Exploring Complex Vowels as Phrase Break Correlates in a Corpus of English Speech with ProPOSEL, a Prosody and POS English Lexicon.

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
Real-world knowledge of syntax is seen as integral to the machine learning task of phrase break prediction but there is a deficiency of a priori knowledge of prosody in both rule-based and data-driven classifiers. Speech recognition has established that pauses affect vowel duration in preceding words. Based on the observation that complex vowels occur at rhythmic junctures in poetry, we run significance tests on a sample of transcribed, contemporary British English speech and find a statistically significant correlation between complex vowels and phrase breaks. The experiment depends on automatic text annotation via ProPOSEL, a prosody and part-of-speech English lexicon.

Index Terms: prosody; real-world knowledge for machine learning; phrase break prediction; text-to-speech synthesis

1. Introduction
The goal of automatic phrase break prediction is to identify prosodic-syntactic boundaries in any given text which, on human evaluation, constitute natural and intelligible phrasing, and which can confidently be used 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 reference dataset from the same corpus; its task is to recapitulate original boundary locations stripped from the test set by classifying tokens in the input text as either breaks or non-breaks.

Real-world knowledge of syntax is seen as integral to this machine learning task but there is a deficiency of a priori knowledge of prosody in both rule-based and data-driven classifiers. We therefore explore prosodic features in the form of complex vowels as potential phrase break correlates, based on the observation that complex vowels tend to occur at rhythmic junctures in poetry.

In a previous paper [1], we have discussed machine-learning techniques and evaluation metrics used in phrase break prediction, plus the inherent problem of prosodic variance: more than one natural and intelligible phrasing (i.e. more than one gold standard) exists for most sentences; and models trained on one corpus may not generalise to other domains. Here we begin with an overview of features and feature sets used when predicting boundaries, before hypothesizing and testing non-traditional, vocalic phrase break correlates in a sample from the Aix-MARSEc corpus of English speech [2] via the chi-squared test for independence. This entails automatic annotation of the dataset with domain knowledge from ProPOSEL, a prosody and syntax English lexicon [3], [4].

2. Features used in phrase break prediction
Syntactic features are integral to phrase break prediction because of the overlap between syntactic and prosodic phrasing. The boundary annotation / | / in the following sentence taken from a landmark psycholinguistic study [5], represents human consensus on the best place to pause:

After the cold winter of that year | most people were totally fed-up.

The least sensitive and most transferable syntactic feature for predicting phrase breaks is content-function word status. Under this rule-based scheme, boundaries are inserted after punctuation and between open-class content words or chunks and closed-class function words or chunks [6].

For our model sentence, function-word groups captured by a standard CFP algorithm match syntactic units delineated by the Link parser [7]:

\[
\begin{align*}
PP & \quad \text{After the cold winter} \\
PP & \quad \text{of that year} \\
NP & \quad \text{most people} \\
VP & \quad \text{were totally fed-up}
\end{align*}
\]

Edinburgh’s Festival speech synthesis system implements a stochastic model for phrase break prediction which requires more discrete syntactic information from part-of-speech (PoS) tags.

After CTS the AT0 cold_AIO winter_NN1 of Pref that DT0 year_NN1 most DT0 people_NNO were VBD totally_AIO fed-up_AIO .

Our sample sentence is annotated with the British National Corpus C5 PoS tag set [23].

The Festival classifier integrates two feature sets: localised observation probabilities of PoS trigrams given juncture type, conditioned on long-distance syntactic information from a high-order n-gram juncture sequence model [8].

Building on the intuition that phrase breaks occur between major syntactic units {NP; VP; PP; ADJP; ADVP}, Koehn et al., (2000) use a sophisticated feature set [9] incorporating binary flags for whether or not the token initiates a major phrase or sub-clause. Their impressive prediction rate of 90.8% for boundary detection is partly accounted for by their incorporation of a feature derived from hand-labelled transcriptions: i.e. accent status of words adjacent to the boundary site; whereas the aim is to predict prosodic events like phrase breaks and accents automatically. Taylor and Black [8], and more recently Ingulf sen et al. [10], have demonstrated that punctuation is the single most important source of information for phrase break classification, finding approximately 50% of all breaks. Other text-based features which have been used to supplement syntactic features, include: word counts denoting length of utterance and distance of potential boundary site from start and
end of sentence [11]: total number of words and syllables, plus
distance from start and finish of utterance in words, syllables
and stressed syllables, plus distance of potential boundary site
from last punctuation mark [9], [12].

Recent work [10], [13] revisits syntactic features to
determine the effectiveness of deep versus shallow linguistic
representations for phrase break prediction. The best
performing models in these studies use a combined set of
long-range parse features and shallow representations
incorporating different levels of granularity: CFP tags and PoS
trigrams.

