Holy smoke: vocalic precursors of phrase breaks in Milton's *Paradise Lost*

Claire Brierley
University of Bolton, UK,
University of Leeds, UK

Eric Atwell
University of Leeds, UK

Abstract

We report on a significant correlation between lexical items containing complex vowels in their present day canonical forms, and prosodic-syntactic boundaries in Milton's *Paradise Lost*, where all line terminals, whether end-stopped or run-on, plus line-medials with associated punctuation, constitute boundary tokens and equate to *gold-standard* phrase break annotations. Real-world knowledge of present day English pronunciation is projected onto each word token in two different versions, constituting two phrasing variants, of Book 1 of the poem via ProPOSEL, a prosody and part-of-speech English lexicon developed by the authors; and pertinent differences in place of articulation of English vowels in Milton's day and ours are also discussed. The chi-squared test for independence returns a two-tailed *P*-value of less than 0.0001 for the association of this vowel subset and phrase breaks in both samples. This leads us to speculate that the poet's *unpremeditated* use of complex vowels—which slow down verse movement in *Paradise Lost* and thus generate rhythmic junctures—may represent a phrasing device habitual not just to poets but to native English speakers in general. Concurrent work on a corpus of present-day British English speech corroborates our findings. Hence, complex vowels may constitute new predictive features in phrasing models for English.

Correspondence:
Claire Brierley, Business and Creative Technologies,
University of Bolton,
Deane Road,
Bolton BL3 5AB, UK
E-mail:
cb5@bolton.ac.uk;
claireb@comp.leeds.ac.uk

1 Introduction

Automatic phrase break classifiers for text-to-speech synthesis systems currently rely on syntactic (e.g. part-of-speech) and text-based (e.g. punctuation) features for recapturing and emulating human parsing and phrasing strategies encapsulated by *gold standard* phrase break annotations in speech corpora used for training and testing such classifiers. In this case, the annotations correspond to listeners' perceptions of pauses in the speech stream as speaker prosody differentiates between syntactically coherent clusters of words: the *chunking* phenomenon (Abney, 1991).

This article is based on observation and intuition: the presence of diphthongs and triphthongs at phrase breaks or *rhythmic junctures* in poetry suggests that new categorical *prosodic* features for boundary prediction may be derived from lexical items which incorporate this subset of English vowels—henceforth referred to as complex vowels for convenience—in their canonical phonetic

doi:10.1093/llc/fqq004 Advance Access published on 14 April 2010

transcriptions. The following examples from English binary verse illustrate this association:

Tyger! Tyger! burning bright
In the forests of the night,
What immortal hand or eye
Could frame thy fearful symmetry?
(The famous opening stanza of Blake's
The Tyger, *circa* 1794)

The dove descending breaks the air
With flame of incandescent terror
Of which the tongues declare
The one discharge from sin and error.
The only hope, or else despair
Lies in the choice of pyre or pyre –
To be redeemed from fire by fire.
(Eliot's Pentecostal invocation in
part IV of *Little Gidding*, 1942)

The term rhythmic juncture is used here to denote in-line caesuras and line ends; and the lexical items of interest are words which immediately precede these boundaries and which bear complex vowels, often in the primary syllable, in their Present Day British English (PresE) canonical forms for both spelling and pronunciation. A vowel is said to be *complex* when vowel quality changes (from initial to target quality) within a single syllable (Maidment, 2009). Our subset includes words like *fire* and *power* where syllabification is dubious (one syllable or two?) and where transcriptions for standard English pronunciation vary between lexica (Section 4). In plain text view, some of these junctures are not physically represented, either by punctuation or by line and verse endings: these are the unmarked, in-line caesuras. Nevertheless, the following pre-boundary tokens with vocalic glides are posited for Blake's stanza: {*Tyger; bright; night; eye; frame*}; and for Eliot's: {*air; flame; declare; hope; despair; choice; pyre; fire*}. In the case of recital, such choices (one might even say *classifications*) reflect '…speakers' perceptions about the divisibility of text…' (Sinclair and Mauranen, 2006, p. xvi); in silent reading, they reflect *projected* prosody (Fodor, 2002).

The present study undertakes an investigation to assess the degree of correlation between words bearing gliding vowels and marked boundaries in a classic Early Modern English (EModE) literary text: Book I of Milton's *Paradise Lost*. Software tools from version 0.9.8 of NLTK, the Natural Language ToolKit (Bird *et al.*, 2009) and Natural Language Processing (NLP) techniques are used to tokenize the text and then annotate it with 'projected prosody' from ProPOSEL, a prosody and part-of-speech English lexicon (Brierley and Atwell, 2008a,b). The principal dataset is drawn from Dartmouth College's eText of the 1674 edition of the poem (Luxon, 2009). The second dataset is a readily available, modern English version of Book 1: the 1992 eText from Project Gutenberg, also distributed in NLTK's corpora; although this does not entirely reflect original punctuation in the 1667 and 1674 editions, it is assumed to be a reliable phrasing variant.[1]

The article discusses: the use of punctuation as a boundary marker in previous studies based on literary corpora (Section 2); the tokenization and classification of each word in the samples as a break or non-break (Section 3); the further annotation of each word token with its phonetic transcription via ProPOSEL, plus pertinent similarities and differences between EModE and PresE pronunciation (Section 4); significance testing of the correlation between complex vowels and boundaries in both samples using the chi-squared statistic (Section 5); and telling examples in Book 1 of Paradise Lost where unmarked conceptual boundaries (i.e. in-line caesuras) are signified by complex vowels (Section 6).

