Word frequency and key word statistics in historical corpus linguistics

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

1. Introduction

Frequency-sorted word lists have long been part of the standard methodology for exploiting corpora. Sinclair (1991: 30) noted that "anyone studying a text is likely to need to know how often each different word form occurs in it". Tribble and Jones (1997: 36) outlined a methodology for using texts in the language classroom, proposing that the most effective starting point for understanding a text is a frequency-sorted word list. A frequency list records the number of times that each word occurs in the text. It can therefore provide interesting information about the words that appear (and do not appear) in a text. A word list can be arranged in order of first occurrence, alphabetically or in frequency order. First-occurrence order serves as a quick guide to the distribution of words in a text, an alphabetic listing is built mainly for indexing purposes, but a frequency-ordered listing highlights the most commonly-occurring words in the text. Frequency dictionaries have appeared for Spanish, Rumanian, French, Portuguese, German and English (Juilland et al 1964, 1965 and 1970; Davies, 2006; Davies and Preto-Bay, 2008; Jones and Tschirner, 2006; Leech et al, 2001). Traditional dictionaries also use frequency information indirectly, in choosing entries for inclusion. Francis and Kučera (1982) took the simple word frequency list one stage further when they reported *grammatical word* frequencies drawn from the tagged version of the Brown corpus. Grammatical word frequencies are associated with a specific part-of-speech (POS) tag.

Although the computer saves us time when processing texts into frequency lists, it presents us with so much information that we need a filtering mechanism to pick out significant items prior to any analysis proper. There are at least two methods that we can use. First, formulae can be applied to adjust the raw frequencies for the distribution of words within a text; in other words, to describe the dispersion of frequencies in subsections of a corpus. Secondly, we can apply statistical procedures to highlight words that occur significantly more or less frequently than expected in a corpus. The frequency profile for a given text can be compared to the profile of a comparable text or to a profile derived from large amounts of text. Since the high frequency items tend to have a stable distribution generally, significant changes to the ordering of the words in the frequency list can flag points of interest to the researcher (Sinclair 1991: 31). For example, Hofland and Johansson (1982) use Yule's K statistic and the chi-squared goodness-of-fit test to pick out statistically significant different word frequencies across British and American English in their comparison of the two language varieties.

This paper examines the technique of key word analysis. This is one of the most widely-used methods for discovering significant words, and is achieved by comparing the frequencies of words in a corpus with frequencies of those words in a (usually larger) reference corpus. It should be noted that the vast majority of key words studies take place using corpora of modern language. However, in this paper, we look at the possible problems that may occur when applying the same technique to historical corpora, and in particular, corpora of Early Modern English, a variety for which there are significant volumes of text already available. In addition, there is a growing body of historical data from this period being scanned and transcribed in the large digitisation initiatives such as Early English Books Online, British Library Newspapers Digitisation Project etc.

The paper continues in section two with further background information on statistical techniques that are used to compare frequencies in corpora of modern languages. We also look at the few key word studies that have been carried out on historical data. In our case study, presented in section three, we first quantify the amount of spelling variation occurring in historical corpora. We then examine the problems of applying the key words technique to historical data and show how much key word lists are affected by issues of spelling variation. Our study quantifies, systematically and on a large scale, how the process of standardisation of written English throughout the Early Modern English period affects the robustness of key words results in historical corpus linguistics. In our conclusion (section four), we highlight possible solutions to this problem and describe directions for further work.

2. Background

2.1. Modern

Although word frequency lists are very useful as a starting point for the analysis of corpora, there are well-known problems with using them. First, the frequencies must be normalised before the lists can be compared directly. Second, high frequency words at the top of any word frequency list are generally of no further interest to those trying to examine the content of corpora. Third, comparing the ranking of words is also misleading. Finally, multiword expressions and inflectional variants of the same lemma are not counted together. For further description of these problems, see Rayson (forthcoming) and Hoffman et al (2008).

