VARD 2: A tool for dealing with spelling variation in historical corpora

Alistair Baron and Paul Rayson
Lancaster University
{a.baron, paul}@comp.lancs.ac.uk

Abstract

When applying corpus linguistic techniques to historical corpora, the corpus researcher should be cautious about the results obtained. Corpus annotation techniques such as part of speech tagging, trained for modern languages, are particularly vulnerable to inaccuracy due to vocabulary and grammatical shifts in language over time. Basic corpus retrieval techniques such as frequency profiling and concordancing will also be affected, in addition to the more sophisticated techniques such as keywords, n-grams, clusters and lexical bundles which rely on word frequencies for their calculations. In this paper, we highlight these problems with particular focus on Early Modern English corpora. We also present an overview of the VARD tool, our proposed solution to this problem, which facilitates pre-processing of historical corpus data by inserting modern equivalents alongside historical spelling variants. Recent improvements to the VARD tool include the incorporation of techniques used in modern spell checking software.

1. Introduction

Spelling variation causes considerable problems for corpus linguistic techniques such as frequency analysis, concordancing and automatic tagging, with a significant impact being made on recall and the accuracy of results (Rayson et al, 2007). This paper will focus on Early Modern English, the most recent period of the English language to include a large amount of inconsistent spelling. Although many corpora of Early Modern English have been constructed, little research has

been completed to deal with the problem of spelling variation within digitised forms of these texts. With the increasing amount of historical data being digitised through current initiatives, including Google Books and Early English Books Online, it is imperative that techniques are found to aid the search and retrieval within such datasets.

The amount of spelling variation within Early Modern English text is due to many different factors, such as adding and removing letters for the justification of lines and the influence of local dialect, but mainly because there were no standard spelling rules and no notion of the importance of a single spelling to represent each word, with individual scribes, authors, editors and printing houses having their own spelling preferences (Vallins and Scragg, 1965).

This paper presents the development of the Variant Detector (VARD) tool which acts as a pre-processor for text containing spelling variation. The tool uses techniques from modern spell checkers to detect spelling variants and find candidate modern equivalents. The tool can be used both interactively and automatically to process spelling variants found within a text and produce an output with modernized forms alongside the original variants, allowing corpus linguistic tools and methods to be more accurately used with the corpora.

2. Early Modern English

Our research mainly focuses upon spelling variation in Early Modern English (henceforth EModE), the period of the English language between 1450 and 1700 – although there is some debate on the precise dating. The EModE period is of significant importance for the study of the English language due to it being influential in the formation of the standard modern English we use today. The introduction of the printing press by William Caxton in 1476 and an increasingly literate public led to book production increasing sharply during the EModE period (Görlach, 1991: 6), the result of this being that EModE is the earliest period of the English Language from which a reasonably large corpus can be constructed and

subsequently studied in detail. Shakespeare's works were also written within the period, adding to its research value.

The English Language was under significant change during the EModE period; French and Latin were rapidly being replaced by English as the preferred choice of language for print and speech for many institutions and individuals (see Singh, 2005: 140-147), especially due to King Henry V's commitment to the vernacular in his official correspondence in 1417 (Richardson, 1980: 727). Spelling variation was a prominent feature in written English during the EModE period. Individual scribes, authors, editors and printing houses had their own spelling preferences, although spelling variation was not solely different depending on the writer or compositor; it is common to find words spelt in a number of different forms in the same text or even on the same page. Generally, there was no notion of the importance for a single spelling for each word, letters would be added or removed to, for example, ease line justification (Vallins and Scragg, 1965: 71). Another problem was that texts were often written by numerous scribes who would use their own spelling preferences resulting in different spelling conventions from one page to the next. Furthermore, spelling tended to be influenced by the local dialect and so could differ between regions (Rayson et al, 2005), this was especially the case earlier in the EModE period, before the spread of London and Chancery English.