Non-traditional features in the form of syllable counts
have previously been implemented in syntax-based phrase
break models for English to regulate the number of syllables in
any one intonational phrase [14]; and as a distance metric for
encoding global information in the sentence [15]. A recent
study by Ananthakrishnan and Narayanan [16] attempts to
integrate the prediction of accents and boundaries based on
combined feature streams (acoustic, lexical and syntactic) and
finds that lexical syllable tokens, augmented with canonical
stress labels derived from an open source pronunciation
lexicon, are effective for accent detection but not for boundary
prediction.

3. Hypothesizing non-traditional phrase
break correlates

Ananthakrishnan and Narayanan conclude that syllable tokens
are poorer indicators of boundary events than PoS tags. However,
this conclusion is based only on word-final syllable
tokens minus stress weightings for the phrase break prediction
task; word-initial and medial syllables are automatically
classed as non-breaks because they are never immediately
followed by boundary tokens.

We wish to question the assumption that non word-final
syllabic nuclei (e.g. the second syllable in secure) have no
influence on boundary placement and to test the hypothesis
that complex vowels – i.e. diphthongs and triphthongs – might
emerge as useful predictive features for phrase break models,
irrespective of where they occur within a word. There is
consensus within the ASR research community that pauses
affect vowel durations in preceding words [17]. We wish to
reverse the perspective on prepausal lengthening and ask to
what extent a domain-independent feature like complex
vowels may be said to induce boundaries.

The intuition that the presence of complex vowels in
(content) words increases the likelihood of their being
classified as breaks comes from poetry [18], where diphthongs
and triphthongs seem to be associated with rhythmic junctures.
This happens within lines and across lines as in Blake’s The
Tyger (circa 1794):

Tyger! Tyger! | burning bright |
In the forests | of the night |

4. Leveraging real-world knowledge
of prosody from the lexicon

One of the thematic programmes for PASCAL-2 (2008)
identifies a current interest in, and trend towards, leveraging
real-world knowledge to enhance performance in machine
learning in a variety of application domains, including text and
language processing, where previously little a priori
knowledge has been assumed on the part of the learning
mechanism. Our survey reveals a deficiency of a priori
linguistic knowledge of prosody in the feature sets typically
used in rule-based and data-driven phrase break models. In
contrast, a competent human reader is able to project holistic
linguistic insights, including projected prosody, onto text and
to treat them as part of the input [19]. It is our contention that
human readers may use the sound patterns inherent in complex
vowels as linguistic signs for phrase breaks in as yet undefined
contexts. Such signs can be extracted from the lexicon and
presented as input features for the phrase classifier in the
same way that real-world knowledge of syntax is represented
in PoS tags.

4.1. ProPOSEL: a prosody and PoS English lexicon

ProPOSEL [3], [4] is a prosody and PoS English lexicon
derived from several widely-used lexical resources for
computer speech and language. ProPOSEL’s multi-field
format classifies 104049 word forms under four variant PoS-
tagging schemes mapped to default closed and open-class
word categories; plus canonical phonetic transcriptions;
syllable counts; consonant-vowel (CV) patterns; and abstract
representations of rhythmic structure or canonical stress labels.
An example entry group for the verb secure is given in Table 1.

<table>
<thead>
<tr>
<th>Field</th>
<th>Sample</th>
<th>Field</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 wordform</td>
<td>secure</td>
<td>9 Penn Treebank tag</td>
<td>VB</td>
</tr>
<tr>
<td>2 C5 tag</td>
<td>VVI</td>
<td>10 content or function word tag</td>
<td>C</td>
</tr>
<tr>
<td>3 Capitalisation flag</td>
<td>0</td>
<td>11 LOB tag</td>
<td>VB</td>
</tr>
<tr>
<td>4 SAM-PA</td>
<td>slˈkjʊər</td>
<td>12 C7 tag</td>
<td>V VI</td>
</tr>
<tr>
<td>5 CV2 tag &amp; frequency rating</td>
<td>h2%</td>
<td>OA%</td>
<td>13 DISC, syllabified transcription</td>
</tr>
<tr>
<td>6 C5 tag &amp; BNC frequency rating</td>
<td>VVI:25</td>
<td>14 DISC syllable-stress mapping</td>
<td>slˈ0%ˈkjʊər:1</td>
</tr>
<tr>
<td>7 syllable count</td>
<td>2</td>
<td>15 CV pattern</td>
<td>[CV][CCVC]</td>
</tr>
<tr>
<td>8 lexical stress pattern</td>
<td>01</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: ProPOSEL’s 15 pipe-separated fields constitute a purpose-
built repository of linguistic concepts in accessible text file format.

