Complex Vowels as Boundary Correlates in a Multi-Speaker Corpus of Spontaneous English Speech

Claire Brierley [1 2], *Eric Atwell* [2]

[1] School of Games Computing and Creative Technologies, University of Bolton, UK
[2] School of Computing, University of Leeds, UK
cb5@bolton.ac.uk, eric@comp.leeds.ac.uk

Abstract

We have found empirical evidence of a correlation in English between words containing complex vowels (diphthongs and triphthongs) and 'gold-standard' phrase break annotations in datasets as apparently different as seventeenth-century verse and a Reith lecture transcript on economics from the late twentieth-century. Spontaneous speech in the form of BBC radio news reportage from the 1980s again exhibits this statistically significant correlation for five out of ten speakers, leading to speculation as to why speakers should fall into two distinct groups. The experiment depends on the automatic annotation of text with *a priori* knowledge from ProPOSEL, a prosody and part-of-speech English lexicon.
Index Terms: prosody; phrase break prediction; pronunciation lexica; ASR; TTS

Introduction

Real-world knowledge of syntax is integral to the machine learning task of phrase break prediction: automatic identification of prosodic-syntactic boundaries in text which, on human evaluation, constitute natural and intelligible phrasing, and which serve as input features to a speech synthesizer for modelling intonation and duration over chunks of text designated by these boundaries. Traditionally, the phrase break classifier is trained on a speech corpus with *gold standard* part-of-speech (PoS) and boundary annotations and tested on an unseen subset from the same corpus, where the task is to recapture original boundary locations stripped from the test set by classifying tokens in the input text as either breaks or non-breaks. The breaks-correct measure (recall) and the proportion of true positives from all boundaries retrieved (precision) are combined in a single performance metric or F-score. However, an ongoing problem is that models trained on one corpus may not generalise to other domains because prosody is inherently variable: more than one natural and intelligible phrasing (*i.e.* more than one gold standard) exists for most sentences [1], [2], [3].

In a recent paper [4], we diagnosed a deficiency of real-world knowledge of *prosody* in a comprehensive survey of both rule-based and data-driven phrase break classifiers. We then argued the case for non-traditional prosodic features in the form of complex vowels (*i.e.* diphthongs and triphthongs) as potential phrase break correlates in English, based on (i) the observation that complex vowels occur at rhythmic junctures in poetry; and (ii) consensus within the ASR community that pauses affect vowel durations in adjacent words [5]. Finally, we obtained empirical evidence that diphthong-bearing content words are highly correlated with phrase breaks in a sample of contemporary British English speech in the form of a scripted lecture from the Aix-MARSEC corpus project [6] and, in a parallel study, seventeenth-century English verse [7]. This finding moderates, rather than contradicts, the study by Ananthakrishnan and Narayanan [8] which concludes that syllable tokens are poorer indicators of boundary events than POS-tags, based on word-final syllables (minus stress weightings) classed as breaks and all preceding syllables classed as non-breaks. The point is, beat-attracting complex vowels will often occur in word-initial or medial position.

In this paper, we extend our investigation of co-occurrence statistics for words containing complex vowels and phrase breaks to a dataset more akin to spontaneous speech, as suggested by Wichmann [9], and involving multiple speakers: namely, informal news commentaries in Section A of the Aix-MARSEC corpus. We present further evidence (mainly in the context of rule-based methods) of the limitations of syntax as a boundary correlate (§2) which then prompted the creation of an extended knowledge source for phrase break prediction in the form of ProPOSEL, our **Pro**sody and **PoS** **E**nglish **L**exicon [10] [11] (§3). We then implemented ProPOSEL as a text annotation tool in building a customised version of the dataset (§4.1). A full discussion of this build is intended for another paper; here we concentrate on how counts were obtained for diphthong-bearing breaks and non-breaks (§4.2, 4.3) before discussing results (§5) and drawing conclusions (§6).

