non-traditional prosodic features for automated phrase break prediction

Claire Brierley and Eric Atwell
University of Leeds, UK

Abstract

It is universally recognized that humans process speech and language in chunks, each meaningful in itself. Any two renditions or assimilations of a given sentence will exhibit similarities and discrepancies in the distribution of phrase breaks. Automated phrase break prediction assigns pauses to plain text as input, evaluated against human performance encapsulated in 'gold standard' boundary annotations in a speech corpus. This article advocates an enhanced feature set for phrase break prediction, incorporating non-traditional prosodic features. The authors have developed ProPOSEL, a **pro**sody and **p**art-**o**f-**s**peech **E**nglish **l**exicon, as text annotation and text analytics tool. Application of ProPOSEL has so far uncovered a statistically significant correlation in English between certain sound patterns (i.e. the diphthongs and triphthongs of Received Pronunciation) and phrase breaks in very different genres. Thus, presence or absence of a complex vowel could easily be incorporated as an extra non-traditional classificatory feature in phrase break models. Our approach also suggests new possibilities for statistical analysis of texts, particularly authorship and genre, via favoured sound and rhythmic patterns in addition to lexis. Moreover, we suggest that our approach of text data-mining descriptive annotations of projected prosody for Text-to-Speech Synthesis and stylistic analysis is applicable to other languages.

Correspondence:
Claire Brierley,
School of Computing,
University of Leeds,
Woodhouse Lane,
Leeds LS2 9JTT, UK
E-mail:
scscb@leeds.ac.uk

1 Introduction

It is universally recognized that humans process speech and language in chunks, each meaningful in itself. This research explores the variety of linguistic cues native speakers use to signal or tune into chunk boundaries via discussion and experimentation. In written English, prominent boundaries are marked by punctuation, which the authors interpret as a form of prosodic-syntactic annotation, denoting chunks of meaning intended by the writer and picked up by the reader. Conversely, in a corpus of transcribed speech, boundary annotations represent *perceived* pauses in the speech stream.

Punctuation may not reflect all such chunks. In this extract from *Jane Eyre*, a succession of elongated vowel sounds—the diphthongs *rain*, *away*, and *wildly*—plus an opportunistic syntactic boundary between two prepositional phrases (in bold) may subconsciously influence readers to pause after *wildly*, as in this example:

> . . . At intervals, while turning over the leaves in my book, I studied the aspect of that winter afternoon. Afar, it offered a pale blank of mist and cloud; near, a scene of wet lawn and storm-beat shrub, **with ceaseless rain sweeping away wildly before a long and lamentable blast** . . .

Literary and Linguistic Computing, Vol. 26, No. 3, 2011.

doi:10.1093/llc/fqr023 Advance Access published on 13 May 2011

(From Chapter 1 of *Jane Eyre* by Charlotte Brontë, 1847)

The reverse might happen in the following breathlessly chunked fragment from a transcript of 1980s BBC radio newsreel in SEC, the Spoken English Corpus (Taylor and Knowles, 1988).

. . . I jumped in | and we set off | at the manic speed | **which for some reason | is a characteristic | of the way all journalists drive** | here in El Salvador | . . .
(From Section A of the *Spoken English Corpus*)

By *elision* across a boundary—*reason's* versus **reason | is**—and *coarticulation* across a boundary—*characteristicof* versus **characteristic | of**—we can craft lengthier, more 'chilled out' phrasing in the bolded section.

2 Phrase Break Prediction and the Challenge of Modelling Human Psycholinguistic Chunking

The goal of automated phrase break prediction is to emulate human performance in terms of naturalness and intelligibility when assigning prosodic-syntactic boundaries to input text. Techniques can be deterministic or probabilistic. Phrase break models exemplifying these two generic approaches are the *chinks 'n' chunks* algorithm (Liberman and Church, 1992) and Taylor and Black's Markov model (1998). The former inserts a boundary after punctuation and whenever the input string matches the sequence: open-class or content word (chunk) immediately followed by closed-class or function word (chink), based on the principle that chinks initiate new prosodic phrases. The latter conditions the probability of juncture type (break or non-break) at any given point on: (i) the *prior probability* of a break or non-break given the immediate syntactic context (i.e. the part-of-speech or PoS trigram in which that juncture is embedded); and (ii) the *likelihood* of a break or non-break at that point given the previous sequence of N juncture types—where, in the best performing model, $N = 6$.

