VARD versus Word

A comparison of the UCREL variant detector and modern spell checkers on English historical corpora

Paul Rayson[1], Dawn Archer[2], Nicholas Smith[1]

Computing Dept., Lancaster University, UK[1]
Department of Humanities, University of Central Lancashire, UK[2]
{p.rayson; n.i.smith}@lancaster.ac.uk, dearcher@uclan.ac.uk

Abstract

Analysis of English historical texts poses a number of obstacles for standard corpus analysis and annotation techniques. In addition to non-standard spellings and contractions, there are difficulties at the morphological, phonetic and syntactic levels. Our response has been to develop a VARiant Detector (VARD). We trained VARD on 16th-19th century data, specifically, the *Nameless Shakespeare* and a selection of texts taken from Chadwyck-Healey's *Eighteenth and Nineteenth Century Fiction* collection. We have chosen to explore data from these centuries as, even though variant usage remains an issue up to the present day (because of the use of dialectal forms/ongoing standardisation), it falls substantially in the 18th-19th centuries. This paper reports on experiments to test the utility of VARD. The experiments compared VARD's performance on unseen data with that of spell checkers for modern English (MS-Word and Aspell). Our hypothesis is that, as these spell checkers are not intended to work on historical data, VARD will be superior at both recognising variants and suggesting modern forms. VARD includes modern equivalents via an XML <reg> tag rather than removing the original variants.

1. Introduction

Spelling in the Middle English period to the 18th century tended to reflect local dialect features, and, as such, could differ from one region to another. Some of these differences were related to differences in pronunciation, whilst others were related to different habits of spelling (Graddol et al 1996: 73). According to McIntosh (1969), variation in spelling is worthy of study, as it can yield valuable information that can help to identify the source of a manuscript. A study of variation in spelling can also tell us about potential changes over time and across different text types (Archer and Rayson 2004) as well as highlighting the general underlying principles of present-day English spelling (Rollings 2004). Yet, such variation can be problematic to handle (and therefore document), as existing corpus tools tend to treat each variant of a word separately. In simple terms, they do not capture the relationship between variants such as *abadoning*, *abandonyng* and *abandonynge* and their standardised form *abandoning* in any meaningful way. One consequence of this is that the findings that existing software generate (in respect to variants) may not be as robust as one might like.

So how might we capture variation whilst also highlighting the relationship between variants – especially given the fact that, from a computational viewpoint, analysis of English historical texts presents a number of challenges for automated corpus

annotation software (Archer et al 2003)? We report here on the development of a tool that allows for the detection and "normalisation" of variants to the modern form. We opted to normalise variants to their modern equivalent so that annotation software founded on models of modern English would need little or no additional retraining to be applied to diachronic corpora. However, it should be noted that in our system original variants are always retained within the texts, as revealed by the following example taken from *A True Narrative of the Late Design of the Papists … 1679*:

> … be gone on *Monday* <reg m="Saith">says he, Would you have the Money beforehand<reg o="before-hand">? then it may be you will not do it. No, said I, that I do not desire. But will you <reg o="deposite">deposit it in a third hand? Truly, said he, that is very fair, and I doubt not but they will do it …

Initial experiments suggest that additional corpus techniques are found to be more accurate when variants are linked to their normalised forms in this way: see for example Culpeper and Kytö (2005), who have used the VARD to revisit an earlier study (Culpeper and Kytö, 2002) and address the methodological issue of spelling variation when studying lexical bundles.

The main focus of this paper, however, is not the VARD's utility in respect to additional corpus techniques. Rather, we will report on experiments we have undertaken which compare the VARD's performance on unseen data with that of two modern spell checkers, MS-Word and Aspell. We should state at this point that we are aware of the mismatch between the principles embodied in modern spell checkers (such as MS-Word and Aspell) and Early Modern English variant spelling detection. Such a comparison is nevertheless a valid research endeavour, as we want to evaluate the VARD's ability to detect spelling variants. That said, we should also point out that the motivation behind VARD is not spell checking per se, since there was no standardised spelling in Early Modern English to provide the 'correct' spelling. Indeed, our ultimate aim is to develop a system that does not merely offer the user possible "suggestions" for spelling variants (as in the case of MS-Word and Aspell), but *automatically* regularises variants within a text to their modernised forms so that historical corpora become more amenable to further annotation and analysis.