2. Why syntax is not enough

Shallow or chunk parsing is a common methodology associated with phrase break prediction; there is consensus that prosodic phrasing is somehow simpler and flatter than syntactic structure. Hence *chink-chunk* [12] or CFP-type (**C**ontent-word, **F**unction-word, **P**unctuation) algorithms are used to identify low-level phrasal units in TTS – as in the Bell Labs speech synthesizer, for example [13]. Noun phrase (NP) chunks are also represented in terms of IOB tags [14], [15] where word tokens are classified as constituents (inside) or non-constituents (outside) of NPs or as initiating (beginning) NPs. Hence, "beginners" correspond to chinks: closed-class or function words immediately preceded by an open-class or content word – the signal for boundary insertion [12].

2.1. Inside or outside the chunk?

Earlier work by the authors [16], which attempted to define likely constituents of *prepositional* phrases using a chunk parser from the Natural Language ToolKit [15], demonstrates the shortcomings of such catch-all rules. We implemented a chunk rule or regular expression pattern over strings of Brown Corpus tags [17], where <IN> (*i.e.* any preposition) initiates a new chunk. The examples in Table 1 from Section A08 (1) and A09 (2-4) of our development set in Aix-MARSEC show prepositions (in **bold**) beginning or not beginning a prosodic phrase, as the speaker decides. Moreover, some forms elude placement inside or outside the prepositional phrase chunk.

1	**on** aeroplanes | *flying* **around** the Middle East and
2	| **on** top **of** a hill | *overlooking* Windhoek |
3	which French authorities | had made in their *handling* **of**
4	fly back **to** South Africa | *leaving* those | internal leaders |

Table 1: Prepositions and particles, plus gerunds and participles are difficult to categorise for prosodic-syntactic boundary placement.

Resolving the problem in (3) would be a straightforward case of re-tagging the word "handling" as a gerund or verbal noun and identifying this new tag as a likely constituent of prepositional phrases. Examples (2) and (4) could not be resolved so easily: we can imagine a legitimate amalgamation of the prosodic chunks in (2) and might wish to retain the option of including participles within prepositional phrase sequences; we would not want this option in (4) however, where the participle initiates a new syntactic chunk and has nothing to do with the prepositional phrase. Finally, what do we make of the chopped-up NP "those internal leaders" in (4)?

2.2. Category blends

Manning and Schutze [18] discuss ambiguity caused by non-categorical behaviour of parts of speech: individual words can be PoS-tagged differently in different syntactic contexts and, though allocated a particular PoS tag in a particular context, may retain and exhibit simultaneous behaviours *e.g.* "-ing" forms blurring the distinction between nouns and verbs. We have identified another blurred category where word forms lean towards "left" (outside) or "right" (inside) behaviours relative to prosodic boundaries depending on whether the token is tagged as a particle <RP> or a preposition <IN> respectively [1]. Such "tagging" is fluid in spontaneous speech.

The prototype rule discussed in section 2.1 inserts a boundary before true prepositions, PoS-tagged <IN>. There are six items tagged as true prepositions in the snippets in Table 1 and only one particle: "*back*" in (4). However, there does not seem to be much difference between the preposition-particles "*flying around*" in (1) and "*fly back*" in (4); and the absence of boundaries in speakers' chunking gives particles the benefit of the doubt here.

2.3 Rhythmic clout

As yet, we do not have a definitive set of content-function word groups mapped to parts-of-speech. The lexicon discussed in Section 3 uses the same default mappings of CF to Penn Treebank [19] tags as Busser *et al.* [20] and Bell [21]. Nevertheless, we are likely to be in accord about the CF category labels allocated to the sentence fragment in row 1 of Table 2.

1	F	F	C	F	F
2	before	the	hijacking	of	the
3	ANA+NRU	ANA	NRU	ANA	

Table 2: Binary classifications for syntax and rhythm

Row 3 represents rhythmic annotations from the Jassem Tier [22] in the Aix-MARSEC dataset. The label NRU (*narrow rhythm unit*) denotes either a stressed syllable in a monosyllabic word or a stressed syllable followed by a number of unstressed syllables in a bi-syllabic or polysyllabic word, while the label ANA (*anacrusis*) denotes an unstressed word-initial syllable or a sequence of unstressed syllables unattached to any NRU. Syntactically, the word "*before*" behaves as a function word in this example but rhythmically it shares attributes with content words, carrying a beat (primary stress) on a long vowel.