2 Punctuation as a Prosodic Template

The symbolic representation of pauses via punctuation has been used in a number of exploratory studies of stylistic evolution in EModE blank verse. Pause patterns in Shakespeare's work, originally obtained from inclusive counts for punctuation at designated within-line positions for each play (Oras, 1960), have recently been subjected to formal statistical analysis (Jackson, 2002) and found to be good guides to chronology: plays of the same period, and in some instances,

Table 1 Inverted accents in bold reinforce midline phrase breaks in Milton's verse

	Juncture type	Example	Line
1	A stop between two accents	'...for ever **dwells! Hail**, horrors...'	250
2	A *de-accented* stop	'...hast **ened: as** when bands...'	675

chronologically adjacent plays, reveal progressive experimentation with the placement of stops (i.e. punctuation) within the line. A similar phenomenon, that of increasing divergence between metrical (the lines of verse) and grammatical units in the Shakespearian chronology, is discussed in a much earlier paper (Langworthy, 1931). Here, a quotient is obtained by dividing the number of parallel line types (e.g. where independent clauses are wholly contained *within* a line) by the number of divergent types in a given play and findings show that, whereas for very early plays the quotient is relatively high (40.00 and above), for later plays like *Hamlet* and *Macbeth* it is much lower (4.21 and 1.89, respectively) and for very late plays like *The Winter's Tale*, lower still (0.47). The following extract from Act I, Scene VII of *Macbeth* illustrates naturalistic prosodic-syntactic chunking both *within* and *between* lines, simulated via shifting placement of marked caesuras, and verse-sentence divergence facilitated by enjambement:

> Macb. If it were done, when 'tis done, then 'twer well,
> It were done quickly: If th' Assassination
> Could trammell vp the Consequence, and catch
> With his surcease, Successe: that but this blow...

Langworthy (1931) observes that the poet '...write[s] his sentence[s] almost as though he had forgotten all about the line, and yet fulfills the line requirements with the off-hand ease of a supreme master of metrics'.

Turning now to *Paradise Lost*, Banks (1927) sets out to identify the 'prosodical devices' by which Milton '...makes the rhythms of his units of thought independent of the single lines and of each other, thus achieving the effect of irregular paragraphs'. Again, punctuation in the form of terminal and medial stops {*periods; colons; question* and *exclamation marks*} is used to delineate verse paragraphs; but Banks' real interest is in classifying these joints in the verse in terms of trigrams consisting of a stop bordered by antecedent and posterior syllables which may or may not carry a beat. He identifies two prosodic patterns—the first of which is high-profile—which reinforce the midline break in Milton's verse through accent inversion, rather in the way of magnets: *like accents repel!* Examples of these are shown in Table 1, with line references for Book I.

The present study also aims to explore *prosodical devices* associated with phrase breaks in *Paradise Lost*: namely, to test the intuition that diphthongs and triphthongs act as *vocalic precursors* of boundaries. It is assumed that punctuation in the principal dataset is sufficiently representative of the poet's phrasing and that *all* punctuation is significant. Such assumptions are supported by precedent; the terms *punctuation* and *pauses* have been used interchangeably in studies considered in this section; and inclusive counts for punctuation have incorporated: (1) major *and* minor boundary types; (2) and *medial* as well as terminal stops. A further point is that punctuation is a primary feature used in language models for the machine learning of task of phrase break prediction: Ingulfsen *et al.* (2005) even make the point that '...punctuation is used by writers to indicate rhythm and pausing'.

Experimentation (Section 5) to determine whether the co-occurrence of complex vowels and pauses in Book I of *Paradise Lost* is statistically significant is based on a boundary count which includes *all* line-terminals in the count, irrespective of whether they are marked by punctuation or not. The mechanics and justification for this are covered in Section 3 and revisited in Section 6, where other types of conceptual boundary are also discussed.

3 Issues of Tokenization and Phrase Break Classification

The authors have experimented with two different approaches to tokenization. Initially, for the Gutenberg sample, CorpusReader and Tokenizer Classes in NLTK 0.9.8 were used to simultaneously read in the unprocessed contents of this eText of *Paradise Lost* and to store these contents as a nested list of line tokens: the variable `milton` in the commented code snippet in Listing 1. The first line of the poem is then accessed via its list index, in this case `milton[2]`; and slice notation is used to assign the whole of Book I to a variable of the same name—`book1`—and to access and print out the first complete sentence: `milton[2:18]` by way of illustration. As an aside, punctuation in the output from Listing 1 accords well with the same excerpt as it appears in an original 1674 edition of the poem, viewable as a *.jpg* image on the internet (Geraghty, 2003).

Listing 1 provides a solution for preserving verse *form* during tokenization. The next step is to transform `book1` so that every word in a line is captured as a separate token which can eventually be counted; each of these tokens is then classified as a break or a non-break, on the basis of two break indicators: associated punctuation and/or line terminal status. For the Gutenberg text, this was initially accomplished using NLTK's `WhitespaceTokenizer()`, which captures any attendant punctuation as part of each word token and thus facilitates the process of break classification. As an example, Listing 2 displays three phrase break tokens highlighted in bold: *Chaos*; *or* and *more*.

However, an alternative approach has since been used by the authors (Brierley and Atwell, 2009b) and has now been applied to both datasets in this study. The customised verse tokenizer in Listing 3 uses a regular expression (*cf.* ibid. for step-by-step decomposition and explanation of this regular expression) to differentiate word-internal from

Listing 1 NLTK's `LineTokenizer()` captures each line of verse in the Gutenberg eText as a separate token of type `string`

```
import nltk, re
from nltk.tokenize import * # import all Tokenizer Classes from tokenize package

tokenizer = LineTokenizer(blanklines='discard') # initialize LINE tokenizer

# Read in & tokenize lines in one step
milton = tokenizer.tokenize(nltk.corpus.gutenberg.raw('milton-paradise.txt'))

book1 = milton[2:800] # start and end LINE indexes for Book I of the poem
```

```
>>> for line in milton[2:18]: print line # gives us the first sentence

Of Man's first disobedience, and the fruit
Of that forbidden tree whose mortal taste
Brought death into the World, and all our woe,
With loss of Eden, till one greater Man
Restore us, and regain the blissful seat,
Sing, Heavenly Muse, that, on the secret top
Of Oreb, or of Sinai, didst inspire
That shepherd who first taught the chosen seed
In the beginning how the heavens and earth
Rose out of Chaos: or, if Sion hill
Delight thee more, and Siloa's brook that flowed
Fast by the oracle of God, I thence
Invoke thy aid to my adventurous song,
That with no middle flight intends to soar
Above th' Aonian mount, while it pursues
Things unattempted yet in prose or rhyme.
```

Listing 2 NLTK's `WhitespaceTokenizer()` captures 3 break tokens in lines 9 to 11 of the Gutenberg eText

```
>>> for line in book1[8:11]: print line # Python lists start at 0

['In', 'the', 'beginning', 'how', 'the', 'heavens', 'and', 'earth']
['Rose', 'out', 'of', 'Chaos:', 'or,', 'if', 'Sion', 'hill']
['Delight', 'thee', 'more,', 'and', "Siloa's", 'brook', 'that', 'flowed']
```