2. Related work

The computational analysis of English spelling has focussed almost exclusively on modern English. Indeed, with the exception of Roger Mitton (1996) who draws on a history of spelling when investigating spelling errors produced by secondary school pupils in 1970, most studies take little or no account of changes to spelling over time. Existing spell checkers for modern English detect spelling errors by comparing each word in a text against a pre-generated word list. This enables detection of *non-word* errors provided the word list has a good coverage of general vocabulary. Fontenelle (2004) describes hybrid methods for building full-form word lists for spell checkers, consisting of a generative morphology step with constraints applied by manual and automatic corpus analysis. By contrast, *real-word* errors are those where a word is not the one intended in the context, but happens to match another correctly spelled word. These real-word errors cannot be detected without examining the contextual environment. Hirst and Budanitsky (2005) attempt to detect such real-word errors by

measuring the semantic distance of words and highlighting those that are semantically unrelated to their context. They also observe that such real-word spelling errors can be introduced by auto-correction mechanisms in word processors which silently replace pre-defined errors while the user is typing. Earlier approaches have built models of part-of-speech bigrams (Atwell and Elliott, 1987) and trigrams (Mays, Damerau and Mercer, 1991) on the hypothesis that unlikely n-gram sequences will indicate real-word spelling errors. More recent work on real-word spelling errors has employed the notion of confusion sets e.g. {then, than} and {weather, whether}. The task is then characterised as 'word disambiguation' between members of the set based on contextual clues (Golding and Schabes, 1996). In contrast, Brill and Moore (2000) present a new error model for noisy channel spelling correction based on generic string to string substitutions and report significant improvement over previous approaches.

Spell checking systems for modern English are greatly aided by the fact that English spelling is now standardised to a large extent. For Early Modern English or varieties whose spelling system is much less standardised - contemporary Jamaican English, for example (Dray, 2004) - the task of detecting/changing spelling variants is considerably more complex. Thus, in the methodology section (3.1), we define our "principles of intervention" as a means of establishing a framework for the systematic evaluation of Early Modern English spell checking. We then provide a description of the VARD tool in section 3.2. Work that focuses on the detection of variants within historical corpus data is rare at present, exceptions being Pilz et al's (2005) research on German (prior to the orthographic reform in 1901), which employs fuzzy dynamic matching techniques to enable user searches of textual databases with variant spellings, and Thomas's (2005) exploration of variants within different electronic scholarly editions of *King Lear*. For 18th century English, Schneider (2002) takes the approach of customising the open source version of Aspell and attempts normalisation without context sensitive rules. As will become clear, our approach involves both the detection and normalisation of spelling variants.

3. Methodology

3.1. Principles of intervention

Determining what needs to be regularised is not as simple as it may seem. Some users will probably want a tool that normalises everything to its modern equivalent. Others (historical and dialectal specialists especially) may take the view that some morphological variation should be left untouched, because of the ongoing use of forms such as *hath*, *sayest*, *thou*, *thine*, *disn't* (= *doesn't*), *gan* (= *gone*), etc. in specialist registers (e.g. religious texts, poetry) and in certain dialects (e.g. Tyneside English). Still others will probably point out that some words in English (e.g. *judg(e)ment*, *connection/xion*, *colonise/-ize*) can legitimately be spelt in a variety of ways. Our intervention policy is to:

(i) Normalise all known variants to their modernised (British English) equivalents.
(ii) Post-process morphological variants that can be used legitimately in specialist/dialectal registers.

(iii) Normalise hyphenated words such as *out-side* and multi-word expressions such as *pallace-yard* that are no longer hyphenated in standard modern English.

(iv) Normalise "open" lexical units (reflexive pronouns, compound adverbs, etc.) to their hyphenated/closed modern equivalent, e.g. *it self* becomes *itself, them selues* becomes *themselves, mee self* becomes *myself, in deede* becomes *indeed*, etc.

