Automatic Standardization of Spelling for Historical Text Mining

Alistair Baron, Paul Rayson (Lancaster University)
Dawn Archer (University of Central Lancashire)

1. Introduction

The use of textual data in humanities research is significantly aided by automated techniques such as key word analysis, collocations and corpus annotation (e.g. part-of-speech). If a text corpus contains a large amount of spelling variation, there is a considerable impact on the accuracy of these automatic techniques. For example, studies in respect to Early Modern English (EModE) corpora - the focus of the study detailed in this paper - have documented the adverse effects of spelling variation on key word analysis (Baron et al., forthcoming), part-of-speech tagging (Rayson et al., 2007) and semantic analysis (Archer et al., 2003).

The problem of spelling variation in corpora needs to be addressed in order for more accurate and meaningful results to be achieved in fields where historical source texts are required. Researchers can side-step the issue by using modernized versions of corpora, of course, but these are not always available. Another potential solution is to manually standardize the spellings; this includes reading through texts, spotting any non-standard spellings and deciding upon a modern equivalent, resulting in the production of a new version of the text with spelling variants replaced. However, a manual standardizing approach is likely to be unworkable when working with some of the larger corpora or online databases that are now available.

This paper details the current version of the VARiant Detector (VARD 2) tool, which can be used in various ways to standardize spelling variation in corpora. In particular, the tool can be used to (partially) standardize spellings automatically, with no restriction on the number of words to be processed. Here, we focus on the ways in which the tool can be trained from manually standardized corpora samples, particularly the letter replacement component of the tool, and evaluate the improvement that this makes to the performance of VARD 2.

2. Early Modern English Spelling Variation

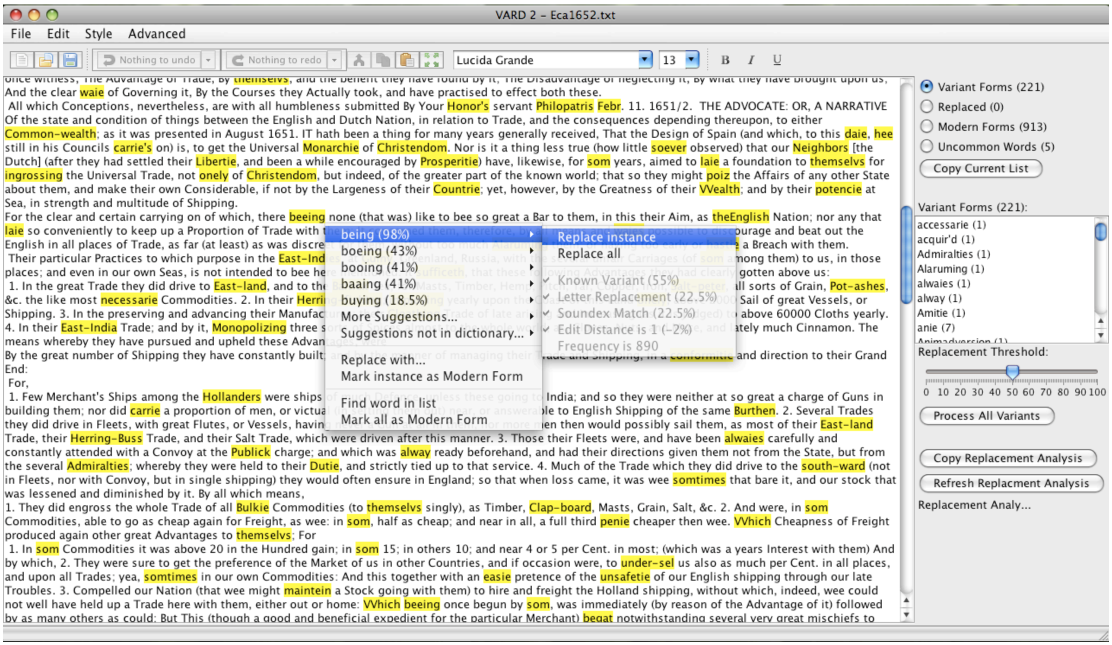
The EModE period is of particular interest in historical text mining studies; book production increased sharply during the period, largely due to the introduction of the printing press (1476) and increasing literacy levels (Görlach, 1991: 6). As such, the EModE period is the earliest period of the English Language from which a reasonably large corpus can be constructed and studied in detail.