Even when they are derived from a large comprehensively-sampled corpus such as the British National Corpus (BNC), the word frequency counts themselves may be misleading. This is not because we might have miscounted the words, but because of how well the frequencies relate to usage in the English language as a whole. If a word has a high frequency count, we may reasonably infer, due to the nature of the BNC, that the word has a similarly high usage in the language. However, it may be the case that the word has a high frequency in the corpus not because it is widely used in the language as a whole but because it is widely used in a small(ish) number of texts, or parts of texts, within the corpus. To reveal these cases, we can calculate range or dispersion statistics. These show how widely distributed the occurrences of a word are within a corpus: i.e. whether it is frequent because it occurs in a lot of text samples in the corpus or whether it is frequent because of a very high usage in only a subset of texts (which may represent particular domains or genres). Frequent words with high dispersion values may be considered to have high currency in the language as a whole; high frequencies associated with low dispersion values should, in contrast, be treated with caution. For example, Church and Gale (1995) term this as the "bunchiness" or "burstiness" of words and show that the occurrences of the "very contagious" word "Kennedy" are not evenly dispersed in the Brown corpus (because he was the president of the United States when the Brown corpus was compiled in 1961).

In the discipline of statistics, the *mean* and *standard deviation* are used as summary measures. In corpus linguistics, these are analogous to *frequency* and *dispersion*. According to Fries and Traver (1950: 21), Thorndike was the first to introduce range values into frequency lists. For further discussion of dispersion statistics, see Lyne (1985). Another way of dealing with the burstiness of words is to combine separate frequency and dispersion values into one measure called *adjusted frequency* (Francis and Kučera, 1982: 464). Words can then be ranked by their adjusted frequencies. A more complex approach for describing variability within corpora is proposed by Gries (2006).

The comparison of word frequency profiles has increasingly been used to examine issues in language variation, that is, to compare language usage across corpora, users, genres, etc. There are two types of corpus comparison. First, a comparison of a sample corpus with a larger 'normative' (or general language standard) corpus (e.g. Scott, 2000b). Second, a comparison of two roughly equal-sized corpora (e.g. Granger, 1998). These two main types can be extended to the comparison of more than two corpora. For example, we may compare one normative corpus to several smaller corpora at the same time, or compare three or more equal-sized corpora with each other. In general, however, this makes the results more difficult to interpret. Homogeneity (Stubbs, 1996: 152) within each of the corpora is important since we may find that the results reflect sections within one of the corpora that are unlike other sections in either of the corpora under consideration (Kilgarriff, 1997). There are a number of different statistics that can be applied in the comparison of word frequency lists. In what follows, we will examine a number of these in order to see how key word analysis operates.

Hofland and Johansson (1982) carried out one of the largest early studies comparing word frequency profiles. This was the comparison of one million words of American English (the Brown corpus) with one million words of British English (the LOB corpus). They used a difference coefficient defined by Yule (1944) to assess the difference in the relative frequency of a word in the two corpora:

$$\frac{Freq_{LOB} - Freq_{Brown}}{Freq_{LOB} + Freq_{Brown}}$$

The value of the coefficient varies between +1 and –1. A positive value indicated overuse in the LOB corpus, a negative value showed overuse in the Brown corpus. A statistical goodness-of-fit test originally suggested by Pearson (1904), the chi-squared test (χ^2), was also used to compare word frequencies across the two corpora. The chi-squared test was calculated as follows:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} \text{ where } E_i = \frac{N_i \sum_i O_i}{\sum_i N_i}$$

where O_i is the observed (actual) frequency, E_i is the expected (averaged) frequency, and N_i is the total frequency in corpus i (i in this case takes the values 1 and 2 for the LOB and Brown corpora respectively). Hofland and Johansson marked any resulting differences that were indicated by chi-squared values showing statistically significant difference at the 5%, 1%, or 0.1% level. The null hypothesis of the goodness of fit test is that there was no difference between the observed frequencies of a word in the two corpora. Note that even if the null hypothesis was not rejected, they could not conclude that it is true. The cut-off value corresponding to the chosen degree of confidence may not be exceeded, but this only indicates there is not enough evidence to reject the null hypothesis (Krenn and Samuelsson, 1997: 36). Critical values for the chi-squared statistic are listed in statistical tables such as those in Barnett and Cronin (1986) and Oakes (1998: 266). For example, the critical value for the 5% level, shown as 0.05 in the tables, is 3.84 at 1 degree of freedom (see below). Leech and Fallon (1992) used the lists produced by Hofland and Johansson to examine evidence of cultural differences between America and Britain in 1961.