The construction of EModE and other historical corpora has become an important focus of research. There are many historical English corpora that have already been developed or are in the process of being developed, these include the Helsinki, ARCHER, Lampeter and ZEN corpora (detailed in Kytö et al, 1994), the Corpus of Early English Correspondence (Nevalainen, 1997), the Corpus of English Dialogues (Culpeper and Kytö, 1997) and also many different versions of Shakespeare's works (for example, the First Folio as printed in 1623, which can be sourced from the Oxford Text Archive[1]). In addition, increasing amounts of

[1] http://ota.ahds.ac.uk/

textual data, large quantities of which are historical texts, are being digitised through current initiatives including: the Open Content Alliance[2], Google Book Search[3], and Early English Books Online[4].

Many automated corpus linguistic functions exist, including key word analysis, collocations, concordances and annotation, problems occur when these functions are applied to historical varieties or dialects of English (and indeed other languages), especially when large levels of spelling variation occurs – as in Early Modern English. Spelling variation poses problems for even simple functions such as a string search in a concordance, with only words spelt in exactly the same way as the search query being returned. A recent examination of the Lampeter corpus has shown that an average of 1 in 5 word types per text are not found in a large modern word list[5]; therefore relying on modern spellings would not return accurate results all of the time. Frequency lists will also be inaccurate due to a word's potential frequency being split between its different spelling forms; *would* for example could be spelt in a variety of forms including: *would, wolde, woolde, wood, wuld, wulde, wud, wald, vvould, vvold,* and so on. Keyword lists could also be obscured by spelling variation, with a word's frequency being reduced due to multiple spellings in a text or corpus affecting the word's 'keyness'.. This problem is potentially intensified when evaluating key word-clusters as even very low frequency word-clusters could be considered key, but if any one of the words within a particular cluster are spelt in different forms throughout a text or corpus the frequency of that cluster will be reduced. Collocations would also be affected in much the same way, with co-occurring words not being detected due to reduced frequencies.

---

[2] http://www.opencontentalliance.org/

[3] http://books.google.com

[4] http://eebo.chadwyck.com/home

[5] This was an individual study conducted using the Variant Detector tool described in Section 3.

Automatic part-of-speech (POS) tagging of English text is possible by producing methods which use well-defined rules of the language, amongst other techniques. However, these methods are based on modern English and problems are encountered when dealing with variations of the language, e.g. EModE. The CLAWS POS tagger (Garside and Smith, 1997), for example, uses a dictionary which includes words (or multi-word units) and suffixes with their possible parts of speech. This dictionary is based upon modern English and does not include the large amount of spelling variants (as previously discussed) and the archaic / obsolete words found in EModE texts. Rayson et al (2007) evaluated the accuracy of CLAWS on EModE corpora, and found a significant drop in POS tagging accuracy (from 96-97% for standard modern English). Interestingly, dealing with spelling variation improved accuracy:

	POS Tagging Accuracy	
	Shakespeare	Lampeter
Spelling variation remaining	81.94%	88.46%
Spelling variation modernised	88.88%	91.24%

Table 1 - POS Tagging Accuracy on EModE Corpora

Semantic tagging can also be carried out automatically, but like POS tagging, accuracy suffers due to spelling variation. One example of an automatic semantic tagger is USAS (Rayson et al, 2004), again this has been developed for processing modern English. Archer et al (2003) discuss developing USAS for EModE, the paper reports on evaluation performed on relatively contemporary texts from 1640. Dealing in part with spelling variation produced an improvement in error rates: 2.9% to 1.2% in one text and 4.0% to 1.4% in the other text processed.

It should be noted that the accuracy of annotation is likely to be affected by additional factors; there were definite differences in the grammar of present-day English and EModE, Kytö and Voutilainen (1995) discuss this in their paper reporting on applying another POS tagger, the ENGCG Parser, to the previously

mentioned Helsinki Corpus. Another point to consider is the possibility of a semantic shift in words from EModE to present-day English. However, the above results show that dealing with spelling variation can achieve substantial improvements in annotation accuracy.