Listing 3 Comparative outputs (**in bold**) from two different approaches to tokenization for the Gutenberg eText

```
import nltk, re
from nltk.tokenize import * # import all Tokenizer Classes from tokenize package

tokenizer = LineTokenizer(blanklines='discard') # initialize LINE tokenizer

# Read in & tokenize lines in one step
milton = tokenizer.tokenize(nltk.corpus.gutenberg.raw('milton-paradise.txt'))

book1 = milton[2:800] # start and end LINE indexes for Book I of the poem

# Tokenize on whitespace
white = WhitespaceTokenizer()
test = [white.tokenize(index) for index in book1]

# INSTANTIATE A CONTAINER AND APPLY A REGULAR EXPRESSION TOKENIZER TO CAPTURE WORD TOKENS
AND PUNCTUATION TOKENS, PRESERVING WORD-INTERNAL PUNCTUATION SUCH AS HYPHENATED FORMS
'sea-monster'

paradise = [] # becomes a deeply nested array

for line in book1:
       paradise.append(re.findall(r"\w+(?:[-']\w+)*|[-.]+|\S\w*", line))
```

```
# OUTPUTS
>>> for line in test[26:28]: print line

['Say', 'first--for', 'Heaven', 'hides', 'nothing', 'from', 'thy', 'view,']
['Nor', 'the', 'deep', 'tract', 'of', 'Hell--say', 'first', 'what', 'cause']
```

```
>>> for line in paradise[26:28]: print line

['Say', 'first', '--', 'for', 'Heaven', 'hides', 'nothing', 'from', 'thy',
'view', ',']
['Nor', 'the', 'deep', 'tract', 'of', 'Hell', '--', 'say', 'first', 'what',
'cause']
```

normal punctuation and effectively combats problems arising from house style punctuation, as in these pauses in lines 27–28 of the Gutenberg variant:

> Say first–for Heaven hides nothing from thy view,
> Nor the deep tract of Hell–say first what cause

Outputs from both the `WhitespaceTokenizer()` (labelled `test`) and the regular expression tokenizer (labelled `paradise`) are juxtaposed in Listing 3, where the existing data structure for `book1` undergoes further nesting to tokenize individual elements within each line.

As stated, the authors have used the customised verse tokenizer for the count (Section 5). Turning now to the principal dataset, Listing 4 operates on Dartmouth's eText and sorts all word tokens into different bags for breaks and non-breaks via a series

Listing 4 Collecting and sorting all word tokens in Dartmouth College's eText of Book I of *Paradise Lost* into five different bags: (1) all line terminals; (2) end-stopped terminals; (3) run-on terminals; (4) marked caesuras; (5) non-breaks.

```
import nltk, re, copy
from nltk.tokenize import *

tokenizer = LineTokenizer(blanklines='discard') # initialize LINE tokenizer

# Dartmouth College version of Book 1 of Paradise Lost, 1674 edition
milton = open('...dartmouth_1674.txt', 'rU').read() # read in PL Book 1 as a string

book1 = tokenizer.tokenize(milton)

paradise = []

# INSTANTIATE A CONTAINER AND APPLY A REGULAR EXPRESSION TOKENIZER TO CAPTURE WORD TOKENS
AND PUNCTUATION TOKENS, PRESERVING WORD-INTERNAL PUNCTUATION SUCH AS HYPHENATED FORMS
'sea-monster'

for line in book1:
       paradise.append(re.findall(r"\w+(?:[-']\w+)*|[-.]+|\S\w*", line))

# (i) CAPTURE THE LAST 2 TOKENS, WHICH COULD BE WORD + PUNCT OR ELSE 2 WORDS
ends = [index[-2:] for index in paradise]

ends_punct = [] # initialises container for end-stopped line-terminal word tokens
ends_nonpunct = [] # initialises container for run-on line terminal word tokens

for index, item in enumerate(ends):
    if '.' in item[-1]: # if the line terminates with punctuation...
        ends_punct.append((index, item[-2])) #...append previous word token
    elif ',' in item[-1]: ends_punct.append((index, item[-2]))
    elif ';' in item[-1]: ends_punct.append((index, item[-2]))
    elif ':' in item[-1]: ends_punct.append((index, item[-2]))
    elif '?' in item[-1]: ends_punct.append((index, item[-2]))
    elif '!' in item[-1]: ends_punct.append((index, item[-2]))
    elif ')' in item[-1]: ends_punct.append((index, item[-2]))
    elif '--' in item[-1]: ends_punct.append ((index, item[-2]))
    elif '"' in item[-1]: ends_punct.append((index, item[-2]))
    else: ends_nonpunct.append((index, item[-1])) # append terminal word token

# (ii) REMOVE LINE TERMINAL WORD AND PUNCTUATION TOKENS
minus_ends = []
for index, item in enumerate(paradise):
    if '.' in item[-1]: minus_ends.append((index, item[:-2]))
    elif ',' in item[-1]: minus_ends.append((index, item[:-2]))
    elif ';' in item[-1]: minus_ends.append((index, item[:-2]))
    elif ':' in item[-1]: minus_ends.append((index, item[:-2]))
    elif '?' in item[-1]: minus_ends.append((index, item[:-2]))
    elif '!' in item[-1]: minus_ends.append((index, item[:-2]))
    elif ')' in item[-1]: minus_ends.append((index, item[:-2]))
    elif '--' in item[-1]: minus_ends.append((index, item[:-2]))
    elif '"' in item[-1]: minus_ends.append((index, item[:-2]))
    else: minus_ends.append((index, item[:-1]))

# (iii) CAPTURE MEDIALS & BAG REMAINING WORD TOKENS AS NON-BREAKS
minus_ends2 = copy.deepcopy(minus_ends) # changes to copy won't affect original
medials = []# initialises container for word tokens marked as caesuras
non_breaks = []# initialises container for remaining non-break word tokens

for index, item in minus_ends2:
    for i, v in enumerate(item):
        if v in [',', '.', ')', '"', '!', '?', ':', ';', '--']:
            medials.append((i, item[i - 1])) # append token prior to punctuation
            del item[i - 1] # remove medial break token from line in minus_ends2

for index, item in minus_ends2:
```

Table 2 Utterances which are *virtually* prosodically identical in the two datasets have different word counts