In corpus linguistics, we usually use a 2 × 2 table to compare frequencies of words or other linguistic features between two corpora. The chi-squared test is applicable to a general table with *r* rows and *c* columns. The number of degrees of freedom (d.f.), which is used when looking up critical values, is defined as the number of independent terms given that the marginal totals in the table are fixed. So, in the 2 × 2 table, d.f., as calculated by (r-1)×(c-1), is equal to 1. In this case, the 2 × 2 'contingency' table is as shown in Table 1 .

	CORPUS ONE	CORPUS TWO	TOTAL
Frequency of feature	a	b	a+b
Frequency of feature not occurring	c	d	c+d
TOTAL	a+c	b+d	N=a+b+c+d

Table 1 - Contingency table for the chi-squared test

Hence, we can calculate the chi-squared statistic (X^2) as follows:

$$X^2 = \frac{N(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

When comparing the frequency distribution of word classes across the two major subdivisions of the Brown corpus, informative prose and imaginative prose, Francis and Kučera (1982: 544) used a normalised ratio value (NR) rather than the chi-squared test. The ratio is normalised to take account of the fact that the informative section is nearly three times larger than the imaginative section of the corpus. An NR value of more than 1 indicates a greater occurrence in informative prose, while a value of less than 1 points to a higher relative frequency in imaginative texts. The greater the NR deviates from 1, the greater the grouping of a particular word class in one of the sections of the corpus. Comparing NR values is problematic since they are not on a linear scale, and the calculation is too generous for smaller relative differences when lower frequency items are compared to higher frequency items. Francis and Kučera (1982: 547) also employed the Mosteller-Rourke (MR) adjustment for chi-squared for large numbers. The MR value is calculated as follows:

$$MR = \frac{1000\chi^2}{n}$$

where *n* is the frequency of an item in the whole corpus (Mosteller and Rourke, 1973: 191). The resulting values cannot be assessed for significance in the chi-squared tables, but they are used to rank items according to their MR value. In effect, MR reduces the chi-squared values for items occurring more than 1000 times, and increases the values for items with a frequency that is less than 1000. This seems a rather arbitrary figure, chosen to show 'nice numbers' and, if anything, the figure should be dependent on the corpus size(s).

Kilgarriff (1996a, 1996b) pointed out that, in the Brown versus LOB comparison, many common words are marked as having significant chi-squared statistics and that, because words are not selected at random in language, we will always see a large number of differences in two such text collections. As an alternative, Kilgarriff selected the Mann-Whitney test that uses ranks of frequency data rather than the frequency values themselves to compute the statistic. Kilgarriff selected the Mann-Whitney test because it "does not give undue weight to single documents with a high [frequency] count" for a particular word. However, he observed that, even with the new test, 60% of words are marked as significant. Ignoring the actual frequency of occurrence, as in the Mann-Whitney test, means discarding most of the evidence we have about the distribution of words. As such, the test will have lower discriminatory power. Due to problems of too many zeros in the Mann-Whitney test, Kilgarriff (2001) reported that his technique omits words with less than 30 occurrences in the joint LOB and Brown corpus. This is a major drawback with the Mann-Whitney test; here it omits 92% of the types in the joint corpus. A further problem is that many words share ranks at the low end of frequency lists, especially for large corpora. For example, Copeck et al (1999) report that 18,630 words occur six times – 10 percent of their list for the BNC. Within each rank, words are ordered alphabetically. Additionally, comparing rank lists between different-sized corpora is also problematic. Copeck et al (1999) note the sizes of their frequency lists for LOB (7,950) and Wall Street Journal (4,550). This means that ranks for middle and lower frequency words in the BNC fall outside this range. These points suggest that the Mann-Whitney ranks test is suitable only for investigating mid- to high-frequency words when comparing corpora of the same size.