Spelling variation was a major feature of EModE texts, the extent of which has recently been quantified in Baron et al. (forthcoming). It is common to find words spelt in a number of different forms in the same text or even on the same page. This was not seen as problematic, however, as there was no notion of the importance for a single spelling for each word; for example, letters would be added or removed to ease line justification. Vallins and Scragg (1965) and Culpeper and Archer (forthcoming) describe the spelling variation trends in more detail.

The effect of this spelling variation on textual analysis techniques has been shown in previous and forthcoming papers: key word analysis (Baron et al, forthcoming), part-of-speech tagging (Rayson et al., 2007) and semantic tagging (Archer et al., 2003). All of the studies showed that spelling variation causes considerable problems to the accuracy and meaningfulness of results, and that dealing with spelling variation (even partially) can achieve substantial improvements in annotation accuracy.[1] The production of standardized or modernized versions of historical corpora therefore allows for more accurate automated text mining techniques to be applied to the corpora.[2]

3. VARD 2 and DICER

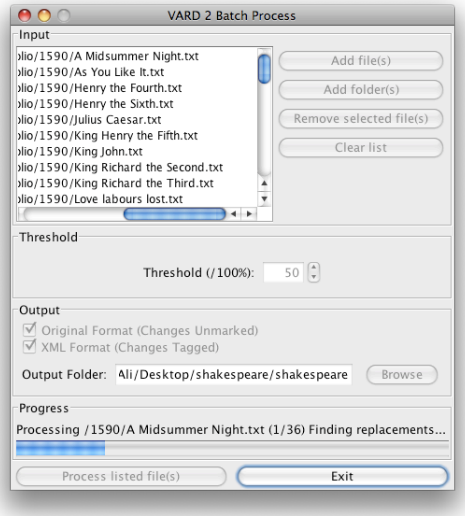
Our solution to the spelling variation problem described in the previous section has been the development of the VARD 2 tool,[3] a piece of software designed to assist researchers in standardizing historical corpora (specifically EModE texts) both manually and automatically. An earlier version of the VARD 2 software is described and evaluated in Rayson et al (2008). The current version can cater for user-created letter replacement rules, which will be used by the tool to find potential variant replacements. In addition, XML provision has been improved, processing speed increased, and a new word reference list[4] added. Screenshots of the latest version, VARD 2.2, are shown in Fig. 1 and Fig. 2.