Gutenberg 311	And broken chariot-wheels. So thick bestrown,	6 word tokens
Gutenberg 340	Waved round the coast, up-called a pitchy cloud	8 word tokens
Gutenberg 460	In his own temple, on the grunsel-edge,	7 word tokens
Dartmouth 311	And broken Chariot Wheels, so thick bestrown	7 word tokens
Dartmouth 340	Wav'd round the Coast, up call'd a pitchy cloud	9 word tokens
Dartmouth 460	In his own Temple, on the grunsel edge,	8 word tokens
Gutenberg 223	"...In billows, leave **i' th'** midst a horrid vale..."	9 word tokens; 10 syllables intended
Dartmouth 223	"...In billows, leave **i'th'** midst a horrid Vale..."	8 word tokens; exactly 10 syllables

Table 3 Comparative representations of SAM-PA and DISC phonetic transcriptions for diphthongs in Received Pronunciation in English

Diphthong	SAMPA	DISC	Example	Example DISC Transcription
✓	/eI/	1	day	/ d1 /
✓	/aI/	2	night	/ n2t /
✓	/OI/	4	boy	/ b4 /
✓	/@U/	5	no	/ n5 /
✓	/aU/	6	now	/ n6 /
✓	/I@/	7	here	/ h7 /
✓	/e@/	8	there	/ D8 /
✓	/U@/	9	sure	/ S9 /

of steps: (1) all 798 line terminal tokens are collected in ends and then subdivided on presence or absence of attendant punctuation (the containers ends_punct and ends_nonpunct); (2) the container minus_ends is then created where line terminal word and punctuation tokens have been removed; (3) a for loop captures medial breaks in minus_ends and then excludes them from consideration before the final iteration bags remaining tokens as non-breaks, ignoring punctuation tokens.

The counts presented in Section 5 of this article are the true counts for (1) this particular version of the corpus and (2) this particular solution for tokenizing blank verse. Even though we are ostensibly working with the same *poem* in the Dartmouth and Gutenberg eTexts, we are not working with the same *text*—or dataset—and subtle differences do emerge which affect the overall counts (but *not* the experimental outcome) for each version. One of the frequent culprits here is hyphenated forms: there are more of them in the Gutenberg version, hence reducing the word count for this dataset (*cf*. unshaded rows in Table 2). Another occasional difference is the representation of elisions—although it is unlikely that a modern reader familiar with the rhythms of blank verse would let such differences spoil the beat (*cf*. shaded rows in Table 2).

4 Projecting Prosody onto Text via ProPOSEL

ProPOSEL is a prosody and part-of-speech English lexicon of 104,049 word forms, where each entry is mapped to a series of fields holding phonetic, syntactic and prosodic information about that word form. Fields of immediate interest to this study are (1) and (13): the headwords and DISC syllabified phonetic transcriptions which, unlike the more familiar International Phonetic Alphabet (IPA) and SAM-PA, use a single character to represent each phonological segment, irrespective of its complexity. Table 3 illustrates the distinctive symbolic equivalents for diphthongs in DISC which are so easy to spot.

Table 4 Prosodic and syntactic annotations acquired via intersection with ProPOSEL for the first six end-stopped line terminals in Dartmouth College's eText of Book I of *Paradise Lost*, where tokens bearing complex vowels appear in bold

```
[['woe', ['1','s', 'C', "'w5", "'w5:1"]],
['seat', ['1','s', 'C', "'sit", "'sit:1"]],
['seed', ['1','s', 'C', "'sid", "'sid:1"]],
['song', ['1','s', 'C', "'sQN", "'sQN:1"]],
[`rhime', 'rhyme', ['1','s', 'C', "'r2m", "'r2m:1"]],
['pure', ['1','s', 'C', "'pj9R", "'pj9R:1"]],...]
```

ProPOSEL was originally designed for the target application of phrase break prediction, for compatibility with Python and NLTK, and for linkage with speech corpora. Projecting *a priori* linguistic knowledge from this lexicon onto corpus text is accomplished automatically. The Python programming language has a dictionary mapping object with entries in the form of (key, value) pairs, and this syntax is exploited by transforming ProPOSEL into a Python dictionary, with headwords (the *keys* in this case) mapped to an array of values from selected fields. During lookup, word tokens in the corpus acquire values associated with matching dictionary keys. In this way, the contents of each bag created in Listing 4 have now been tagged with prosodic annotations for further analysis. Table 4 shows the first six line terminal breaks in `ends_punct` after intersection with an instance of ProPOSEL holding the following symbolic values: syllable count (field 7); lexical stress pattern[2] (field 8); content-function word tag (field 10); DISC transcription (field 13); stressed and unstressed values mapped to DISC syllable transcriptions (field 14).

4.1 What about the Great Vowel Shift?

In this study, present day pronunciation is projected onto EModE text—which is normally what contemporary readers/speakers do anyway with literary classics of this period—and findings are based on canonical forms: the eight diphthongs, plus the triphthongs, of Received Pronunciation (Roach, 2000, pp. 21–4), or standard English speakers' '*wacky vowels*' (BBC, 2009). Some of these sound patterns would not have been used in Milton's day. The so-called Great Vowel Shift was a sound change over a prolonged period (roughly 1500–1800) that affected long vowels in English, such that their place of articulation shifted upwards: mid vowels were raised and high vowels became diphthongized. The double vowels in *day* and *now* are one outcome of this process; the dialect we now know as *Standard English* ("RP") is another (Giancarlo, 2001).

Barber (1997, pp. 139–40) offers a possible pronunciation for an extract from *To His Coy Mistress* by one of Milton's contemporaries, the poet Andrew Marvell. His transcriptions (presented in their equivalent SAM-PA forms in this section) for diphthongs in the following words: *time*; *coyness*; *down* accord with PresE pronunciations, while transcriptions for *day* and *no* simply indicate long vowels, the latter only becoming diphthongized in the late eighteenth century (ibid, p. 107). An online simulation for EModE (Menzer, 2000) suggests that the diphthong `/aU/` in *loud* would have sounded much the same with an advanced or progressive speaker in 1650 as it does today but that the vowel in *name* was still in flux and reminiscent of the French sound *même*; this agrees with Barber's transcription for *day*: /dE:/. Another online simulation suggests that words like *time/bite* and *now/loud* did contain diphthongs but that these sounded more like hybrids of the combination *but* and *beet* and the combination *but* and *boot* respectively (Rogers, 2000). Similar shifts in phoneme combinations are posited for the diphthong in *boat* and *blow* (Görlach in Giancarlo, 2001).