3. VARD 2

The previous section highlighted the problem that spelling variation causes for automatic corpus linguistic tools when dealing with texts which contain a large amount of spelling variation. Our solution to this problem was to develop a tool which acts as a pre-processor for corpus linguistic tools which 'standardizes' spelling variation found within texts. This led to the production of the VARD (Variant Detector) software which inserted a modern equivalent alongside the original spelling for any variants detected (see Rayson et al, 2005). The processed text can then be passed on to corpus linguistic software such as Wmatrix (Rayson, 2007) and WordSmith Tools (Scott, 2004). It should be noted that the spelling of EModE texts is not being "corrected", there was no "correct" spelling at the time and the spelling variants are important linguistic features. The original variant is retained and it is a simple process to switch between the original and modernised forms. The modern equivalents are inserted for the benefit of the automated software for retrieval and annotation purposes.

The original VARD tool used a large manually created list of variant to modern equivalent mappings in order to search for and replace any spelling variants found within a text. This technique successfully deals with a substantial amount of spelling variation, however due to the extensive variety in spelling variant forms it is impossible to include all possible spelling variants in a pre-defined list, and the list was generated solely to deal with EModE spelling variation, the tool would therefore not be of use when dealing with the spelling variation found in other varieties of English (and other languages). The tool also permitted little user control over whether a variant was replaced, if a word in the text was listed as a variant it would have the modern equivalent listed along with it inserted alongside the word in the text; whilst this may be desirable in some cases, the

user may wish to have more control over which variants are replaced, for example, they may wish certain forms to remain.

VARD 2[6] was developed which employs techniques derived from modern spell checking software to find candidate replacements for spelling variants within a text. This more flexible approach allows the tool to deal with a much larger variety of spelling variants; any word not found in the tool's modern lexicon is marked as a potential variant, a list of candidate modern equivalents ranked by 'confidence' is produced for each potential variant and is presented to the user for consideration. The system can also be instructed to choose the top candidate for each variant providing its 'confidence' score is over a user-defined threshold.

For any potential variant found, the tool uses three methods to search for candidate modern equivalents:

- The manually created list used in the original VARD tool.
- A phonetic matching technique (modified SoundEx) which assigns a phonetic code to each word; any words in the tool's modern lexicon with the same phonetic code as the variant form are listed as candidate modern equivalents.
- A series of letter replacement rules which can be used to transform the spelling variant into a variety of forms, any created forms which equate to a word found in the tool's modern lexicon are listed as candidate modern equivalents. The letter replacement rules represent common patterns of spelling variation, these include the doubling of certain characters, interchanging characters in confusion sets such as {'v', 'u'} and {'i', 'j'}, and the addition and removal of certain letters, e.g. the final 'e'.

[6] Full details on VARD 2 are available at http://www.comp.lancs.ac.uk/~barona/vard2/

The ‘confidence’ score is calculated by summing the weights associated with each method which successfully found the modern candidate, a small amount is subtracted based on the edit distance (Kukich, 1992: 393-395), and a percentage figure is produced to present to the user.

The weights associated with each method are not static and will change each time a method is successful over another method in finding the chosen modern equivalent; this results in the tool ‘learning’ which methods are more appropriate for finding modern equivalents and thus giving a higher ‘confidence’ score to those candidates found with these methods. This capability makes the tool much more flexible when dealing with different varieties of text; training the tool for a particular corpus by processing a sample text first will allow the tool to better find and rank candidate modern equivalents for variants found in the remainder of the corpus.