(v) Ignore case distinctions (for this experiment at least; but see section 4.3)

Section 4.3 discusses a range of further issues that are routinely encountered in texts from Early Modern English (EModE) or earlier periods.

3.2. Compiling the EModE regularisation list

Our initial explorations of historical English texts focussed on newspapers dating from 1653 to 1654. A list of variant terms was compiled in part by manual inspection, in part by use of a special tag (Z99) in the USAS semantic annotation system that flagged unknown word forms (see Archer et al, 2003). With the help of students from Northwestern University, Chicago, this approach was then extended to other sources, including the *Nameless Shakespeare* and a selection of texts taken from Chadwyck-Healey's *Eighteenth and Nineteenth Century Fiction* collection. The OED and other historical sources were used to verify and extend the list of historical variants. Entries in the list of variants, now totalling 45,805, have been categorised according to whether they are morphological ("reg m"), orthographical ("reg o"), phonological ("reg ph"), fuzzy (that is, belong to multiple categories; "reg f") or problematic (that is, are difficult to categorise precisely; "reg p"). This type of categorisation scheme allows us to identify patterns relating to a particular type of variation (e.g. the use of *(e)s* where one would expect the genitive today, the doubling of the consonants *l* and *n* in mid-position or the use of *z* for *s*), which can then be used to develop fuzzy-matching rules in future extensions of the software.

The VARD consists of several components. The first component incorporates a search and replace script which 'matches' spelling variants within the pre-processed list to their 'normalised' equivalent and replaces them with an SGML 'reg' tag. Thus, 'addes' is replaced by '<reg o= "addes">adds'. We do this so that the original spelling is retrievable (because of it being encoded in the corpus markup).

The second component utilises fuzzy matching techniques and context rules to identify those real-word variants that require some form of contextualisation to be normalised appropriately. Effectively, the component is designed to identify significant (i.e. potentially *problematic*) sequences of text, and apply some specified annotation to that text. In the case of the variant *bee*, for example, we have a rule that treats the variant as a verb (infinitive or base form) when (i) preceded by a general preposition or a modal auxiliary, and (ii) followed by an article or the past tense/past participle form of a lexical verb. The rule for recognising 'bee' as an infinitive looks something like the following (for an explanation of the template rule format we adopt see Fligelstone *et al* 1996):

II {bee} (RR*n) VV*

Notice that we have allowed for the possibility that a number of adverbs (RR*n) may occur before the past tense/past participle form of the lexical verb. We also use a wild card (*) so that the rule will match any of several strings (i.e. any form of adverb and the past tense or the past participle of the lexical verb).

The VARD may stand alone as a text pre-processor or serve as part of a larger corpus annotation or retrieval system. In our research, the VARD is one component within the USAS grammatical and semantic annotation system (Rayson et al, 2004). By regularising historical spellings to a modern standard, more accurate grammatical and semantic tagging is attained. There is, moreover, a cyclical character to the USAS processing stream, since part-of-speech pattern-matching rules are used in turn to resolve ambiguous forms, distinguishing for example the use of *then* for *than* and the inconsistent use of the genitive.

In the experiment reported in this paper, only the first component was used, as the manually created templates remain to be tested over a larger corpus.

3.3 Methodology for evaluation

The primary software against which VARD was compared was Microsoft Word 2002.[1] MS-Word was chosen because it is the most widely used word processing program, and its spell-checking component provides a reasonable baseline against which to measure the performance of VARD. The spell checker in MS-Word is intended for use on modern language data, rather than historical data. It is run interactively and therefore not convenient for large-scale automatic corpus processing. It also does not explicitly mark where in the text modifications have been made. An additional comparison was made against the less widely used Aspell program[2], which is intended to replace Ispell, an interactive spell-checking tool for Unix. Aspell is still in beta-testing, therefore the analysis of Aspell was limited to a smaller range of text samples, and the analysis here provides only a general indication of the performance level of Aspell.

3.4 First Experiment: VARD vs. WORD

The VARD spelling regulariser was run on a series of texts from four register categories in the Lampeter Corpus of English Tracts (Schmied, 1994). To enable cross-register comparison of performance all texts sampled are from the late seventeenth century (between 1666 and 1679). In addition, we used a selection of prose novels from the eighteenth century. Regularisations were marked in place with an XML <reg> tag; the original form of the word was retained as an attribute of the <reg> tag, e.g.