Phonetic transcriptions in ProPOSEL show English vowels still in flux today and variation in source pronunciation lexica. The SAM-PA and CELEX transcriptions in fields 4 and 13—derived from CUVPlus (Pedler and Mitton, 2002) and CELEX (Baayen *et al.*, 1996)—are in agreement for the following instance of the diphthong `/U@/`

Table 5 Instances of variant syllabification and phonetic transcription for the same orthographic form in pronunciation lexica show English vowels still in flux today

	Word form	Syllable count	SAM-PA	DISC
CUVPlus	fire	1	‘faI@R	
CELEX	fire	2		‘f2-@R
CUVPlus	power	2	‘paU@R	
CELEX	power	2		‘p6-@R

in *pure*; interestingly, the CELEX notation for *pure* (*cf.* Table 4) incorporates a *y-glide* (*cf.* Bridges, 1921, p. 24) and the same goes for the SAM-PA field: / pjU@R /. On the other hand, the word form *moor* is realised with a diphthong in one source lexicon (CUVPlus) and a monophthong in the other (CELEX): / mU@R / versus / 'm$R /; the same speaker may also happily switch from one variant to the other. Finally, triphthongs are particularly unstable, as shown in the mismatches in syllabification in Table 5: CELEX does not appear to use any triphthongs and therefore *fire* and *power* are bi-syllabic; the CUVPlus transcription for *fire* may be interpreted as a triphthong, given the syllable count, but on the same basis, the transcription for *power* may not. This variance, even in canonical forms, is part of our language today.

When we read Milton's verse, we are not, for example, put off by unfamiliar spellings (*thir; o're; highth; strait; dores; foulds; suttle; yeilded*) but render these on-the-fly as their modern equivalents. Similarly, when reading any text, we are cognisant of template (canonical) pronunciations which underlie regional and speaker variation. Temporal variation (EModE versus PresE) might affect the pronunciation of *highth* and *foulds* in the following sentence from Book 1 (lines 722–730) for example, but does not detract from our central insight: the poet uses vowel duration as an explicit (when accompanied by punctuation) and implicit phrasing and highlighting device; and we can simulate this stratagem in phrasing models for English. Complex vowels are a finite set and easily identified; we therefore start with these.

> ‘. . .Th' ascending **pile**
> Stood fixt her stately **highth,** and strait the dores
> Op'ning thir brazen **foulds** discover **wide**
> Within, her ample **spaces,** o're the smooth
> And level **pavement:** from the arched roof
> Pendant by suttle Magic many a **row**
> Of Starry Lamps and blazing Cressets fed
> With Naphtha and Asphaltus yeilded **light**
> As from a **sky**. . .'

In response to Milton's verse here, we note in particular the revelatory effect of an outgliding diphthong, /aI/, at the line turns in 724 and 729: giant metal doors slowly *wide*ning, to reflect and reveal eerie, dazzling *lights* as an influx of hope for the awestruck devils, like the numinous heavenly *light* eternally lost.

5 Significance Testing: the Correlation of Complex Vowels with Boundaries

The motivation for this article has been to explore, formally, the degree of correlation between two variables which apparently co-occur: namely, complex vowels and phrase breaks. Noticing the presence of diphthongs and triphthongs at rhythmic junctures in poetry, and hypothesizing that (i) complex vowels may constitute a phrasing device habitual to native English speakers *because* exemplified by English poets; and (2) that therefore this sound pattern may be used to automatically detect phrase breaks in text (i.e. prose) is our creative insight. This in turn, is partly supported by consensus in the research field of Automatic Speech Recognition over the fact that pauses affect vowel durations in adjoining words (Vergyri *et al.*, 2003). We propose a reverse perspective on prepausal lengthening by interpreting complex vowels as phrase break *signifiers* (Brierley and Atwell, 2009a, 2010).

Both complex vowels and phrase breaks are high frequency events and as a consequence, might often co-occur by chance. Significance testing is therefore used to determine whether high frequency and low variance between these variables is accidental or not. In this case, we apply the chi-squared test because we are dealing with language and cannot assume a

Table 6 Shaded rows provide data for a chi-squared test based on a break count which includes *all* line terminals plus marked caesuras

Queries	Containers	Counts
Number of LINE TERMINAL tokens	ends	798
Number of END-STOPPED lines	ends_punct	266
Number of RUN-ON lines	ends_nonpunct	532
Total number of MEDIAL BREAKS	medials	553
Number of NON-BREAKS which are *not* line-end tokens	non_breaks	4649
Total number of WORD TOKENS	ends + medials + non_breaks	6000
Total for TOKENS with attendant punctuation	ends_punct + medials	819
Total for TOKENS without attendant punctuation	ends_nonpunct + non_breaks	5181
Total number of BREAKS	ends + medials	1351
Total number of NON-BREAKS	non_breaks	4649
Total for unmatched diphthongs + triphthongs after ProPOSEL lookup	MANUAL INSPECTION OF: ends_punct; ends_nonpunct; medials; non_breaks	294
Total for unmatched diphthong_triphthong BREAKS after ProPOSEL lookup	MANUAL INSPECTION OF: ends_punct; ends_nonpunct; medials	106
Count for GLIDES as BREAKS, excluding unmatched items	ends_punct; ends_nonpunct; medials	419
Count for GLIDES as NON-BREAKS, excluding unmatched items	ends_nonpunct	874
Total count for GLIDES as BREAKS		419 + 106
Total count for GLIDES as NON-BREAKS		874 + 188
Total count for complex vowels		1587

normal distribution, a characteristic of natural language being that the majority of words occur very infrequently (*cf.* Zipf's law).

5.1 Collecting counts

Sections 3 and 4 of this article have described how each word in Book I of *Paradise Lost* has been tokenized; then classified as a break or non-break, depending on the presence or absence of attendant punctuation, and as a further refinement, line-terminal status; and finally tagged with its modern-day phonetic transcription. Correspondence between the pronunciation of complex vowels in Milton's day and ours has also been discussed (Section 4).

Table 6 shows counts for the five different containers in Listing 4: {all line terminals; end-stopped terminals; run-on terminals; marked caesuras; non-breaks}, together with various counts for diphthongs and triphthongs obtained through dictionary lookup. These figures represent final counts after manual inspection and correction of totals for complex vowels due to unmatched items during lookup, where the latter generally comprise: proper nouns (e.g. *Nile; Sinai; Horonaim; Aonian*); and compounds (e.g. *sound-board; love-tale; straw-built; dove-like; night-founder'd*), in addition to archaic words and forms (e.g. *compeer; scape; know'st; ere-while; extreams; battel; choyce*).