once witness, The Advantage of Trade, By themselvs, and the benefit they have found by it, The Disadvantage of neglecting it, By what they have brought upon us,
And the clear waie of Governing it, By the Courses they Actually took, and have practised to effect both these.
All which Conceptions, nevertheless, are with all humbleness submitted By Your Honor's servant Philopatris Febr. 11. 1651/2. THE ADVOCATE: OR, A NARRATIVE
Of the state and condition of things between the English and Dutch Nation, in relation to Trade, and the consequences depending thereupon, to either
Common-wealth; as it was presented in August 1651. IT hath been a thing for many years generally received, That the Design of Spain (and which, to this daie, hee
still in his Councils carrie's on) is, to get the Universal Monarchie of Christendom. Nor is it a thing less true (how little soever observed) that our Neighbors [the
Dutch] (after they had settled their Libertie, and been a while encouraged by Prosperitie) have, likewise, for som years, aimed to laie a foundation to themselvs for
ingrossing the Universal Trade, not onely of Christendom, but indeed, of the greater part of the known world; that so they might poiz the Affairs of any other State
about them, and make their own Considerable, if not by the Largeness of their Countrie; yet, however, by the Greatness of their VVealth; and by their potencie at
Sea, in strength and multitude of Shipping.
For the clear and certain carrying on of which, there beeing none (that was) like to bee so great a Bar to them, in this their Aim, as theEnglish Nation; nor any that
laie so conveniently to keep up a Proportion of Trade with t ourage and beat out the
English in all places of Trade, as far (at least) as was discre a Breach with them.
Their particular Practices to which purpose in the East-Ind mong them) to us, in those
places; and even in our own Seas, is not intended to bee he gotten above us:
1. In the great Trade they did drive to East-land, and to the ll sorts of Grain, Pot-ashes,
&c. the like most necessarie Commodities. 2. In their Herri Sail of great Vessels, or
Shipping. 3. In the preserving and advancing their Manufac above 60000 Cloths yearly.
4. In their East-India Trade; and by it, Monopolizing three s ately much Cinnamon. The
means whereby they have pursued and upheld these Advan
By the great number of Shipping they have constantly built and direction to their Grand
End:
For,
1. Few Merchant's Ships among the Hollanders were ships India; and so they were neither at so great a charge of Guns in
building them; nor did carrie a proportion of men, or victua ble to English Shipping of the same Burthen. 2. Several Trades
they did drive in Fleets, with great Flutes, or Vessels, havin en then would possibly sail them, as most of their East-land
Trade, their Herring-Buss Trade, and their Salt Trade, which were driven after this manner. 3. Those their Fleets were, and have been alwaies carefully and
constantly attended with a Convoy at the Publick charge; and which was alway ready beforehand, and had their directions given them not from the State, but from
the several Admiralties; whereby they were held to their Dutie, and strictly tied up to that service. 4. Much of the Trade which they did drive to the south-ward (not
in Fleets, nor with Convoy, but in single shipping) they would often ensure in England; so that when loss came, it was wee somtimes that bare it, and our stock that
was lessened and diminished by it. By all which means,
1. They did engross the whole Trade of all Bulkie Commodities (to themselvs singly), as Timber, Clap-board, Masts, Grain, Salt, &c. 2. And were, in som
Commodities, able to go as cheap again for Freight, as wee: in som, half as cheap; and near in all, a full third penie cheaper then wee. VVhich Cheapness of Freight
produced again other great Advantages to themselvs; For
1. In som Commodities it was above 20 in the Hundred gain; in som 15; in others 10; and near 4 or 5 per Cent. in most; (which was a years Interest with them) And
by which, 2. They were sure to get the preference of the Market of us in other Countries, and if occasion were, to under-sel us also as much per Cent. in all places,
and upon all Trades; yea, somtimes in our own Commodities: And this together with an easie pretence of the unsafetie of our English shipping through our late
Troubles. 3. Compelled our Nation (that wee might maintein a Stock going with them) to hire and freight the Holland shipping, without which, indeed, wee could
not well have held up a Trade here with them, either out or home: VVhich beeing once begun by som, was immediately (by reason of the Advantage of it) followed
by as many others as could: But This (though a good and beneficial expedient for the particular Merchant) begat notwithstanding several very great mischiefs to

Variant Forms (221):
accessarie (1)
acquir'd (1)
Admiralties (1)
Alaruming (1)
alwaies (1)
alway (1)
Amitie (1)
anie (7)
Animadversion (1)

Fig. 1 Screenshot of VARD 2.2 showing the interactive mode which allows the user to manually standardize texts and train the tool on samples of a corpus



olio/1590/A Midsummer Night.txt
olio/1590/As You Like It.txt
olio/1590/Henry the Fourth.txt
olio/1590/Henry the Sixth.txt
olio/1590/Julius Caesar.txt
olio/1590/King Henry the Fifth.txt
olio/1590/King John.txt
olio/1590/King Richard the Second.txt
olio/1590/King Richard the Third.txt
olio/1590/Love labours lost.txt

Threshold (/100%): 50

☑ Original Format (Changes Unmarked)
☑ XML Format (Changes Tagged)

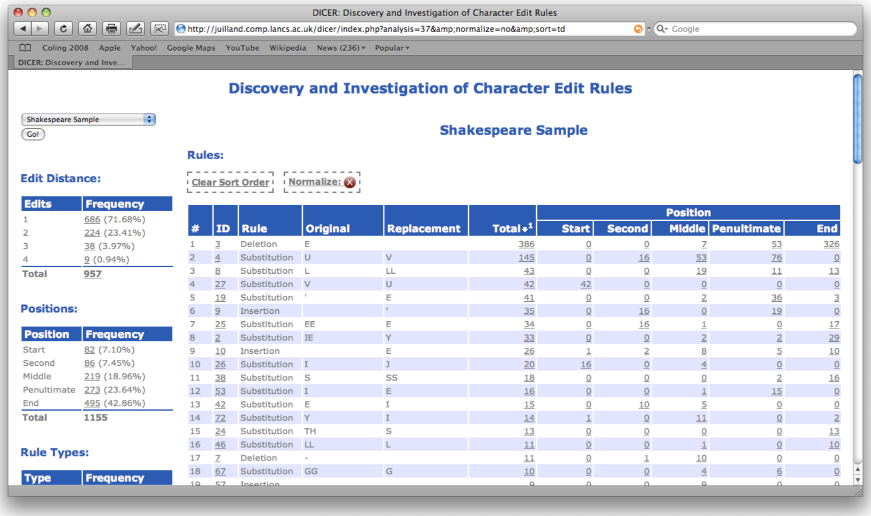
Output Folder: Ali/Desktop/shakespeare/shakespeare

Processing /1590/A Midsummer Night.txt (1/36) Finding replacements...