To investigate the correlation between complex vowels
and phrase breaks, we have automatically tagged an extract
from the Aix-MARSEC corpus with shallow parse features
and canonical phonetic transcriptions from ProPOSEL, and
run a chi-squared test to determine whether this correlation is
statistically significant or not. We have used the same
development sets as in previous studies [1], [20]: a BBC radio
recording from the 1980s of a Reith lecture in Section C of the
corpus, with illustrative examples drawn from sections A08
and A09: informal news commentaries.

Preparing the dataset prior to dictionary lookup was non-
trivial and involved several stages. The first task was to map
annotation tiers in overlapping subfiles in the Aix-MARSEC
corpus to a set of word breaks in the PoS-tagged text in the
Spoken English Corpus [21], discrepancies intervene: compounds and abbreviations are handled
differently in both datasets, for example ($\xi.4.3$). Next, the
corpus was re-tagged with the PoS tag scheme used in the
lexicon i.e. a discriminating tagset (LOB) was collapsed into a
sparser one (C5) ($\xi.4.4$). Finally, desired information from the
lexicon was projected onto the dataset by matching word
C5 pairings ($\xi.4.5$).
4.2. Mapping tiers in Aix-MARSEC

The Aix-MARSEC Corpus has multi-level prosodic annotation tiers aligned with the speech signal; the two tiers used in this study are for plain text plus intonation units (IUs) delineated by phrase break mark-up / \ . The SAMP-PA transcriptions from the syllables tier were not used in our study because we are interested in predictive features derived from speaker-independent and domain-independent citation forms in ProPOSEL which can be superimposed on any unseen English text – for example, seventeenth century English verse cf. [18].

Each section in Aix-MARSEC is split up into a series of much smaller, overlapping TextGrid files. Merging the text and IUs tiers was therefore accomplished on a file-by-file basis, using interval tokens to retrieve a match between tiers. The resulting list objects were concatenated in a final list – listAllText – ready for merger with the corresponding file in the Spoken English Corpus (SEC) to capture PoS-tags.

4.3. Merging Aix-MARSEC and SEC files

The target data structure for dictionary lookup (§4.5) is a nested list where each index holds values for: word token; break class; punctuation; and PoS-tag. Capturing PoS tags from SEC entailed looping over two parallel lists of unequal length – listAllText and a list of word_PoS pairings from SEC – a process complicated by the fact that compound words are represented differently in both datasets, and furthermore, that punctuation in SEC does not always correspond to boundaries or placeholders in Aix-MARSEC. Such problems are exemplified in Listing 1 (section A09 of the corpus), where we find different representations for the compound adjective: cross-ethnic; variant phrasing for the fragment: who two years ago; no apparent placeholder in Aix-MARSEC following the boundary after ago; no punctuation in SEC after the word together, which is marked as a phrase break in AIX-MARSEC.

4.4. Mapping between PoS tag sets using ProPOSEL

List indices in listAllText have now acquired PoS tags and, if present, punctuation from the semi-automated process just described. However, the recommended lookup strategy with the prosody and PoS lexicon is via compound dictionary keys comprising word_C5 pairings. A range of tagsets (Penn, LOB and C7) were mapped to C5 as part of lexicon build; and ProPOSEL’s software tools provide solutions for mapping between schemes (Brierley and Atwell, 2008a). In the present study, a more discriminating tagset – LOB [22] – is collapsed into a sparser scheme (C5). As part of this process, enclitics in LOB are re-formatted in a style compatible with the lexicon; instances such as: [BEDZ, ‘was’, ‘>’, ‘\XNOT’, ‘\nt’, ‘\ct’] and [‘WP’, ‘who’, ‘>’, ‘\n’, ‘\nt’, ‘\ve’, ‘\ct’] are transformed into: [BEDZ\XNOT, ‘\wasn’t’] and [‘WP+\n’, ‘\who’\ve’].

4.5. Dictionary lookup and text annotation

Nested arrays in listAllText are finally augmented with domain knowledge of prosody (e.g. DISC fields in ProPOSEL) and coarse-grained syntactic information (default content-function word tags) via interaction with ProPOSEL. Listing 2 first builds an instance of the dictionary object proPOSEL with compound keys word_C5 tuples mapped to selected values. Python’s itertools module is then used to loop through two parallel iterables: listAllText and match, a sequence of word_C5 tuples from the same dataset. Items in the latter are compared against ProPOSEL’s keys; a successful match appends dictionary values associated with those keys to the parallel nested position in listAllText.