5.2 Applying the chi-squared test for independence

Based on figures from the shaded rows in Table 6 and entered in **bold** in Table 7, it is now possible to assign each word in the sample to one of four different categories and to compute and enter totals for each category in a 2 × 2 contingency table (*cf.* Table 7) ready for the chi-square test. The category label of diphthongs is used here to denote *all* complex vowels.

Table 7 juxtaposes observed and expected frequencies for all four categories obtained from the data in Table 6 and/or calculated from marginal totals in rows and columns for each category. Expected frequencies are given in *italics*; for example, the expected frequency for items in the sample which exhibit the following attribute-value pairings: diphthong {yes}; break {yes} is 357.34 (i.e. 1351/6000 * 1587).

Table 7 Observed and expected frequencies are computed from the **raw counts** obtained in Listing 4

GROUPS	OUTCOMES		TOTALS
	Breaks	Non-breaks	
Diphthongs	**525** *357.34*	**1062** *1229.66*	**1587**
No diphthongs	826 *993.66*	3587 *3419.34*	4413
TOTALS	**1351**	4649	**6000**

To assess the degree of randomness in our data, we first assume that complex vowels and phrase breaks are independent phenomena: such an assumption is known as a *null hypothesis* H_o. If the null hypothesis is true, then the distributions resulting from observed (f_o) and expected frequencies (f_e) in the shaded area in Table 7 will be very similar. If, on the other hand, this differential exceeds some pre-determined decision-level, then we can reject H_o and surmise that the observed distribution is unlikely to have occurred by chance, and that diphthongs and boundaries are not independent of each other.

To compare observed frequencies with frequencies expected for independence, we calculate the chi-squared χ^2 statistic for all squares (i.e. the shaded area) in Table 7, via the following formula, which sums the differences between these frequencies scaled by the theoretical (i.e. expected) values.

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

In this case, the association between groups and outcomes is deemed to be extremely statistically significant[3]: chi squared equals 138, with 1 degrees of freedom, and a two-tailed *P*-value or odds ratio which is less than 0.0001. Since a *P*-value of 0.01 would mean that there is only a one per cent chance of getting the same result if H_o were true, we can be confident that the perceived association of complex vowels and boundaries is not random.

The break count in Table 6 for Dartmouth's eText of the 1674 edition of Book 1 is lower than that of the more heavily punctuated Gutenberg version: 1351–1447, respectively. Nevertheless, the Gutenberg text is a reliable phrasing variant, encapsulating an alternative parsing and phrasing strategy for the reader or speaker. It is of consequence, therefore, that the statistically significant correlation between complex vowels and phrase breaks is corroborated by this dataset. Experimental replication returns a chi-squared statistic of 123, with 1 degree of freedom, and a two-tailed *P*-value of 0.0001.

6 Textual Analysis of Unmarked Conceptual Boundaries in Book I of *Paradise Lost*

This investigation is based on rhythmic junctures in Book I of *Paradise Lost* marked by punctuation and by line ends. That the latter represent conceptual boundaries and reflect *performance structure* (*cf.* Gee and Grosjean, 1983; Abney, 1992) in theatre or for recital is apparent in this directive on verse-speaking from the Royal Shakespeare Company (Hall, 2004, p. 28): '. . .[t]he end of each line is in fact a punctuation often more crucial than the regular punctuation itself'. An alternative view or segmentation of the text is implemented via XML markup in Durusau and O'Donnell (2002); their *sentence view* differs from traditional presentation, which preserves the integrity of each line (*cf.* the tokenization process in Section 3), and instead segments on punctuation, so that chunks often run from one line to the next and sometimes incorporate constituents from more than two lines. In the following example from lines 10–12 of Book I, segments (i.e. strings between <seg></seg> XML tags) reflect punctuation in the Raben[4] version.

Table 8 An instance of verse-sentence divergence in Book 1 of *Paradise Lost* captured via XML mark-up (Durusau and O'Donnell, 2002)

```
...or, </seg><seg> if Sion hill
Delight thee more, </seg><seg> and Siloa's brook that flowed
Fast by the oracle of God, </seg><seg> I thence
Invoke thy aid to my adventurous song, </seg><seg>
```

Table 9 Phrasing in 17th-century editions of *Paradise Lost* is more open-ended

1667 and 1674 editions	Project Gutenberg eText
. . .Him the Almighty Power	'. . .Him the Almighty Power
Hurld headlong flaming from th' Ethereal Skie	Hurled headlong flaming from th' ethereal sky,
With hideous ruine and combustion down	With hideous ruin and combustion, down
To bottomless perdition, there to dwell	To bottomless perdition, there to dwell
In Adamantine Chains and penal Fire,	In adamantine chains and penal fire,
Who durst defie th' Omnipotent to Arms.	Who durst defy th' Omnipotent to arms. . .'

Run-on lines are common in blank verse. Borrowing terminology from Durusau and O'Donnell (ibid.), line terminals which are not end-stopped are members of *overlapping hierarchies*. They represent the logical relation of *intersection* between two different sets within the sentence: the metrical line and the prosodic-syntactic chunk. The token *down*, for example, in '. . .With hideous ruin and combustion down. . .' (*Paradise Lost*, Book I, line 46) is unmarked with punctuation in the original 1667 and 1674 editions of the poem, as well as the Project Gutenberg eText, and exhibits this kind of duality. It is also part of a wider context: the sentence container spanning lines 44—49 (*cf.* Table 9), where the majority of terminals are run-on and carry diphthongs or triphthongs.

While Raben's version faithfully reflects poetic elisions (*th'etheral*; *th'Omnipotent*), it is more prescriptive in its punctuation such that, in sentence view, *down* would be assigned to a different segment from the 17th-century versions. In the latter, the token *down* is highly ambiguous; syntactically, it is probably part of the compound preposition *down to* and attached to the subsequent noun phrase *bottomless perdition*, but the absence of punctuation seems to preserve an almost uncapturable, long-distance syntactic and semantic relationship to the verb *Hurld*, in which case, *down* would be a particle as in: *Satan was hurled down from heaven*. By twice separating *down* from *hurled*, with commas after *sky* and *combustion*, Raben has edited out some poetic effects: *down* as a particle lost in space, as a long sound lamenting the terrible violence of Satan's severance from God.