Fig. 2 Screenshot of VARD 2.2 showing the batch-processing mode where users can automatically standardize a chosen group of texts

The most useful way in which VARD 2 can be used is to automatically standardize spelling variation in an entire corpus. For EModE texts, this can be done immediately, with no training. However, for better results and to use the tool with other varieties of English, the user can train the software on a particular corpus by using the interactive version to manually process samples from the corpus. The tool will improve its ability to deal with a corpus based on decisions made by the user in the interactive version. It does this by learning which of its methods are most successful in finding the correct replacement for variants and adjusting its method weights accordingly (these are used when ranking potential replacements). The tool will also edit its dictionary and its list of specific variant replacements based on changes made by the user.

A new development to allow for further training of VARD 2 on a corpus is a tool named DICER (Discover and Investigation of Character Edit Rules). DICER can search XML output from VARD 2 for variant – replacement mappings or be provided with a list of such mappings. Each mapping is analyzed and a set of character edit rules are produced which can transform the spelling variant into its modern equivalent. The details of these character edit rules are then collated into a database, which can be viewed through a set of web pages.[5] The main table produced by the analysis, shown in Fig. 3, displays details of the individual character edit rules along with various frequencies. By clicking on individual rules, further information is available such as which characters typically occur before and after the rule occurs; this is shown for the rule ‘Delete e’ in Fig. 4. Any frequency in the tables can be clicked to view a list of occurrences producing that frequency. The data produced in the DICER analysis is vast and thus cannot be detailed in full here.


DICER: Discovery and Investigation of Character Edit Rules
http://juilland.comp.lancs.ac.uk/dicer/index.php?analysis=37&amp;normalize=no&amp;sort=td
Google
Coling 2008
Apple
Yahoo!
Google Maps
YouTube
Wikipedia
News (236)
Popular
DICER: Discovery and Inve...
Shakespeare Sample
Go!

Discovery and Investigation of Character Edit Rules

Shakespeare Sample

Edit Distance:

Edits	Frequency
1	686 (71.68%)
2	224 (23.41%)
3	38 (3.97%)
4	9 (0.94%)
Total	957

Positions:

Position	Frequency
Start	82 (7.10%)
Second	86 (7.45%)
Middle	219 (18.96%)
Penultimate	273 (23.64%)
End	495 (42.86%)
Total	1155

Rule Types:

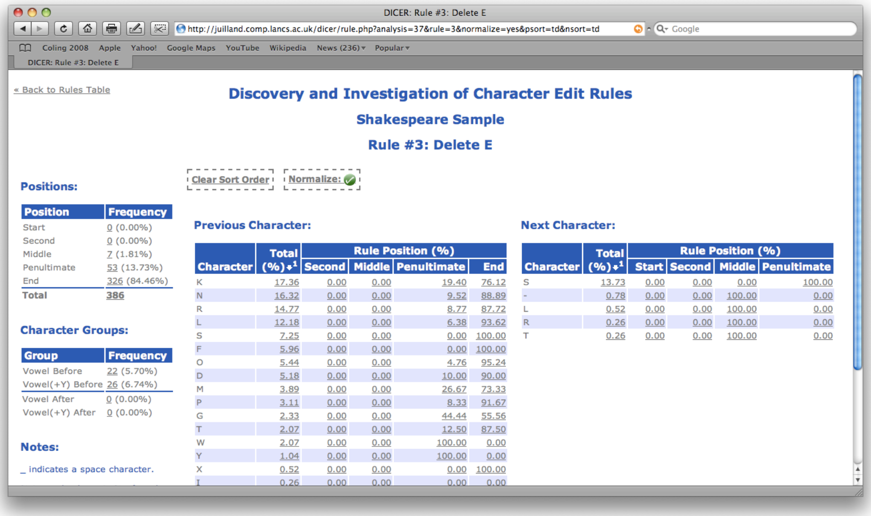
Type	Frequency

Rules:

Clear Sort Order Normalize: ☒

						Position				
#	ID	Rule	Original	Replacement	Total↓1	Start	Second	Middle	Penultimate	End
1	3	Deletion	E		386	0	0	7	53	326
2	4	Substitution	U	V	145	0	16	53	76	0
3	8	Substitution	L	LL	43	0	0	19	11	13
4	27	Substitution	V	U	42	42	0	0	0	0
5	19	Substitution	'	E	41	0	0	2	36	3
6	9	Insertion		'	35	0	16	0	19	0
7	25	Substitution	EE	E	34	0	16	1	0	17
8	2	Substitution	IE	Y	33	0	0	2	2	29
9	10	Insertion		E	26	1	2	8	5	10
10	26	Substitution	I	J	20	16	0	4	0	0
11	38	Substitution	S	SS	18	0	0	0	2	16
12	53	Substitution	I	E	16	0	0	1	15	0
13	42	Substitution	E	I	15	0	10	5	0	0
14	72	Substitution	Y	I	14	1	0	11	0	2
15	24	Substitution	TH	S	13	0	0	0	0	13
16	46	Substitution	LL	L	11	0	0	1	0	10
17	7	Deletion	-		11	0	1	10	0	0
18	67	Substitution	GG	G	10	0	0	4	6	0
19	57	Insertion			9	0	0	9	0	0

Fig. 3 Screenshot of DICER analysis on a manually standardized 5,000-word sample of Shakespeare's First Folio



Discovery and Investigation of Character Edit Rules

Shakespeare Sample

Rule #3: Delete E

Clear Sort Order Normalize: ✓

Positions:

Position	Frequency
Start	0 (0.00%)
Second	0 (0.00%)
Middle	7 (1.81%)
Penultimate	53 (13.73%)
End	326 (84.46%)
Total	386

Character Groups:

Group	Frequency
Vowel Before	22 (5.70%)
Vowel(+Y) Before	26 (6.74%)
Vowel After	0 (0.00%)
Vowel(+Y) After	0 (0.00%)

Notes:

_ indicates a space character.

Previous Character:

		Rule Position (%)			
Character	Total (%)↓1	Second	Middle	Penultimate	End
K	17.36	0.00	0.00	19.40	76.12
N	16.32	0.00	0.00	9.52	88.89
R	14.77	0.00	0.00	8.77	87.72
L	12.18	0.00	0.00	6.38	93.62
S	7.25	0.00	0.00	0.00	100.00
F	5.96	0.00	0.00	0.00	100.00
O	5.44	0.00	0.00	4.76	95.24
D	5.18	0.00	0.00	10.00	90.00
M	3.89	0.00	0.00	26.67	73.33
P	3.11	0.00	0.00	8.33	91.67
G	2.33	0.00	0.00	44.44	55.56
T	2.07	0.00	0.00	12.50	87.50
W	2.07	0.00	0.00	100.00	0.00
Y	1.04	0.00	0.00	100.00	0.00
X	0.52	0.00	0.00	0.00	100.00
I	0.26	0.00	0.00	0.00	0.00

Next Character:

		Rule Position (%)			
Character	Total (%)↓1	Start	Second	Middle	Penultimate
S	13.73	0.00	0.00	0.00	100.00
-	0.78	0.00	0.00	100.00	0.00
L	0.52	0.00	0.00	100.00	0.00
R	0.26	0.00	0.00	100.00	0.00
T	0.26	0.00	0.00	100.00	0.00

Fig. 4 Screenshot of DICER showing the rule 'Delete E' in a manually standardized 5,000-word sample of Shakespeare's First Folio

By using DICER to analyze manually standardized samples of a corpus, a list of common character edit rules can be viewed. These character edit rules can then be added to VARD 2 and the tool will be better equipped to make judgments on the correct replacement for variants found whilst automatically standardizing the corpus.

In order to test VARD 2 and DICER's training ability a 5,000-word sample of Shakespeare's First Folio[6] was manually standardized in the interactive-mode of VARD 2 as training data, the entire corpus was then automatically standardized. Using this small amount of training data (6% of the entire corpus) increased the proportion of spelling variants replaced[7] from 70.33% to 73.75%.