VARD 2 has two user interfaces available: an interactive version and a batch processing version. The interactive version, shown in Figure 1 below, allows the user greater control over dealing with variants in the text. The full text is available in the main window to examine, with words grouped into four categories: variants (words not found in the tool’s modern lexicon), replaced words (variants which have been dealt with so far), modern forms (words in the tool’s modern lexicon) and uncommon words (words in the tool’s modern lexicon but at a low frequency – based on the British National Corpus (Leech et al, 2001)); and displayed in an alphabetical list in the sidebar on the right-hand side. A user can make their way through a text, right-clicking on any highlighted variant to be presented with the ranked list of candidate replacements. Clicking on an offered replacement changes the variant to the modern form selected, the original form is stored for reference during output or if the user decides to reverse a replacement operation. The process of replacing a variant is shown in Figure 2 below. As can be seen, the tool displays details of how it arrived at the confidence score by indicating which methods were used to find the candidate and giving the edit distance between the candidate and the variant. There is also an option to manually

replace a variant if the correct replacement is not listed by the system. Other options available in the interactive version include the ability to join words separated by white space (e.g. line breaks), undo or redo any edit made and add / remove letter replacement rules.

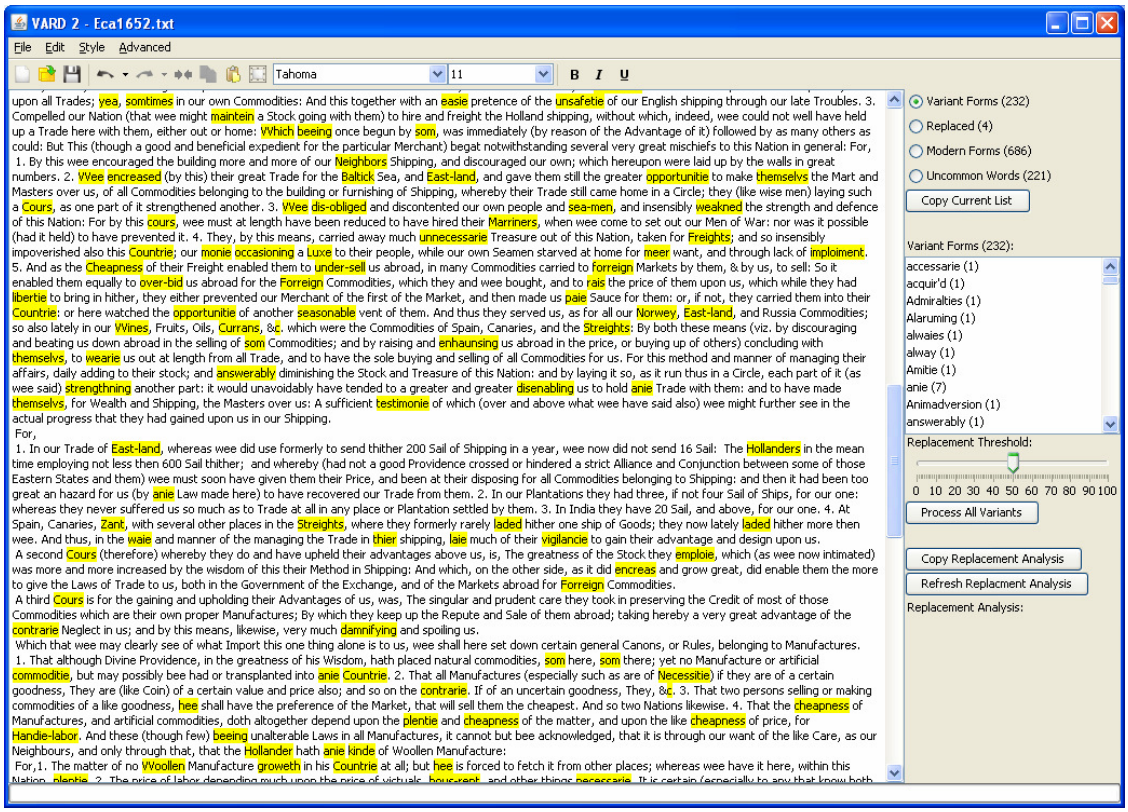

Figure 1 - VARD 2 interactive interface

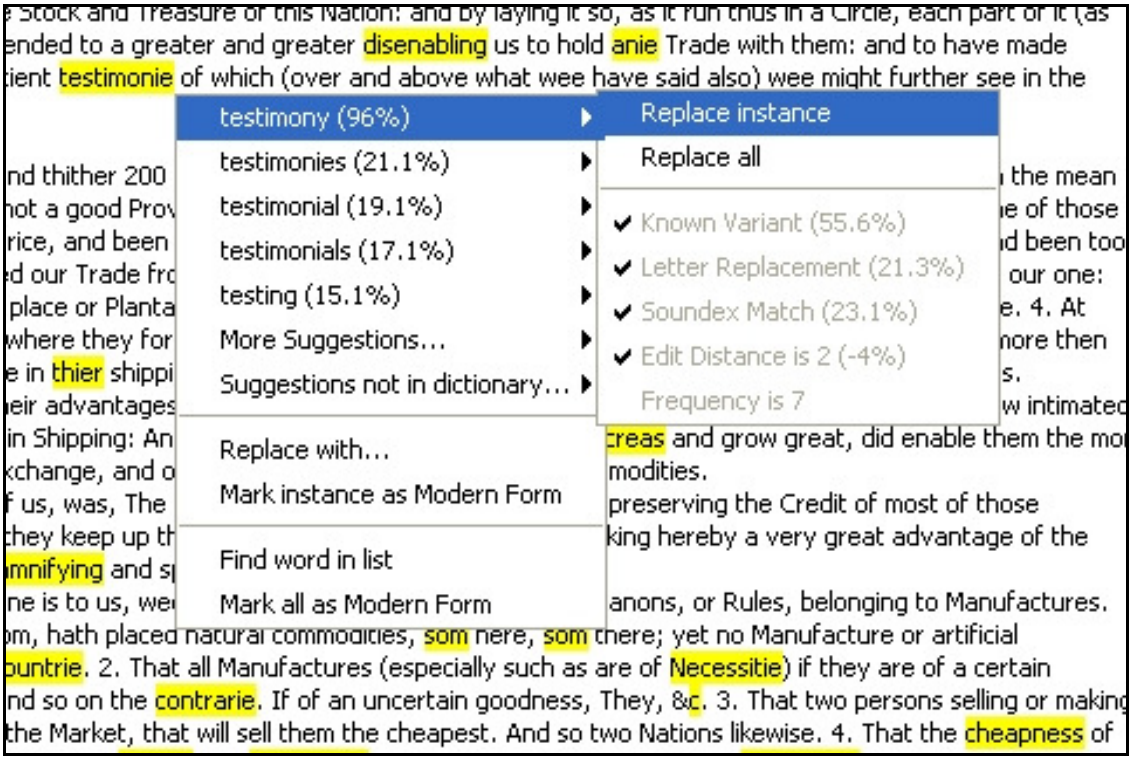

Figure 2 - Replacing a variant in VARD 2

As well as manually dealing with the variant forms, the user can choose to automatically replace all variant forms with their highest ranked candidates, the user can also provide a threshold ‘confidence’, which is the minimum score the candidate must reach for it to be used. By using this feature with a relatively high threshold, the user can automatically replace most common variant forms, thereby saving a substantial amount of time processing the text. This function allows a semi-automatic approach; a user can spend some time training the tool, allowing method weights to be adjusted accordingly. The automatic replace option can then be used to deal with a large amount of cases after which the user can manually deal with any variants still remaining, if they wish.

An extension to the automatic replace option is a batch processing user interface which can be used to process as many text files as the user desires, for example a whole corpus. This interface uses the same background processes as the main interactive version but only has the option to automatically process variants. As in

the interactive tool, a threshold confidence measure can be set for replacements. The batch processing version can be used in conjunction with the interactive tool; a user can manually process a sample of the texts to be processed, the replacement methods will duly have their weights adjusted, some words not previously found in the dictionary will be added and some common variant-to-replacement mappings will be added. The user can then use the batch processing tool to automatically process the remaining texts. The user interface of the batch processing tool is shown in Figure 3 below.