The Fall—and Milton's depiction of it—is indelible from our imaginations; a recent stunning re-enactment is the opening sequence of Peter Jackson's film adaptation of *The Two Towers* (2002) and that long shot of the Balrog falling flaming from the bridge of Khazad-Dum into the pit of Moria. Images of falling abound in Book I. There is the famous Mulciber passage where again, a gathering of complex vowels and long vowels in the hinterland between lines delays the verse movement as we witness the protagonist's fall from grace—a beautiful slow-motion arc:

> '. . .**thrown** by angry **Jove**
> **Sheer o'er** the crystal battlements: from **morn**
> To **noon** he fell, from **noon** to **dewy eve**,
> A summer's **day**. . .'

6.1 Caesuras as conceptual boundaries

The boundary concept in verse may be extended to caesuras or rhythmic junctures within the line (*cf.* Introduction), though it has not been possible to include all candidates in the boundary count for the present study because the location of *unmarked*

Table 10 Again, phrasing in the most popular 17th-century edition of *Paradise Lost* is less directive than a contemporary edition

1674 version	1992 version
And chiefly Thou O Spirit, that dost prefer	And chiefly thou, O Spirit, that dost prefer
Before all Temples th' upright heart and pure,	Before all temples th' upright heart and pure,
Instruct me, for Thou know'st; Thou from the first	Instruct me, for thou know'st; thou from the first
Wast present, and with mighty wings outspread	Wast present, and, with mighty wings outspread,
Dove-like satst brooding on the vast Abyss	Dove-like sat'st brooding on the vast Abyss,
And mad'st it pregnant: What in me is dark	And mad'st it pregnant: what in me is dark
Illumine, what is low raise and support;	Illumine, what is low raise and support;
That to the highth of this great Argument	That, to the height of this great argument,
I may assert Eternal Providence,	I may assert Eternal Providence,
And justifie the wayes of God to men.	And justify the ways of God to men.

caesuras is open to interpretation and we have no agreed gold standard to work from. Nevertheless, complex vowels may signal optimal phrase break opportunities within the line, especially when enjambement encourages the reader or speaker to process phrases like '. . .I thence / Invoke thy aid to my adventurous song. . .' as one chunk (*cf.* XML segmentation in Table 8). Would chunking or pausing somewhere within this phrase enhance a reader's or listener's understanding? If so, where *is* the best place to pause? Is it after *thence* or is it after the diphthong-bearing *aid*?

One final extract (*cf.* lines 17–26 in Table 10) from Book I of *Paradise Lost* may serve to highlight how complex vowels signify conceptual boundaries which are pivotal to the parsing strategy for that sentence; and how the correlation of complex vowels and boundaries seems, in fact, to fit Saussure's model of the sign: '. . .[a] linguistic sign is not a link between a thing and a name, but between a concept [signified] and a sound pattern [signifier]. . .' (Saussure in Chandler, 2002, p. 18). Diphthongs act as *precursors* or signifiers of phrase breaks.

Punctuation is again more subtle in the 17th-century version and assumes a poetic sensibility and a poetic ear. In the original, for example, *Dove-like* belongs both to *outspread* and to *sat'st*, whereas the modern edition eliminates one of these paths. Moreover, a comma after *Abyss* at the end of line 21 is perhaps redundant because we cannot produce a succession of sibilants {sat'st; vast; Abyss; mad'st} without slowing down. The section of interest, however, is lines 22–23, where both versions agree.

Assuming that punctuation represents the poet's phrasing, we are meant to pause at the comma in: '. . .What in me is dark / **Illumine,** what is low raise and support. . .' Nevertheless, despite the status of *dark* as a run-on line terminal, and despite its proximity to the marked boundary in *Illumine*, the syntax requires a break at this point; the bigram <adjective><verb> is unusual and the line-break alerts us to this fact. In the subsequent clause, we have a repetition of this uncommon template but instead of a line-break, we have two consecutive diphthongs: **low raise** inhibiting normal phonotactics. A *gold standard* phrasing of this section is hypothesized as follows: '. . .What in me is dark | Illumine, | what is low | raise and support; |. . .' Thus adjacency of complex vowels has been interpreted as a textual cue or *text-based feature* in a difficult syntactic context and in the absence of explicit permission to pause.

7 Conclusions and Further Work

This study uses punctuation, as in previous work on pause patterns in English verse, plus line endings, as equivalents for gold standard phrase break annotations and discovers a significant correlation between complex vowels (i.e. diphthongs and triphthongs) and prosodic-syntactic boundaries, a result which is replicated in *two* naturalistic phrasing variants of the same poem. We believe this finding has several implications. First, complex vowels (like punctuation itself) constitute a domain-independent phrase break feature. Thus, what works for verse may also work for prose; and the authors have already reported on similar findings in parallel experiments for PresE using an extract from the Aix-MARSEC[5] dataset (Brierley and Atwell, 2009a, 2010). Second, while punctuation is a top-performing phrase break feature, it does not capture all perceived prosodic-syntactic boundaries. The use of additional run-on line endings as conceptual boundaries substantiates our findings and also highlights the ambiguous status of some phrase break tokens as constituents of more than one syntactic grouping, where the groups are not always immediately adjacent, even in plain text view. The article also considers complex vowels as boundary precursors, as textual cues signifying optimal parsing and phrasing strategies, and enhancing understanding, for readers and speakers alike. Finally, the *prosodical devices* used deliberately or subconsciously by poets (Milton did say his verse was *unpremeditated*[6]) may provide generic insights into prosodic-syntactic chunking. Banks (Section 2) detects accented and deaccented stops which can be parameterised for experiments with PresE speech corpora (*cf.* Aix-MARSEC); and he leaves us with an intriguing observation: that *units of thought* are *rhythmical*.

References

Abney, S. (1991). Parsing by chunks. In Berwick, R.C., Abney, S., and Carol, T. (eds), *Principle-Based Parsing: Computation and Psycholinguistics*. Dordrecht: Kluwer Academic Publishers, pp. 257–78.

Abney, S. (1992). Prosodic structure, performance structure and phrase structure. In *Proceedings, Speech and Natural Language Workshop*. San Mateo, CA: Morgan Kaufmann Publishers, pp. 425–8.

Auran, C., Bouzon, C., and Hirst, D. (2004). The Aix-MARSEC Project: an evolutive database of spoken British English. In *Proceedings of Speech Prosody (SP 2004)*, Nara, Japan, March 2004, pp. 561–4.

Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1996). *CELEX-2*. Linguistic Data Consortium: Philadelphia.

Banks, T.H., Jr (1927). A study of the relation of the full stops to the rhythm of paradise lost. In *Proceedings of the Modern Languages Association*, Vol. 42.1. New York: Modern Language Association, pp. 140–5.

Barber, C. (1997). *Early Modern English*. Edinburgh: Edinburgh University Press.

BBC. (2009). *The Great Vowel Shift.* (Creative content for online encyclopedia from h2g2 researcher community.) http://www.bbc.co.uk/dna/h2g2/classic/A964578 (accessed 13 November 2009).

Bird, S., Klein, E., and Loper, E. (2009). *NLTK: Natural Language ToolKit* version 0.9.7 http://www.nltk.org/Home (accessed 2 February 2009).

Bridges, R. (1921). *Milton's Prosody: with a Chapter on Accentual Verse and Notes by Robert Bridges*. Oxford: Oxford University Press.

Brierley, C. and Atwell, E. (2008a). ProPOSEL: a prosody and POS english lexicon for language engineering. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008).* Marrakech, Morocco, May 2008. ELRA, pp. 2849–53.

Brierley, C. and Atwell, E. (2008b). A human-oriented Prosody and PoS english lexicon for machine learning and NLP. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), Workshop on Cognitive Aspects of the Lexicon.* Manchester, UK, August 2008, pp. 25–31.

Brierley, C. and Atwell, E. (2009a). Exploring complex vowels as phrase break correlates in a corpus of english speech with ProPOSEL, a prosody and POS english lexicon. In *Proceedings of InterSpeech 2009.* Brighton, UK, September 2009, pp. 868–71.

Brierley, C. and Atwell, E. (2009b). Exploring imagery in literary corpora with the natural language ToolKit. *Proceedings of Corpus Linguistics 2009.* Liverpool, UK, July 2009. http://ucrel.lancs.ac.uk/publications/cl2009/ (accessed 7 April 2010).

Brierley, C. and Atwell, E. (2010). Complex vowels as boundary correlates in a multi-speaker corpus of spontaneous English speech. In *Proceedings of Speech Prosody 2010.* Chicago, Illinois, US, May 2010, in press.

Chandler, D. (2002). *Semiotics: The Basics.* London: Routledge.

Durusau, P. and O'Donnell, M. (2002). Concurrent Markup for XML Documents. Presentation at XML Europe 2002. http://www.idealliance.org/papers/xmle02/dx_xmle02/papers/03-03-07/03-03-07.html (accessed 2 February 2009).

Fodor, J. D. (2002). Psycholinguistics cannot escape prosody. In *Proceedings of Speech Prosody (SP-2002),* Aix-en-Provence, France, April 2002, pp. 83–90.

Gee, J. P. and Grosjean, F. (1983). Performance structures: a psycholinguistic and linguistic appraisal. *Cognitive Psychology*, **15**: 411–58.

Geraghty, J. (2003). *Digital Facsimile Project* http://www.johngeraghty.com/Literature/Texts/Milton/P_Lost_1674/PL_6.jpg (accessed 2 February 2009).

Giancarlo, M. (2001). The rise and fall of the great vowel shift? The changing ideological intersections of philology, historical linguistics, and literary history. *Representations*, **76**(Autumn): 27–60.

Hall, P. (2004). *Shakespeare's Advice to the Players*. London: Oberon Books Ltd.

Ingulfsen, T., Burrows, T., and Buchholz, S. (2005). Influence of syntax on prosodic boundary prediction. In *Proceedings of INTERSPEECH 2005.* Lisbon, Portugal, April 2005, pp. 1817–20.

Jackson, M. P. (2002). Pause patterns in Shakespeare's verse: canon and chronology. *Literary and Linguistic Computing*, **17**(1): 37–46.

Langworthy, C. A. (1931). A verse-sentence analysis of Shakespeare's plays. In *Proceedings of the Modern Languages Association*, Vol. 46.3. New York: Modern Language Association, pp. 738–51.

Luxon, T. H. (ed.) *The Milton Reading Room*. http://www.dartmouth.edu/~milton (accessed 29 October 2009).

Maidment, J. (2009). *The Speech Internet Dictionary*. http://www.phon.ucl.ac.uk/home/johnm/sid/sidhome.htm (accessed 5 November 2009).

Menzer, M. J. (2000). *The Great Vowel Shift* http://facweb.furman.edu/~mmenzer/gvs/index.htm (accessed 2 February 2009).

Oras, A. (1960). *Pause Patterns in Elizabethan and Jacobean Drama*. Gainesville, FL: University of Florida Press.

Pedler, J. and Mitton, R. (2002). *CUVPlus* [Electronic Resource] Oxford Text Archive. http://ota.ahds.ac.uk/textinfo/2469.html (accessed 21 June 2007).

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

Rogers, W. E. (2000). *The History of English Phonemes* http://facweb.furman.edu/~wrogers/phonemes/ (accessed 2 February 2009).

Sinclair, J. M. and Mauranen, A. (2006). *Linear Unit Grammar: Integrating Speech and Writing*. Amsterdam: John Benjamins Publishing Company.

Vergyri, D., Stolcke, A., Gadde, V. R. R., Ferrer, L., and Shriberg, E. (2003). Prosodic knowledge sources for automatic speech recognition. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP 2003)*, Hong Kong, April 2003, pp. 208–11.

Notes

[1] Jackson (2002) observes that '. . .agents of transmission may prefer heavy or light punctuation, [but] tend not to diverge too markedly in where they place the stops. . .' (Section 2).

[2] A *lexical stress pattern* is an abstract representation of the rhythmic structure of a word, using the characters 1 and 2 to denote primary and secondary stress, and 0 for unstressed or weakly stressed syllables.

[3] Using data from Table 7, we have calculated the chi-squared statistic thus: $\chi^2 = \mathtt{SUM(((525-357.34)^2/357.34)+((1062-1229.66)^2/1229.66)+((826-993.66)^2/993.66)+((3587-3419.34)^2/3419.34))}$.

[4] The eText used in Project Gutenberg's Paradise Lost was originally created by Dr. Joseph Raben of Queen's College, NY circa 1964–5.

[5] The Aix-MARSEC Corpus comprises circa 55,000 words of English speech transcribed from BBC radio recordings from the 1980s (Auran *et al.*, 2004).

[6] *Paradise Lost*, Book 9, line 24.