The automatic standardization (after training) resulted in 10,601 unique variant replacements. 70.35% of these replacements could be achieved through VARD 2's original set of character edit rules alone. DICER analysis was then produced on the manually standardized Shakespeare sample; this is shown in Fig. 3. VARD 2's rule list was augmented with additional rules from the DICER analysis: any rule occurring 10 or more times was added, if not already present. Using this new rule list 77.66% of the 10,601 unique replacements could now be found, an increase of 7.31%.

The results are extremely promising, and increasing the size of the manually standardized sample should improve these figures even further. DICER can also be used to provide probabilities dictating how likely a character edit rule should be applied in a certain position with specified surrounding characters. Modifying VARD to use these probabilities could see even greater improvements in performance.

4. Conclusion

This paper has described the problems that variant spellings cause for historical text mining, particularly for automated methods in historical corpus linguistics, such as part-of-speech tagging and key words analysis. In previous and forthcoming work, we have quantified the errors or differences that result from the application of untrained tools and techniques on historical data that has not been standardized. Our proposed solution is the VARD tool, which offers the potential to standardize spelling in historical texts automatically and with high accuracy. We have described recent improvements to VARD 2, such as the inclusion of a much larger modern dictionary that enables better detection of historical variants and matching with modern forms.

VARD 2 has been developed to deal with spelling variation in EModE texts; the tool can be used with its default settings to achieve partial standardization automatically. However, with some training, we have shown that VARD 2's performance is enhanced. Further training could allow the tool to be used with other varieties of non-standard English (e.g. SMS corpora and weblogs).

In the future, we will evaluate the extent to which variation that can only be detected contextually (e.g. 'then' for 'than' and 'bee' instead of 'be') contributes to the problem. Dealing with this problem requires more advanced techniques, e.g. POS tagging, to be used in the detection phase.

Notes

1. Of course, spelling variants themselves are important linguistic features and thus worthy of study: as such, although our focus relates to how we might deal with spelling variation within historical data as a means of enabling the (more)

effective use of automated analytical techniques, we advocate that any solution to this problem should always retain the original spelling.

2. It should be noted that the accuracy of annotation is likely to be affected by additional factors, including differences in the grammar of the EmodE period when compared to present-day English (see Kytö and Voutilainen, 1995) and the possibility of a semantic shift in words from EModE to present-day English (see, for example, Knapp, 2000).
3. The tool is available to download online, with a user guide also provided. The software is free to use for academic purposes from http://www.comp.lancs.ac.uk/~barona/vard2/
4. Derived from the Spell Checking Oriented Word List (SCOWL). See http://wordlist.sourceforge.net/scowl-readme
5. Available at http://juilland.comp.lancs.ac.uk/dicer/
6. Available from the Oxford Text Archive: http://ota.ahds.ac.uk/
7. Variants here are words which VARD 2 deems to be variants, i.e. words which are not in its modern lexicon. It should be noted that words will be incorrectly marked as variants (particularly proper names) and some variants will be incorrectly marked as modern words (particularly *read-word errors*, such as ‘bee’ for ‘be’ and ‘doe’ for ‘do’).

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.

Baron, A., Rayson, P. and Archer, D. (forthcoming). Word frequency and key word statistics in historical corpus linguistics. To appear in *Anglistik. International Journal of English Studies, Special Issue: Corpus Linguistic – Methods and Approaches.*

Culpeper, J. and Archer, D. (forthcoming). The History of English Spelling. In Culpeper, J., Katamba, F., Kerswill, P., Wodak, R. and McEnery, T. (eds), *English Language and Linguistics*. Palgrave Macmillan, Basingstoke, UK.

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

Knapp, P. A. (2000). *Time-Bound Words: Semantic and Social Economies from Chaucer's England to Shakespeare's.* Anthony Rowe Ltd, Chippenham, Wiltshire, UK.

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

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. (2008). Travelling Through Time with Corpus Annotation Software. In Lewandowska-Tomaszczyk, B. (ed) *Corpus Linguistics, Computer Tools, and Applications – State of the Art. Palc 2007*. Peter Lang, Frankfurt am Main.

Vallins, G. H., and Scragg, D. G. (1965). *Spelling*. André Deutsch.