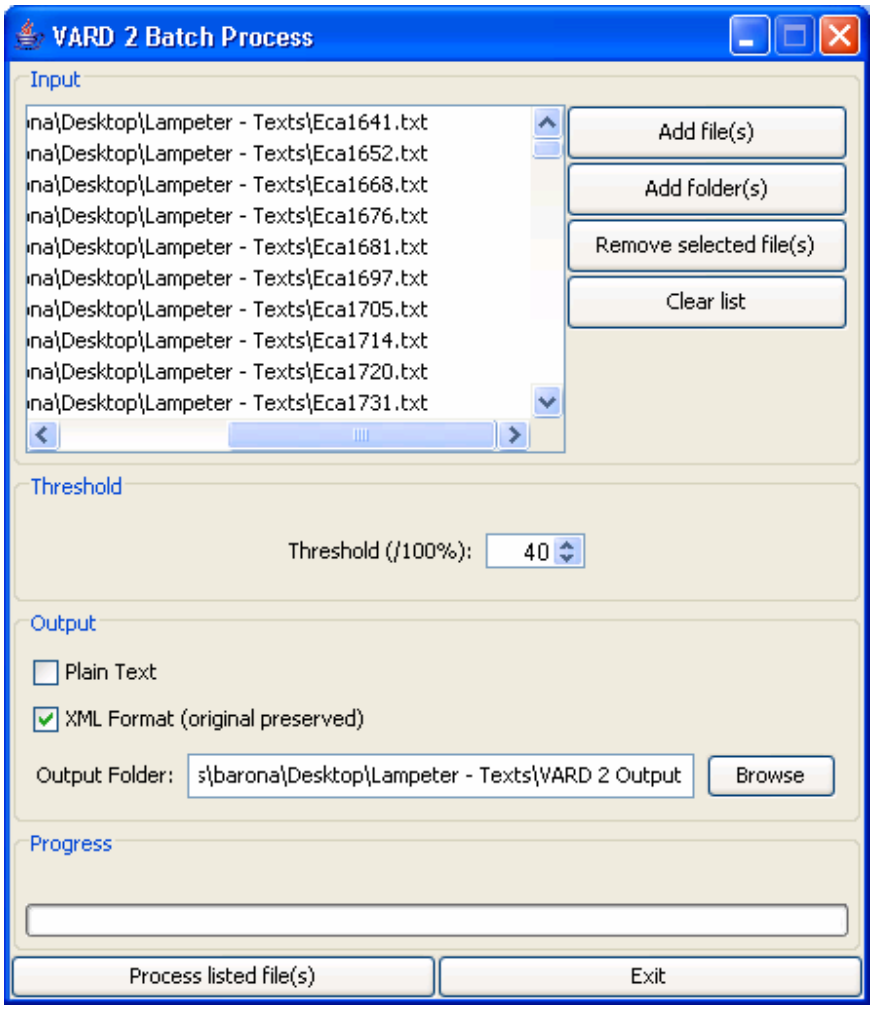

Figure 3 - VARD 2 batch processing interface

The interactive and batch versions of the software both produce output of the processed text. Two formats can be chosen from: plain text and xml. Plain text

simply returns the text in the original format but with modern equivalents present in the place of variants, where they have been replaced by the user or the system; the original spellings are lost in this format. A more useful output is an xml version of the text; here tags are included for remaining variant forms, and those which have been dealt with. Where a modern equivalent has been selected for a variant, the original spelling is stored as an attribute in a *replaced* tag around the modern equivalent which replaces the variant in the text. When other software reads the xml output only the modern equivalent is processed, however the tool can still have access to the original spelling through the xml tag attributes.