We therefore trusted ourselves to the Mercy of the <reg o="Waves">waives

Microsoft Word's spell-checker was then run interactively on the same texts. In MS-Word, suspect words are presented to the user one at a time, with a candidate list of spelling corrections. Since the candidates are listed in descending order of probability,

[1] MS-Word 2002 is part of Office XP, version 10.4109.3501 (SP-1), copyright Microsoft Corporation. According to the 'About Microsoft Word' information, portions of the International CorrectSpell™ spelling correction system were developed by Lernout & Hauspie Speech Products.
[2] Aspell version 0.50.3, downloaded in April 2005 as part of Cygwin Tools (www.cygwin.com)

we selected the first-choice replacement in every case. (Words which were queried but for which no candidates were suggested were simply skipped.)

To enable direct comparison of the effectiveness of VARD and MS-Word in detecting variants and suggesting standard forms, the output files of the two programs were verticalised then aligned with the GNU sdiff program. Sdiff flags all differences with a pipe symbol, including regularisations made by MS-Word but not VARD. A simple code, indicating the accuracy of an "intervention", was manually assigned to forms flagged as variant and regularised by the respective programs, as follows:

v1 = VARD correct,
v0 = VARD incorrect
w1 = MS-Word correct
w0 = MS-Word incorrect

The scoring system is illustrated in the following text excerpt, in which the VARD output appears on the left and the MS-Word output on the right.

```
(1)
      That                                      That
      the                                       the
      Bankers                                   Bankers
      are                                       are
      not                                       not
      Men                                       Men
      of                                        of
      greater                                   greater
      Abilities                                 Abilities
      nor                                       nor
      acquired                                  acquired
      Parts                                     Parts
      than                                      than
      other                                     other
      Tradsmen                              |   Tradesmen     v0w1
      ,                                         ,
      nor                                       nor
      better                                    better
      instructed                                instructed
      than                                      than
      others                                    others
      to                                        to
      <reg o="imploy">employ                |   employ        v1w1
      greater                                   greater
      Stocks                                    Stocks
      in                                        in
      an                                        an
      <reg o="advantagious">advantageous    |   advantageous  v1w1
      Trade                                     Trade
      ,                                         ,
      &c.                                       &c.
```

Problematic cases (see further, section 4.3) were marked with a question mark – i.e. evaluated as neither "correct" nor "incorrect", even though the VARD sometimes regularised these, e.g. *doth* in example (2).

(2)

Nor	Nor
<reg d="doth">does	**doth v?w?**
the	the
time	time
of	of
emitting	emitting
this	this
Paper	Paper
,	,
favour	favour
less	less
of	of
a	a
Peaceable	Peaceable
intention	intention

3.5 Second Experiment: VARD vs. ASPELL

The same procedure was carried out to spell-check, align and score the outputs of VARD and the ASPELL software. Due to time constraints, and the generally lower success rate of Aspell than MS-Word, the analysis of Aspell was limited to two texts only, Msca1676 and Scia1666. In the evaluation, the following codes were used to rate regularisations made by Aspell:

a1 = Aspell correct
a0 = Aspell incorrect
a? = Aspell unclear evaluation

These are illustrated in extract (3): (VARD output appears on the left, Aspell output on the right).

(3)

they	they	
govern	govern	
more	more	
securely	securely	
,	,	
nay	nay	
,	,	
more	more	
absolutely	absolutely	
than	than	
the	the	
Sultan	Sultan	
<reg d="doth">does	doth v?w?a?	
with	with	
his	his	
Scemitar	Scimitar	v0w1a1
;	;	
<reg o="rendring">rendering	rendering	v1w1a1
small	small	
Territories	Territories	
equivalent	equivalent	
to	to	
Monarchies	Monarchies	
.	.	

4. Results and discussion

The results of the comparison of VARD and MS-Word on the Lampeter corpus texts is presented in Table 1.