However, as we have just illustrated, there are usually alternative parsing and phrasing strategies for a given sentence and therefore evaluation against one prosodic-syntactic variant is problematic (Taylor and Black, 1998; Atterer and Klein, 2002; Brierley and Atwell, 2008a). Moreover, as illustrated in the Section 1, other linguistic features may influence boundary placement in addition to the syntactic and text-based cues traditionally used. These 'non-traditional' features emerge from our internalized experience (and hence knowledge) of the *physicality* of language: the sound system and rhythm of our native tongue; it is these features which engage our interest here.

Developing a better model of human psycholinguistic chunking for applications like text-to-speech synthesis (Sanderman, 1994) may well, in our view, entail reinstating *a priori* linguistic knowledge of *prosody* to complement traditional features (i.e. syntax and punctuation). This echoes current thinking in the wider research community: extending knowledge sources to supplement raw training data, and hence improve performance, is seen as a challenge for automatic speech recognition[1] and machine learning in general (PASCAL, 2008).

3 ProPOSEL: A Prosody and PoS Text Annotation Tool

To this end, the authors have developed ProPOSEL (Brierley, forthcoming), a *pro*sody and *PoS E*nglish *l*exicon of 104,049 entries, for text annotation and text analytics, supported by a comprehensive software tutorial. The latter expounds the computational steps necessary to annotate text with descriptive categories for syntactic, phonetic, and prosodic attributes. Table 1 gives an example of the finished article, a fragment of newsreel tagged via ProPOSEL, with word tokens mapped to symbolic representations as follows: two variant PoS-tagging schemes (LOB[2] and C5[3]); DISC[4] phonetic transcriptions; default open and closed-class word categories; syllable counts; lexical stress patterns (i.e. abstract representations of word-internal rhythm, as in 201 for *disappear*, with secondary stress on the first syllable and primary stress on

the final syllable); and finally an additional boundary classification derived from corpus annotations in SEC.

4 ProPOSEL and Text Analytics

In the Section 1, we cite a relatively long (12 words) unpunctuated fragment from *Jane Eyre*, where the writer can, in a sense, depend on readers to partition it naturally because their innate sense of prosody compels them to. In recital, we would probably need to pause between *wildly* and *before* in this fragment because of the effort required to produce the succession of long vowels and diphthongs in: '*. . . ceaseless rain sweeping away wildly . . .*'. It is our contention that: (i) even in silent reading, people are sensitive to the prosody inherent in text (cf. Donald, 1993 p. 248; Fodor, 2002); (ii) certain sound patterns are tantamount to *linguistic signs* for phrase breaks; (iii) such patterns can be extracted from the lexicon and reconceptualized as categorical features for phrase break classification in the same way that real-world knowledge of syntax is represented in PoS tags; and (iv) such features are potentially very useful because they are domain independent and can be projected onto *any* corpus.

In a recent paper for *Literary and Linguistic Computing* (Brierley and Atwell, 2010), we use significance testing to evaluate the association between one such sound pattern, the subset of complex vowels (i.e. the eight diphthongs and triphthongs of Received Pronunciation in Roach, 2000: 21–4), and phrase breaks in Book 1 of Milton's *Paradise Lost* and find they are highly correlated. Table 2 shows a fragment from Book 1 (lines 17–18) tagged with ProPOSEL. Here, two out of three pre-boundary tokens (in bold) contain a complex vowel.

Table 1 Newsreel fragment annotated via ProPOSEL, mapping word tokens to: PoS tags from LOB and C5; DISC phonetic transcriptions; content-function word tags; syllable counts; lexical stress patterns; and phrase break annotations

```
the, ATI, AT0, 'Di, F, 1, 1, nonBreak
most, QL, AV0, 'm5st, C, 1, 1, nonBreak
eerie, JJ, AJ0, '7-rI, C, 2, 10, break
the, ATI, AT0, 'Di, F, 1, 1, nonBreak
scariest, JJT, AJS, 'sk8-rI-Ist, C, 3, 100,
  break
```

5 Significance Testing

We have observed this same marked association between junctures and diphthong/triphthong-bearing words (see items in bold) in spontaneous

Table 2 Extract from *Paradise Lost* (Book 1, lines 17–18) tagged via ProPOSEL with symbolic tokens for: syllable count; lexical stress, content-function word status; DISC phonetic transcription; DISC syllables mapped to stress weightings; and phrase break annotations