A similar situation arises if we view the whole of this opening sentence in A08.

A few days before the hijacking | of the TWA aircraft | soon after it took off from Athens airport || I was catching a similar TWA flight | from the same airport. ||

Here we have two instances of the preposition "*from*" – another grammatical or function word – which have different phonetic and rhythmic properties. We can verify this by inspecting the TextGrid file for section A0801 in Aix-MARSEC.

Tonic Stress Marks Tier	Jassem Tier
5.0099999999999998 "from" 5.0099999999999998	5.0099999999999998 "ANA" 5.0099999999999998
9.1639999999999997 "~from" 9.1639999999999997	9.1639999999999997 "NRU" 9.1639999999999997

Table 3: Even grammatical words exhibit prosodic variance.

Vowel reduction in the first occurrence of `/fr@m/` makes it an anacrusis. Conversely, the second instance of `/frQm/` is a narrow rhythm unit and even carries a pitch accent.

2.4. Taking stock

In summary, our example sentence exhibits all sorts of recalcitrant prosodic-syntactic behavior. A syntax-based rule which inserts a boundary before true prepositions or between content and function words, or between major syntactic groupings (NP/AVP: *A few days* versus PP: *before the hijacking*) is insensitive to speaker evidence here, where the adverbial qualifier is being treated prosodically as part of the prepositional phrase chunk since its role is to enhance the specificity of that phrase. The perceived need for prosodic features to complement syntax and punctuation in phrase break models, thus extending the knowledge source for this classification task, (*cf.* the recommendation that improvements in ASR depend on better knowledge sources [23]; and the trend towards supplementing raw training data with a priori knowledge [24]) has been the motivation behind our ProPOSEL lexicon discussed in the next section.

3. ProPOSEL: a linguistic repository

ProPOSEL assembles information from several widely-used lexica into one resource and is equally applicable to the range of computer speech and language applications and research projects which utilise such lexica. The lexicon comes in accessible text file format and the current version already classifies 104049 word forms under four variant PoS-tagging schemes to maximise linkage with speech corpora. Its multi-field format further maps each word form to default closed and open-class word categories; plus canonical phonetic transcriptions in SAM-PA and DISC; syllable counts; consonant-vowel (CV) patterns; and abstract representations of rhythmic structure or canonical stress labels. Moreover, the fine-grained syntactic, morphological and phonological information in ProPOSEL serves as a guide for developing lexica for new languages. An example entry group for the verb *secure* is given in Table 4.

Field	Sample	Field	Sample
1 wordform	secure	**4** SAM-PA	sI'kjU@R
2 C5 tag	VVI	**5** CUV2 tag & frequency rating	H2%,OA%
3 Capitalisation flag	0	**6** C5 tag & BNC frequency rating	VVI:25

Field	Sample	Field	Sample
7 syllable count	2	**12** C7 tag	VVI
8 lexical stress pattern	01	**13** DISC syllabified transcription	sI-'kj9R
9 Penn Treebank tag	VB	**14** DISC syllable-stress mapping	sI:0 'kj9R:1
10 content or function word tag	C	**15** CV pattern	[CV][CCVVC]
11 LOB tag	VB		

Table 4: ProPOSEL's 15 fields constitute a purpose-built repository of linguistic concepts in accessible text file format.

4. Significance Testing

So far, we have gathered empirical evidence from seventeenth century verse [7], and read speech from the twentieth century [4] which highlights a statistically significant correlation between words carrying complex vowels and phrase breaks in English via the chi-squared test for independence. We now extend this investigation to spontaneous speech, while reminding readers that the gold-standard phrase break annotations used still denote *intentional* as opposed to disfluent pauses.