Table 1: Accuracy of VARD vs MS-Word in modernising historical spelling variants (percentages refer to the proportion of all variants detected in each text, by both programs)

	Eca1676	Msca1676	Rela1679	Scia1666	Total
VARD correct, MS-Word incorrect	29 27.6%	18 30.0%	52 45.2%	63 23.2%	162 29.4%
VARD incorrect, MS-Word correct	3 2.9%	5 8.3%	7 6.1%	22 8.1%	37 6.7%
Both correct	55 52.4%	27 45.0%	37 32.2%	111 41.0%	230 41.7%
Both incorrect	1 1.0%	0 0.0%	5 4.3%	0 0.0%	6 1.1%
Borderline/problematic cases	17 16.2%	10 16.7%	14 12.2%	75 27.7%	116 21.1%
Total variants detected by either program	105 100.0%	60 100.0%	115 100.0%	271 100.0%	551 100.0%

4.1 Comparison of performance of VARD and MS-Word

Depending on the text type, between a third and a half of variants are correctly modernised by both programs. However, in those cases where only one program successfully modernises a variant (rows 1 and 2 of Table 1), the VARD is much more effective than MS-Word.

In general it seems likely that the greater accuracy of the VARD can be attributed to its hand-crafted regularisation table, founded on careful scrutiny of historical sources (see section 3.2).

A few generalisations can be made about the kinds of historical variants correctly and incorrectly modernised by the respective programs. Variants correctly modernised by both VARD and MS-Word are mostly high and medium-frequency words. It seems also that some of these are liable to be mis-spelt by modern-day writers. Examples include:

> *alwaies* (*always*), *behinde* (*behind*), *coyn* (*coin*), *daies* (*days*), *errours* (*errors*), *fourtieth* (*fortieth*), *intire* (*entire*), *knowes* (*knows*), *onely* (*only*), *publick* (*public*), *severall* (*several*), *suddain* (*sudden*), *surprize* (*surprise*).

Words that VARD alone successfully modernised included:

> *busie* (standard form: *busy*), *daies* (*days*), *deterre* (*deter*), *disturbe* (*disturb*), *scape* (*escape*), *expresse* (*express*), *fewel* (*fuel*), *gon* (*gone*), *lookt* (*looked*), *publikely* (*publicly*), *strangly* (*strangely*), *sute* (*suit*)

MS-Word identified the same items as variants, but did not accurately regularise them, as extracts (4) and (5) illustrate:

(4)

nor	nor
can	can
any	any
Pile	Pile
of	of
wealth	wealth
afford	afford
<reg o="Fewel">fuel	| Feel v1w0a0
for	for
such	such
a	a
Flame	Flame

(5)

And	And
such	such
causes	causes
sought	sought
for	for
,	,
as	as
might	might
best	best
<reg o="sute">suit	| suet v1w0a0
with	with
such	such
a	a
Supposition	Supposition

An important respect in which VARD outperforms MS-Word is in its *non*-interventions. That is, MS-Word frequently flags as variants forms that are either standard English, or foreign language forms that we would not wish to alter. By leaving these forms alone, VARD attains a much higher rate of precision. Common cases in this category include proper names, e.g.:

> *Joab* (MS-Word: *Jab*), *Horrocks* (MS-Word: *Herrick's*), *Baron Capel of Hadham* (MS-Word: *Baron Carpel of Had Ham*)

and passages of Latin and French:[3]

[3] This is not to say, of course, that the frequent phenomenon of codeswitching between English and Latin/French etc. found in Early Modern texts does not pose problems for natural language processing. However, for the present purpose, the important thing is that the text is left intact.

(6)

And		And	
therefore		therefore	
,		,	
if		if	
as		as	
to		to	
my		my	
self		self	
any		any	
thing		thing	
should		should	
humanitus	|	humanities	v1w0a0
accidere	|	acrider	v1w0a0

A small proportion of regularisations – fewer than 10% of variants detected in any one file – were correctly regularised by MS-Word, but not VARD. They included the following:

> *atending* (regularised to: *attending*), *Brittain* (*Britain*), *Scemitar* (*scimitar*), *periodick* (*periodick*), *Saturne* (*Saturn*), *substract* (*subtract*), *thred-bare* (*thread-bare*)

These items are all absent from the VARD regularisation list. We suspect that MS-Word identifies and corrects some of these items (e.g. *periodick*, *atending*) on the basis of algorithms as opposed to simple "search and replace" word patterns.