Numerous other authors have used the chi-squared test to determine significant frequency differences of individual words or other linguistics features, rather than whole frequency profiles, between two corpora (for example Woods et al 1986: 140, Virtanen 1997, Oakes 1998: 26, Roland et al 2000, Wikberg 1999). Many authors also apply Yates' continuity correction (1934), developed to improve the approximation of the continuous probability distribution (chi-squared) to the discrete probability distribution of the observed frequency (multinomial). The Yates' corrected chi-squared statistic (Y^2) is calculated as follows (from Table 1):

$$Y^2 = \frac{N(|ad - bc| - 0.5N)^2}{(a+b)(c+d)(a+c)(b+d)}$$

In some texts, its use has been recommended (Everitt 1992: 14, Butler 1985: 122, and Woods et al 1986: 146), but current statistical textbooks report that the correction is less important than it was once thought (Agresti 1990: 68). Fisher's exact test may be used for tables with small expected frequencies, as an alternative to the chi-squared test. It uses the observed frequencies themselves to find the probability (P) of obtaining any particular arrangement of frequencies a, b, c, and d (again from Table 1):

$$P = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!N!}$$

where *a!* is 'a factorial' (the product of *a* and all the whole numbers less than it, down to one, 0! = 1). The P value is then compared directly to the probability level, e.g. 0.05 for 5%, or 0.01 for 1%, to indicate departure from the null hypothesis in a specific direction. It is a *one-tailed test* whereas the chi-squared is two-tailed. The P value may be doubled in order to compare it with the probability obtained through the chi-squared test. Fisher's exact test is computationally expensive since it involves calculating factorials, and the value of P for every possible arrangement of frequencies keeping the marginal totals fixed.

Dunning (1993) reported that we should not rely on the assumption of a normal distribution when performing statistical text analysis and suggested that parametric analysis based on the binomial or multinomial distributions is a better alternative for smaller texts. Dunning also went on to propose the log-likelihood ratio as an alternative to Pearson's chi-squared test, and he demonstrated this for the extraction of significant bigrams from text. Conversely, Mosteller and Rourke (1973: 162) state that the chi-squared statistic assumes a multinomial distribution, as do Cressie and Read (1994). Woods et al (1986: 188) describe the chi-squared test for association as non-parametric and state that it makes no special distributional assumptions of normality. There seems to be some confusion in the literature. Everitt (1992: 5-8) explains the situation more clearly. It is the observed frequencies that are assumed to follow a multinomial distribution, whereas the chi-squared distribution (which is used to calculate and tabulate critical values) arises from the normal distribution. Some papers in the literature report that the chi-squared statistic becomes unreliable when the expected frequency is *too small*, and *possibly* overestimates significance with high frequency words and when comparing a relatively small corpus to a much larger one. The former of these vague terms has been taken as meaning that all *expected* values must be greater than 5 (for example, Butler 1985: 117, Woods et al 1986: 144), and sometimes the same limit is applied to the *observed* frequencies (De Cock, 1998 and Nelson et al, 2002: 277). It was Cochran (1954) who suggested a rule that 4 in 5 (80%) of the expected values in an r × c table should be 5 or more. In the 2 × 2 table case, this means all cells should have expected values of 5 or more. Everitt (1992: 39) cites other more recent work than Cochran, which suggests that this rule is too conservative. Butler (1985: 117) suggests one possible solution to this is to combine frequencies until the combined classes have an expected frequency of 5 or more; likewise Nelson et al (2002: 277) for the observed frequencies, but Everitt (1992: 41) argues against this practice.

Everitt (1992: 72) also mentions that the chi-squared statistic is "easily shown to be an approximation to" the log-likelihood for large samples. The two statistics take similar values for many tables. Williams (1976) notes that the log-likelihood is preferable to Pearson's chi-squared in general. Everitt (1992: 18) also notes that the chi-squared test, Yates' corrected chi-squared and Fisher's exact test are equivalent in large samples. The obvious question, then, is: what constitutes a large sample? Kretzschmar et al (1997) start to answer the question by estimating sample sizes for various confidence levels. Scott (2001b) uses the log-likelihood statistic in his keywords procedure, as we shall see below. For the 2 × 2 case (in Table 1), the log-likelihood ratio is calculated as follows:

$$G^2 = 2\,(a\ln a + b\ln b + c\ln c + d\ln d + N\ln N - (a+b)\ln(a+b) - (a+c)\ln(a+c) - (b+d)\ln(b+d) - (c+d)\ln(c+d))$$