4.1. Custom-built dataset

Our dataset has been custom-built to align word tokens and phrase break information from Aix-MARSEC, with syntactic information (*i.e.* LOB PoS-tags) from SEC and ProPOSEL (*i.e.* C5 PoS-tags), plus punctuation from SEC, plus shallow parse features (*i.e.* content-function word tags) and canonical phonetic transcriptions, again from ProPOSEL. The dataset of 7762 *word* tokens is compiled from ten different speakers, both male and female, and two different annotators: Gerry Knowles and Briony Williams, and is outlined in Table 5.

Section A file no.	Word count	Break count	Speaker gender	Annotator
A01	791	135	Female	Williams
A03	635	120	Male	Williams
A04	984	283	Male	Knowles
A05	803	200	Male	Knowles
A06	827	126	Male	Williams
A07	714	163	Male	Knowles
A08	629	120	Male	Williams
A09	789	199	Male	Knowles
A10	801	132	Male	Williams
A11	789	147	Male	Knowles

Table 5: Overview of dataset used

4.2. Obtaining the counts

Word and phrase break totals for each Section A sub-file in Table 5 constitute initial values for a 2 x 2 contingency table exploring the relationship between two distinct *groups*: diphthong-bearing words versus words with no diphthong (where the label 'diphthong' stands for *all* complex vowels); and two distinct *outcomes*: breaks versus non-breaks. Word counts were obtained by subtracting the break count (number of pauses) from the length of each file. Each word token was then classified as a break or non-break, depending on whether or not it was followed by a pause.

The total counts for diphthong and non-diphthong-bearing words were generated automatically for the most part but subject to manual inspection where prosodic information from ProPOSEL was (or appeared to be) missing. Missing information was due to a variety of factors. The dataset is spattered with proper nouns which do not appear in the lexicon. Furthermore, there are omissions passed down from source lexica: the noun *hijackings* from A08 does not appear as a plural in ProPOSEL, for example; and while the verb *rely* (in A11) carries a lexical stress pattern generated from one source, it has no values for fields 13-15 simply because they are generated from an alternative source which, surprisingly, does not include that word. Finally, there are some 'freaks of nature' such as the misspelling of *disillusioned* in Section A09 of the corpus: (A09|**dissillusioned**|non_break|AJ0|No_match). There are, in fact, several opportunities for a match here in ProPOSEL, depending on whether the word has been tagged in context as an adjective, past participle or past preterite.

4.3. Running the chi-squared test

Four counts were used to populate each 2 x 2 contingency table: word and break counts from Table 5 and total counts for diphthong-bearing (content and function) word *breaks* versus diphthong-bearing (content and function) word *non-breaks*. The remaining counts were generated from these as in this example from Section A09.

GROUPS	OUTCOMES		
	Breaks	Non-breaks	Totals
Diphthongs	**57**	**129**	186
No diphthongs	142	461	603
Totals	**199**	590	**789**

Table 6: A 2 x 2 contingency table records the observed frequency distribution for target groups and outcomes from corpus sample A09.

The chi-square test in this experiment determines whether the distribution resulting from observed frequencies in the shaded area in Table 6 is significantly different from the chance distribution anticipated from expected frequencies. The latter are calculated via marginal totals for rows and columns in the table: for example, the expected frequency for diphthongs classified as breaks is given by (199 / 789) * 186.

5. Discussion of results

Table 7 presents a summary of our findings. On the evidence of this study, the correlation between words carrying complex vowels and phrase breaks in English is a very significant stylistic feature of some speakers (at least 50%) but not others.