4. Conclusion

This paper has highlighted the effect spelling variation (especially in EModE) has on the accuracy of corpus linguistic tools. An interactive and flexible piece of software has been created which can pre-process texts containing spelling variation, producing a 'standardized' text which can be parsed more accurately by corpus linguistic software whilst retaining the original spelling variants for reference.

The VARD 2 tool is designed with EModE spelling variation in mind; however its learning capabilities and flexibility allow the tool to be used with potentially any form of spelling variation. This is an area which requires further investigation. VARD 2 is still under development and various improvements are planned for the near future. Evaluation of the tool's accuracy (Rayson et al, forthcoming) and its effect on part-of-speech tagging (Rayson et al, 2007) has already taken place. With future improvements to the tool, further evaluation of its accuracy, usability and effect on corpus linguistic tools will be necessary.

References

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

Culpeper, J. and Kytö, M. (1997). Towards a corpus of dialogues, 1550-1750. In Ramisch, H. and Wynne, K. (eds.). *Language in Time and Space. Studies in Honour of Wolfgang Viereck on the Occasion of His 60th Birthday* (Zeitschrift für Dialektologie und Linguistik - Beihefte, Heft 97). pp 60-73. Franz Steiner Verlag, Stuttgart.

Garside, R., and Smith, N. (1997). A hybrid grammatical tagger: CLAWS4, in Garside, R., Leech, G., and McEnery, A. (eds.) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Longman, London, pp. 102-121.

Görlach, M. (1991). *Introduction to Early Modern English,* Cambridge University Press, Cambridge.

Kukich, K. (1992). Techniques for Automatically Correcting Words in Text. *ACM Computing Surveys*, Vol. 24, No. 4, pp. 377-439.

Kytö, M., Rissanen, M. and Wright, S. (1994). *Corpora across the Centuries: Proceedings of the First International Colloquium on English Diachronic Corpora*, Cambridge, March 1993. Rodopi, Amsterdam.

Kytö, M. and Voutilainen, A. (1995). Applying the Constraint Grammar Parser of English to the Helsinki Corpus. *ICAME Journal* 19, pp. 23-48.

Leech, G., Rayson, P. and Wilson, A. (2001). *Word Frequencies in Written and Spoken English: based on the British National Corpus*. Longman, London.

Nevalainen, T. (1997). Ongoing work on the Corpus of Early English Correspondence. In Hickey, R., Kytö, M., Lancashire, I. and Rissanen, M. (eds.) *Tracing the Trail of Time: Proceedings from the Second Diachronic Corpora Workshop*. Rodopi, Amsterdam.

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.

Rayson, P., Archer, D. and Smith, N. (2005). VARD versus Word: A comparison of the UCREL variant detector and modern spell checkers on English historical corpora. In proceedings of the Corpus Linguistics 2005 conference, July 14-17, Birmingham, UK. *Proceedings from the Corpus Linguistics Conference Series on-line e-journal*, Vol. 1, no. 1, ISSN 1747-9398.

Rayson, P. (2007). Wmatrix: a web-based corpus processing environment, Computing Department, Lancaster University. Available on the internet at http://ucrel.lancs.ac.uk/wmatrix/.

Rayson, P., Archer, D., Baron, A., Culpeper, J. and Smith, N. (2007). Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. In proceedings of *Corpus Linguistics 2007*, July 27-30, University of Birmingham, UK.

Rayson, P., Archer, D., Baron, A. and Smith, N. (forthcoming). Travelling Through Time with Corpus Annotation Software. In *Proceedings of Practical Applications in Language and Computers (PALC) 2007*, The Department of English Language at Łódź University, Poland, 19th-22nd April 2007.

Richardson, M. (1980). Henry V, the English Chancery, and Chancery English. In *Speculum*, Vol. 55, No. 4, pp. 726-750.

Scott, M. (2004). *WordSmith Tools version 4*. Oxford University Press.

Singh, I. (2005). *The History of English*. Hodder Arnold, London.

Vallins, G. H. and Scragg, D.G. (1954). *Spelling*. André Deutsch.