Cases where *neither* VARD nor MS-Word provide an accurate regularisation are few in number (less than 5% in any file). Many items in this category *cannot* be regularised by a simple search and replace entry, because they match different modern-day forms depending on context. Note the different functions of *then* in (7) and (8):

(7)

If		If	
this		this	
confident		confident	
accusing		accusing	
of		of	
them		them	
,		,	
be		be	
more		more	
then		**then**	
he		he	
<reg h="hath">has	|	hath v?w?	
grounds		grounds	
for		for	

(8)

How	How
then	**then**
in	in
these	these
ways	ways
,	,
is	is
the	the
Integrity	Integrity
and	and
Generosity	Generosity
of	of
a	a
<reg mg="Mans">man's	| Mans
Dealings	Dealings
better	better
discovered	discovered
?	?

Clearly context-sensitive rules are required to disambiguate *then* in such cases. Similarly, a word ending in *-s* may be either a plural noun or the genitive form of a singular noun, as in:

(9) I thought it fittest for your Lordships Patronage …

We are currently developing pattern-matching rules to disambiguate these and other recurrent items that cannot be fixed by fixed string matching.

4.2 Comparison of performance of VARD and Aspell

Table 2 presents the results of the comparison of VARD with Aspell on two of the Lampeter corpus texts.

Table 2: Accuracy of VARD vs Aspell in detecting and regularising historical spelling variants

	Msca1676	Scia1666	Total
VARD correct, Aspell incorrect	12 16.7%	92 33.3%	104 29.9%
VARD incorrect, Aspell correct	4 5.6%	15 5.4%	19 5.5%
Both correct	29 40.3%	75 27.2%	104 29.9%
Both incorrect	3 4.2%	5 1.8%	8 2.3%
Borderline/problematic cases	24 33.3%	89 32.2%	113 32.5%
Total variants detected by either program	72 100.0%	276 100.0%	348 100.0%

Again, VARD outperforms Aspell in detecting and regularising historical spelling variants. The margin of difference in scores is higher than that between the VARD and MS-Word (cf. rows 1 and 2 of Table 1 and Table 2). Aspell's suggestions for

regularisation are less accurate than either the VARD's or MS-Word's. Some commonly occurring variants are accurately identified but inaccurately regularised, e.g.:

> *dayes* (regularised to *dais*; should be *days*), *daies* (regularised to *dyes*; should be *days*), *neer* (regularised to *nee*; should be *near*), *turnes* (regularised to *turners*; should be *turns*), *tydes* (regularised to *tades*; should be *tides*), *waies* (regularised to *Wise*; should be *ways*).

(10)

```
in                                          in
which                                       which
it                                          it
<reg o="turnes">turns                   |   turns v1w1a0
upon                                        upon
its                                         its
own                                         own
Axis                                        Axis
```

Moreover, Aspell more frequently than the other two programs suggests changes where they are not required.[4] This applies not only to proper names and foreign words (which are problematic also for MS-Word), but also to formulae; cf *ABCDE* in example (11):

(11)

```
Now                                         Now
supposing                                   supposing
ABCDE                                   |   ABCDE v1w1a0.a
to                                          to
be                                          be
a                                           a
part                                        part
of                                          of
the                                         the
great                                       great
Orb                                         Orb
of                                          of
the                                         the
Annual                                      Annual
motion                                      motion
```

4.3 Problematic cases

Up to a third of variants detected in any text may be problematic for modernisation (see row 5 of Table 1 and Table 2). Foremost among such items are:

1. the archaic *–eth* and *–(e)st* verb suffixes, e.g. *doth* (see example (2)), *hath*, *hast*, *sayeth*, etc., which persist in specialised contexts: religious and poetic usage
2. the fused form *'Tis* (*It is*)

[4] We suspect that the less accurate performance of Aspell is because it has a smaller lexicon than MS-Word and VARD, and it has had little or no historical training.