Listing 1: Transcriptions of the same utterance in two different versions of the corpus exhibit variant phrasing.

```
Aix-MARSEC SEC
['ethnic', '48.69', 'I'] JJ ethnic
['#', '48.74', 'P'] ,
['cross', '49.12', 'non-break'] JJ cross-ethnic
['ethnic', '49.53', 'I'] ,
['#', '49.62', 'P'] ,
['and', '49.88', 'non-break'] CC and
['political', '50.41', 'non-break'] JJ political
['parties', '50.88', 'I'] ,
['#', '51.39', 'P'] ,
['who', '51.59', 'non-break'] NNS parties
['two', '51.73', 'non-break'] NNS two
['years', '52.04', 'non-break'] RB ago
['ago', '52.44', 'I'] ,
['came', '52.70', 'non-break'] VBD came
['together', '53.12', 'I'] RB together
['to', '53.34', 'non-break'] TO to
```

Listing 2: Intersection between the dictionary object proPOSEL and the sequence object match appends dictionary values to the parallel position in listAllText.

```
proPOSEL = dict(zip(lex_keys, lex_values))
match = [index[0], index[5]] for index in listAllText
for x, y in itertools.zip(match, listAllText):
    if x in proPOSEL.keys():
        y.append(buildDict(x))
else:
    y.append(’No match’)  
```

5. Significance Testing

Each word in the sample was assigned to one of four different categories and counts for each category were entered in a 2 x 2 contingency table (Table 2) ready for the chi-square test. The category label of diphthongs is used here to denote all complex vowels. The total word count is simply the length of listAllText minus the count for unmatched items; these were not included in the final calculation and figures used in Table 2 reflect this.

<table>
<thead>
<tr>
<th>GROUPS</th>
<th>OUTCOMES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Breaks</td>
</tr>
<tr>
<td>Diphthongs</td>
<td>201</td>
</tr>
<tr>
<td>No diphthongs</td>
<td>437</td>
</tr>
<tr>
<td></td>
<td>638</td>
</tr>
<tr>
<td></td>
<td>(696 – 58)</td>
</tr>
</tbody>
</table>

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

The chi-square test in this experiment determines whether the distribution resulting from observed frequencies in the shaded area in Table 2 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 (638 / 2293) * 499. Table 3 presents observed versus expected frequencies (given in bold and expressed as whole numbers for clarity of presentation) for all four categories.
transitions between high and low tones: the hallmark of salient tone groups) because vocalic glides facilitate sudden pitch
according to the formula:

$$\chi^2 = \sum \frac{(f_i - f_e)^2}{f_e}$$

The null hypothesis $H_0$ assumes that the distributions will be the same or that the difference will not exceed some critical value. In our case, however, $H_0$ can be rejected because the association between groups and outcomes turns out to be extremely statistically significant: chi squared equals 49.28, with one degree of freedom, and a two-tailed p-value which is less than 0.0001. This p-value represents the odds ratio for achieving the same result through random sampling. Finally, since there are only four diphthong-bearing function words which are also classified as breaks in this sample, we can hypothesize that the significant correlation is actually between diphthong-bearing content words and phrase breaks.

6. Conclusion

Our survey of features used in phrase break prediction highlights a deficiency of a priori knowledge of prosody in both rule-based and data-driven language models. The authors concur with studies that recognise how, even in silent reading, humans project prosody onto text and treat it as part of the input. Hence we have developed ProPOSEL, a domain-independent lexical resource and prosodic-syntactic text annotation tool.

There is consensus in the ASR community that pauses affect vowel durations in adjacent words. Based on intuitions from poetry and concurrent work [18], we have redefined this causal relationship and interpreted complex vowels as phrase break signifiers. From significance tests on a sample of contemporary British English speech from the Aix-MARSEC Corpus, plus seventeenth century English verse (ibid.), we now have empirical evidence that diphthong-bearing content words are highly correlated with phrase breaks.

Since accent status of pre-boundary words has already proved effective in phrase break prediction, future work will focus on the correlation of complex vowels, salient pitch accents and boundaries to explore a linguistically-motivated hypothesis: native English speakers subconsciously favour diphthong-bearing words as tonics (i.e. nuclear prominences in tone groups) because vocalic glides facilitate sudden pitch transitions between high and low tones: the hallmark of salient pitch accents.

7. References


<table>
<thead>
<tr>
<th>GROUPS</th>
<th>OUTCOMES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Breaks</td>
</tr>
<tr>
<td>Diphthongs</td>
<td>201</td>
</tr>
<tr>
<td>No diphthongs</td>
<td>437</td>
</tr>
<tr>
<td>Total</td>
<td>1398</td>
</tr>
</tbody>
</table>

Table 3: Observed and expected frequencies are used to find the value of $\chi^2$ in this test for independence.