...th' upright heart and **pure,**
Instruct **me,** for Thou **know'st**...

```
['th', 'No_match‘, nonBreak]
['upright', ['2', ‘10', 'C', "'Vp-r2t", "'Vp:1 r2t:0"], nonBreak]
['heart', ['1', '1', 'C', "'h#t", "'h#t:1"], nonBreak]
['and', ['1', '1', 'F', "'{nd", "'{nd:1"], nonBreak]
['pure', ['1', '1', 'C', "'pj9R", "'pj9R:1"], break]
['instruct', ['2', '01', 'C', "In-'strVkt", "In:0 'strVkt:1"], nonBreak]
['me', ['1', '1', 'F', "'mi", "'mi:1"], break]
['for', ['1', '1', 'F', "'f$R", "'f$R:1"], nonBreak]
['thou', ['1', '1', 'F', "'D6", "'D6:1"], nonBreak]
["know'st", 'No_match', break]
```

speech from the twentieth century (see Listing 2 for this same fragment tagged with ProPOSEL).

> . . . every correspondent **here** | agrees | that the **final** | six mile stretch | through **Suchitoto** | is the most **eerie** | the **scariest** | bit of **road** | in El Salvador | . . .

However, what *we* construe as a pattern may in fact be quite random, simply because both phenomena crop up all over the place and are therefore bound to co-occur. Thus, we undertake significance testing to verify this insight using the chi-squared statistic.

As summarized in Table 3, we now have empirical evidence that the perceived correlation *is* statistically significant in three very different styles of speech: a scripted lecture; informal news commentary; and seventeenth century verse (Brierley, forthcoming). The rightmost column in the table gives the chi-squared χ^2 statistic for each experiment, calculated from raw counts for four groups of data (word tokens as breaks, non-breaks, diphthongs, or non-diphthongs), plus the total word count. Basically, we compare frequency distributions for target phenomena in our experimental dataset(s), namely the *observed* frequencies, to *theoretical* frequencies, those expected for chance co-occurrence; then, if observed and expected frequencies are sufficiently different and the difference exceeds some pre-determined confidence level, we surmise that the observed pattern/correlation is, in all probability, *not* random. In all our experiments, this pre-determined confidence level is very rigorous: 99% plus. This is given by the two-tailed *p-value* in the 'Result' column. A low *p*-value is a measure of the *im*probability of obtaining a similar result if our variables are truly independent; a *two-tailed* *p*-value considers the discrepancy in population means when either group has the larger mean. For 99% confidence, we are looking for a *p*-value of 0.01; our results (*p*-values of < 0.0001) out-perform this.

6 Conclusions

Through automated phrase break prediction, we model our understanding of human psycholinguistic chunking by identifying textual cues and superimposing *a priori* linguistic knowledge to develop feature sets for training classifiers. We then evaluate our classifier or model via the number of 'correct' boundaries retrieved in unseen text via comparison with that same text complete with gold standard

Table 3 Summary of empirical evidence of a statistically significant correlation between complex vowels and phrase breaks in three contrasting styles of British English speech

Text genre	Corpus size	'Vintage'	Participants	Result
Transcribed speech: Reith lecture	2293 words	British English 1980s	Single speaker Two annotators	$\chi^2 = 49$ 1 degrees freedom 2-tailed *p-value* *<0.0001*
Transcribed speech: Informal news commentary	7762 words	British English 1980s	Ten speakers Both genders Two annotators	$\chi^2 = 71$ 1 degrees freedom 2-tailed *p-value* *<0.0001*
Poetry: Milton's blank verse	6000 words	Early Modern English 1674	Poet *and* his publishers *and* his readers	$\chi^2 = 138$ 1 degrees freedom 2-tailed *p-value* *<0.0001*

phrase break annotations. The latter only represents one chunk parse for a given sentence, however, and classifier insertion and deletion errors may in fact capture variant intelligible and naturalistic phrasing. Moreover, models trained and tested on one corpus (i.e. one phrasing variant) perform less well in other domains.

One factor which may influence speakers' dynamic phrasing strategies and account for boundary placement or withholding in identical part-of-speech sequences is the immediate phonetic and rhythmic boundary context. Moreover, we support the view, cited earlier in this article (Section 4), that competent human readers are sensitive to the sounds and rhythms symbolized by the orthographic form, and that *internalized* and *projected* prosody informs silent reading. We therefore contend that phrase break models are insufficiently equipped to emulate human performance without prosodic information. This has yet to be tested.