3. spellings that are variable even in modern-day usage, e.g. *center*/*centre*, skilful/skillful/skilfull, suffixes in *-or*/*-our*, *-ise*/*-ize*
4. archaic forms like *howbeit*, *betwixt*, for which no obvious modern equivalent exists
5. compound words, e.g. *superadded* (*if nothing else were superadded*), *Newmoon* (*about the Full-moon and Newmoon*), *neaptides* (*As likewise of the Spring tides and Neaptides*), *it self* (*whose thrift is itself a yearly, nay weekly revenue*), *now adays* (*those Pleasure-boats now adays carrying such sail*), *in stead* (*in stead of being burthened*)
6. proper names of Latin origin that are sometimes modernised, e.g. *Tycho* (*Tyco*), *Galilaeo* (*Galileo*)

The variant list within VARD tends to regularise most cases falling within types 1 to 3, while MS-Word and Aspell "correct" most types except 1 and 4. The absence of a clear, *uniform* modern-day practice with respect to types 1 to 6 above suggests that it is appropriate to err on the side of caution – which explains why we have chosen to count all such cases separately in this evaluation exercise. As we have highlighted in this paper, we will be normalising some of the above via fuzzy matching algorithms, and incorporating a post-processing component within VARD that will reintroduce the variant forms, whilst signalling a relationship between the latter and their modernised equivalents. The motivation for our approach is to ensure we can make use of important contextual information (that would have been lost had we not initially normalised them).

5. Conclusion and Future Work

The experiment has confirmed that compilation of the list of historical spelling variants with their modern equivalents, although labour-intensive, has proven worthwhile. The VARD is significantly more accurate than either MS-Word or Aspell in modernising historical variants.

The main value of modern spell-checkers such as MS-Word and Aspell for processing historical English texts is that they flag potential additions to the VARD regularisation table. In this respect they serve a similar function to the Z99 tag assigned to unmatched forms by the USAS annotation software. Even if the spelling suggested by MS-Word and Aspell is incorrect, the marking of the form as a potential historical variant is a significant time-saver for the linguist seeking to extend the coverage of the VARD.

In the USAS corpus annotation software, the VARD list of variants and modern equivalent spellings is an important process of regularisation run in conjunction with other methods, notably context-based rules and fuzzy-matching algorithms (cf. Robertson and Willet 1991, 1993). Our future work will build on all three fronts. It will contribute to making historical English texts more accessible and amenable to the standard corpus linguistic techniques, such as frequency profiles, concordances, collocations and extraction of n-grams.

Issues remain as to where precisely modernisation of historical spellings is necessary or desirable. To maximise consistency in our processing of variants, we are documenting all types of variant that the VARD encounters, and ranking them as

either clearly archaic (in which case we provide a modernised form) or as obsolescent or specialised (in which case we leave them "as is").

Acknowledgements

The work presented in this paper was carried out within two projects: 1. *Unlocking the Word Hoard* funded by the Andrew W. Mellon Foundation with Martin Mueller of Northwestern University and 2. *Scragg Revisited* funded by the British Academy (under the small research grant scheme).

References

Archer, D., McEnery, T., Rayson, P., Hardie, A. (2003). Developing an automated semantic analysis system for Early Modern English. In Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery (eds.) *Proceedings of the Corpus Linguistics 2003 conference*. UCREL technical paper number 16. UCREL, Lancaster University, pp. 22 - 31.

Archer, D. and Rayson, P. (2004) Using an historical semantic tagger as a diagnostic tool for variation in spelling. Presented at *Thirteenth International Conference on English Historical Linguistics (ICEHL 13)* University of Vienna, Austria 23-29 August, 2004.

Atwell, E. and Elliott, S. (1987) Dealing with ill-formed English text. In: Garside, R., Leech, G. and Sampson, G. (eds.) *The Computational Analysis of English: A Corpus-Based Approach*, Longman, London, pp. 120–138.

Brill, E. and Moore, R. C. (2000) An improved error model for noisy channel spelling correction. In *Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics (ACL2000)*, Hong Kong, pp. 286-293.