Section A file number	Ratio: words to breaks	Value of χ^2	2-tailed p-value	Significant?
A01	5.86 : 1	0.356	0.5510	No
A03	5.29 : 1	0.095	0.7585	No
A04	3.48 : 1	25.354	< 0.0001	Yes
A05	4.02 : 1	15.976	< 0.0001	Yes
A06	6.56 : 1	1.358	0.2439	No
A07	4.38 : 1	10.947	0.0009	Yes
A08	5.24 : 1	30.090	< 0.0001	Yes
A09	3.97 : 1	3.795	0.0514	Not quite
A10	6.07 : 1	0.873	0.3502	No
A11	5.37 : 1	7.885	0.0050	Yes

Table 7: Results per file for the chi-squared test

The presence or absence of this habit of speech seems to be independent of speaker gender and discernible (albeit subconsciously) to different listeners: both Knowles' and Williams' phrase break annotations are consistent with the findings. There also seems to be a link to phrasing density: on balance, the significant correlation occurs with speakers who pause more often. The densest phrasing occurs in A04, where dramatic reportage covers war-torn El Salvador. What is interesting in these findings is: (i) there is a stark contrast between these two types of speaker; and (ii) a multi-speaker corpus of spontaneous speech corroborates findings from

previous studies [4] [7], where the datasets might be described as 'composed speech'.

The diphthong counts err on the side of caution. The category of diphthong-bearing non-breaks is skewed somewhat by the high frequency of indefinite articles tagged with a full vowel, the canonical pronunciation: /eɪ/. Bearing this in mind, we re-calculated the value of chi-squared for files with non-significant correlations (*i.e:* A01, A03, A06, A09, A10), subtracting occurrences of /aːeɪ/ from the count for diphthong-bearing non-breaks and adding them to the non-diphthong-bearing non-breaks group. This made no difference to the result for each sub-file in all but one case: for A09, with 18 occurrences of /aːeɪ/, the re-calculated value of χ^2 is 8.579, with a two-tailed p-value of 0.0034.

Finally, calculating the chi-squared statistic for the correlation between diphthong-bearing words and breaks for the *whole* of Section A, we get a very significant result, for the data in Table 8: chi-squared equals 70.887 with one degrees of freedom and a two-tailed p-value which is less than 0.0001.

GROUPS	OUTCOMES		
	Breaks	Non-breaks	Totals
Diphthongs	550	1447	1997
No diphthongs	1075	4690	5765
Totals	1625	6137	7762

Table 8: A 2 x 2 contingency table records the observed frequency distribution for target groups and outcomes over all Section A files

6. Conclusion

We now have empirical evidence from three very different styles of speech (seventeenth century verse, a scripted lecture on economics, and informal news commentary) of a significant correlation between complex vowels and phrase breaks in English. Each dataset is relatively small, but the fact that this correlation is common to all suggests that this is a generic habit of English speech.

We believe this correlation merits further investigation via different genres, different corpora and different prosodic annotation schemes. We also believe other prosodic correlates will emerge from such work. Words bearing complex vowels can easily be identified via phonetic transcriptions, such as those in ProPOSEL, and like content-function word status, constitute domain-independent features. If incorporated into phrase break models, such a feature may be used to qualify over-predictive behaviour.

Given the conference theme of the 'universality' of prosody, we might hypothesise that while complex vowels seem to constitute phrase break *signifiers* in English, this may translate to a subset of the vowel system in other languages.

There are further questions and lines of enquiry. Why is it that speakers have fallen into two distinct groups? Is there an association between complex vowels and dramatic speech, in which phrasing density also plays a part? Do complex vowels have a particular emotive quality for English speakers? How does this translate to other languages? If some speakers favour diphthong-bearing words as *tonics* (*i.e.* nuclear prominences in tone groups) can this intrinsic quality be used in speaker identification? What about the ethical implications of this?

7. References

[1] Brierley, C. and Atwell, E., "Prosodic Phrase Break Prediction: Problems in the Evaluation of Models against a Gold Standard", in Traitement Automatique des Langues, 48(1):187-206, 2007b.

[2] Atterer, M. and Klein, E., "Integrating Linguistic and Performance-Based Constraints for Assigning Phrase Breaks" in Proc. 19th International Conference on Computational Linguistics (Coling 2002), 29-35, 2002.

[3] Taylor, P. and Black, A. W., "Assigning Phrase Breaks from Part-of-Speech Sequences" in Computer Speech and Language, 12(2):99-117, 1998.