We have therefore created ProPOSEL, a tool for annotating text with canonical and categorical representations of sound and rhythm as potential additional features for phrase break prediction. Such features constitute an alternative to acoustic representation of prosody, and are innately domain independent. We have a fully annotated dataset of spontaneous speech (1980s BBC radio newsreel from SEC) which we have used along with samples from two contrasting spoken genres to verify a perceived association between complex vowels and phrase breaks. Thus, presence or absence of a complex vowel could easily be incorporated as an extra non-traditional classificatory feature in phrase break models. The significant correlation between this vowel subset and syntactic-rhythmic junctures in seventeenth century English verse, for example, also advocates further application of ProPOSEL for text analytics and stylistic analysis of (literary) texts and authorship attribution via prosodic as well as lexical, syntactic, and semantic attributes. Moreover, this principle of uncovering intrinsic phrase break signifiers and stylistic patterns is transferable to Arabic via a pronunciation lexicon modelled on ProPOSEL, to complement traditional models of Arabic prosody (El-Qawasmeh and Aref, 1999).

References

Atterer, M. and Klein, E. (2002). Integrating linguistic and performance-based constraints for assigning phrase breaks. In *Proceedings of 19th Conference on Computational Linguistics – Volume 1*. Stroudsburg PA, USA: Association for Computational Linguistics, pp. 1–7.

Brierley, C. (forthcoming). *Prosody Resources and Symbolic Prosodic Features for Automated Phrase Break Prediction.* PhD thesis, School of Computing, University of Leeds.

Brierley, C. and Atwell, E. (2008a). Prosodic phrase break prediction: problems in the evaluation of models against a gold standard. *Traitement Automatique des Langues*, **48**(1).

Brierley, C. and Atwell, E. (2010). Holy Smoke: Vocalic Precursors of Phrase Breaks in Milton's *Paradise Lost*. *Journal of Literary & Linguistic Computing*, **25**(2): 137–51.

Donald, M. (1993). *Origins of the Modern Mind: Three Stages in the Evolution of Culture and Cognition*. Cambridge, Massachusetts: Harvard University Press.

El-Qawasmeh, E. and Aref, O. (1999). Prosody-analyzer: an intelligent expert system for prosody science in Arabic art. In *Proceedings of IASTED International Conference Software Engineering and Applications 1999*. Arizona, USA: ACTA Press, pp. 109–13.

Fodor, J.D. (2002). Psycholinguistics cannot escape prosody. In *Proceeding of Speech Prosody (SP-2002)*. Aix-en-Provence, FR: International Speech Communication Association (ISCA), pp. 83–90.

Liberman, M.Y. and Church, K.W. (1992). Text analysis and word pronunciation in text-to-speech synthesis. In Furui, S. and Sondhi, M. M. (eds), *Advances in Speech Signal Processing*. New York: Marcel Dekker, Inc.

***PASCAL* Thematic Programme.** (2008). http://www.cs.man.ac.uk/~neill/thematic08.html (accessed January 2010).

Roach, P. (2000). *English Phonetics and Phonology: A Practical Course,* 3rd edn. Cambridge: Cambridge University Press.

Sanderman, A. (1994). How can prosody segment the flow of (synthetic) speech? In *Proceedings of ESCA/IEEE Workshop on Speech Synthesis*. NY, USA: International Speech Communication Association (ISCA), pp. 147–50.

Taylor, L.J. and Knowles, G. (1988). *Manual of Information to Accompany the SEC Corpus: The machine*

readable corpus of spoken English. http://khnt.hit.uib.no/icame/manuals/sec/INDEX.HTM (accessed January 2010).

Taylor, P. and Black, A.W. (1998). Assigning Phrase Breaks from Part-of-Speech Sequences. *Computer Speech and Language*, **12**(2): 99–117.

Notes

1 Furui, S. *Selected Topics from 40 Years of Research on Speech and Speaker Recognition*. PowerPoint presentation. Keynote speech at *InterSpeech 2009*, Brighton, UK.

2 LOB: ***L***ancaster, ***O***slo, ***B***ergen tag set developed for the Spoken English Corpus. Online. http://khnt.hit.uib.no/icame/manuals/lobman/INDEX.HTM%20 (accessed 12 October 2010).

3 C5: CLAWS 5 tag set developed for the British National corpus. Online. http://ucrel.lancs.ac.uk/claws5tags.html (accessed 12 October 2010).

4 DISC: Computer phonetic characters. Online. http://www.let.uu.nl/~Hugo.Quene/personal/phonchar.html (accessed 12 October 2010).