Culpeper, J. and Kytö, M. (2002). Lexical bundles in Early Modern English dialogues: a window into the speech-related language of the past. In T. Fanego, B. Méndez-Naya, and E. Seoane (eds.) *Sounds, words, texts and change. Selected papers from 11th ICEHL, Santiago de Compostela, 7-11 September 2000*. John Benjamins, Amsterdam, pp. 45-63.

Culpeper, J. and Kytö, M. (2005). Exploring speech-related Early Modern English texts: lexical bundles re-visited. Presented at the *26th conference of ICAME (International Computer Archive of Modern and Medieval English)*, University of Michigan, USA, May 2005.

Dray, S. (2004) *Writes of passage: exploring non-standard texts, writing practices and power in the context of Jamaica*. Ph.D. thesis, Lancaster University, UK.

Fligelstone, S., Rayson, P., and Smith, N. (1996). Template analysis: bridging the gap between grammar and the lexicon. In J. Thomas, and M. Short (eds.), *Using corpora for language research: Studies in the Honour of Geoffrey Leech*. pp 181-207. Longman, London.

Fontenelle, T. (2004). Lexicalisation for Proofing Tools. In Williams G. and Vessier S. (eds.) *Proceedings of the 11th EURALEX (European Association for Lexicography) International Congress (Euralex 2004), Lorient, France, 6-10 July 2004. Université de Bretagne Sud*. Volume I, pp. 79-86.

Golding, A. R. and Schabes, Y. (1996) Combining trigram-based and feature-based methods for context-sensitive spelling correction. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL96)*, Santa Cruz, pp. 71–78.

Graddol, D., Leith, D. and Swann, J. (1996) *English: History, Diversity and Change.* Routledge.

Hirst, G. and Budanitsky, A. (2005) Correction real-word spelling errors by restoring lexical cohesion. *Natural Language Engineering*, 11 (1), pp. 87-111.

McIntosh, A. (1969) 'The analysis of written Middle English'. In Lass, R. (ed.) *Approaches to English Historical Linguistics.* New York: Holt, Rhinehart & Winston.

Mays, E., Damerau, F. J. and Mercer, R. L. (1991) Context based spelling correction. *Information Processing and Management* 27(5), pp. 517–522.

Mitton, R. (1996). *English spelling and the computer*. Longman, London.

Pilz, T., Luther, W., Ammon, U., and Fuhr, N. (2005) Rule based search in text databases with non-standard orthography. *Presented at the 17th joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing (ACH/ALLC 2005)*, Victoria, BC, Canada.

Rayson, P., Archer, D., Piao, S. L., McEnery, T. (2004). The UCREL semantic analysis system. In *proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks in association with 4th International Conference on Language Resources and Evaluation (LREC 2004)*, 25th May 2004, Lisbon, Portugal, pp. 7-12.

Robertson, A. M. and Willet, P. (1991) Digram and trigram matching for the identification of word variants in historical text databases. In McEnery, T (ed). *13th Information Retrieval Colloquium, Lancaster 1991*. Chippenham: Antony Rowe Ltd, pp. 12-21.

Robertson, A. M. and Willet, P. (1993) Evaluation of techniques for the conflation of modern and seventeenth century spelling. In McEnery, T and Chris P (eds.) *14th Information Retrieval Colloquium, Lancaster 1992*. London: Springer-Verlag, pp. 155-168.

Rollings, A. (2004) *The Spelling Patterns of English.* Lincom Europa.

Schmied, J. (1994) The Lampeter Corpus of Early Modern English Tracts. In Kytö, M., Rissanen, M. and Wright, S. (eds.) *Corpora Across the Centuries*. Rodopi, Amsterdam, pp. 81-89.

Schneider, P. (2002) Computer assisted spelling normalization of 18th century English. In P. Peters, P. Collins, and A. Smith (eds.) *New frontiers of corpus research: papers from the 21st International Conference on English Language Research on Computerized Corpora, Sydney 2000*, Rodopi, Amsterdam, pp. 199-211.

Thomas, S. (2005) Finalizing the multiple-text electronic King Lear for use in the classroom. *Presented at the 17th joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing (ACH/ALLC 2005)*, Victoria, BC, Canada.