[4] Brierley, C. and Atwell, E., "Exploring Complex Vowels as Phrase Break Correlates in a Corpus of English Speech with ProPOSEL, a Prosody and PoS English Lexicon" in Proc. INTERSPEECH'09, 2009.

[5] Vergyri, D., Stolcke, A., Gadde, V.R.R., Ferrer, L. and Shriberg, E., "Prosodic Knowledge Sources for Automatic Speech Recognition", in Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP 2003), 208-211, 2003.

[6] Auran, C., Bouzon, C. and Hirst, D., "The Aix-MARSEC Project: an Evolutive Database of Spoken British English", in Proc. Speech Prosody (SP-2004), 561-564, 2004.

[7] Brierley, C. and Atwell, E., "Holy Smoke: Vocalic Precursors of Phrase Breaks in Milton's *Paradise Lost*", submitted to Literary and Linguistic Computing, 2009.

[8] Ananthakrishnan, S. and Narayanan, S.S., "Automatic Prosodic Event Detection Using Acoustic, Lexical, and Syntactic Evidence", in IEEE Transactions on Audio, Speech, and Language Processing (TASLP 2008), 16(1):216-228, 2008.

[9] Wichmann, A. Personal communication. ICAME 30, Lancaster. 2009.

[10] Brierley, C. and Atwell, E., "ProPOSEL: A Prosody and POS English Lexicon for Language Engineering" in Proc. 6th Language Resources and Evaluation Conference (LREC 2008), 2008a.

[11] Brierley, C. and Atwell, E., "A Human-oriented Prosody and PoS English Lexicon for Machine Learning and NLP' in Proc. 22nd International Conference on Computational Linguistics (Coling 2008), Workshop on Cognitive Aspects of the Lexicon, 2008b.

[12] Liberman, M. Y. and Church, K. W., "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., 1992.

[13] Abney S., "Introduction to Computational Linguistics: Chunk Parsing." PowerPoint presentation. Online. Accessed: Dec. 2006 www.cs.um.edu.mt/~mros/csa2050/ppt/chunking.ppt 2006

[14] Ramshaw, L. A. and Marcus, M. P., "Text chunking using transformation-based learning", in Proc. 3rd. ACL Workshop on Very Large Corpora, pp.82-94, 1995.

[15] Bird, S. Klein, E. and Loper, E., "Natural Language Processing with Python", Sebastopol: O'Reilly Media Inc. 2009.

[16] Brierley, C. and Atwell, E., "An Approach for Detecting Prosodic Phrase Boundaries in Spoken English", in ACM Crossroads Journal, 14(1), Online: http://www.acm.org/crossroads/xrds14-1/nltklite.html, accessed 15 July 2009

[17] Greene, B. B. and Rubin, G. M. Rubin. "Automatic grammatical tagging of English". Providence, R.I.: Department of Linguistics, Brown University. 1981.

[18] Manning, C. D. and Schutze, H., "Foundations of Statistical Natural Language Processing". Cambridge, Massachusetts: The Massachusetts Institute of Technology. 1999.

[19] Santorini, B. "Part-of-speech tagging guidelines for the Penn Treebank Project". Technical report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania. 1990

[20] Busser, B., Daelemans W. and Van den Bosch, A., "Predicting phrase breaks with memory-based learning." 4th ISCA Tutorial and Research Workshop on Speech Synthesis, Edinburgh. 2001.

[21] Bell, P. "Adaptation of Prosodic Phrasing Models." MPhil Thesis, University of Cambridge Online: Accessed Sept. 2009 http://homepages.inf.ed.ac.uk/s0566164/m_thesis.pdf. 2005

[22] Bouzon, C. and Hirst, D., "Isochrony and Prosodic Structure in British English", in Proc. Speech Prosody 2004. 2004.

[23] Furui, S. "Selected topics from 40 years of research on speech and speaker recognition". Keynote speech in InterSpeech 2009, Brighton. 2009.

[24] PASCAL Thematic Programme 2008. Online: Accessed August 2008 http://www.cs.man.ac.uk/~neill/thematic08.html