Cressie and Read (1984) show that Pearson's X^2 (chi-squared) and the likelihood ratio G^2 (Dunning's log-likelihood) are, in fact, two statistics in a continuum defined by the power-divergence family of statistics. They go on to describe this family in later work (1988, 1989). Here, they also make reference to the long and continuing discussion (since 1900) of the normal and chi-squared approximations for X^2 and G^2, and 2 × 2 contingency tables, during which many alternative tests have been devised (Yates, 1984).

Finally, we can present the key word approach taken by Scott, which takes a systematic approach to the comparison of word frequency lists (Scott 1997, 1998, 2001a). Tribble (2000: 79-80) describes the way that the WordSmith tool finds keywords as follows:

1. frequency sorted wordlists are generated for the 'reference' corpus and for the research text or texts
2. each word in the research text is compared with its equivalent in the reference text and the program evaluates a statistical test based on the log-likelihood procedure to calculate the keyness
3. the wordlist for the research corpus is reordered in terms of the 'keyness' of each word

Scott (1997) sets a minimum threshold of two occurrences for each word in the research text, although this does result in manually identified keywords being omitted from the keywords database (Scott, 2001b: 118). Other words with frequencies that violate the Cochran rule are still included in the keyword listing since, in practice, they are still interesting. The resulting keyword list contains two types of keyword: *positive* (those which are unusually frequent in the target corpus in comparison with the reference corpus), and *negative* (those which are unusually infrequent in the target corpus). These correspond to the terms *overuse* and *underuse* used in the learner corpus literature to describe the same observations. Tribble compares the list of positive and negative keywords against the frequency list for his corpus and demonstrates the improved usefulness of the keyword technique over simple frequencies for extracting interesting lexical items for stylistic studies. Scott also uses the notion of key-keywords. These are words that are key in all, or a large percentage, of the texts that are contained in the corpus under investigation. Tribble uses this feature to select lexical items that (may) give pedagogical insights in respect to (particular) genre(s). Key-keywords give us an insight into the dispersion of a key word in the corpus.

Having reviewed how the key words procedure works and the different possibilities for the statistical apparatus that is used in the procedure, we now turn our attention to (some of the) studies which have applied the word frequency and key word approaches to historical data.

2.2. Key word studies relating to historical data

Sub-branches within the field of historical linguistics have a long tradition of utilising corpus-based techniques (see, e.g., Risannen et al 1993 for an example of early studies made possible because of the advent of the computer/computer-based techniques, and Jucker et al. 1999: 16-20 for an overview of the impact of computerisation on historical linguistic research methods). As such, it may surprise the reader to learn that there are relatively few studies of historical data that make use of the key words approach. Several of these (e.g., Culpeper 2002, sections of Culpeper and Archer 2008, Mahlberg 2007a, Mahlberg 2007b, Mahlberg forthcoming, Archer et al. forthcoming) explore classic English literature, whilst others explore specific activity types such as the historical English courtroom (see, e.g., Archer 2006 and sections of Culpeper and Archer 2008) or specific topics such as swearing (see, e.g., McEnery 2005, McEnery forthcoming).

It is worth noting that the majority of these studies make use of Mike Scott's *WordSmith Tools* programme. Yet, only two of the above - Culpeper (2002) and Scott and Tribble (2006) – are mentioned in Scott's online lists of key word studies.[1] Both explore *Romeo and Juliet*: Scott and Tribble compare the latter to a number of reference corpora (other Shakespearean tragedies, the *Complete Works of Shakespeare* and the British National Corpus (BNC)) to determine the extent to which the choice of reference corpora affected the key word results for *Romeo and Juliet*; Culpeper (2002) explores the extent to which key words can be used to identify the characteristics of six characters from the play. His choice of reference corpus was thus the *Romeo and Juliet* play itself minus the particular character's speech he was investigating at that time. Interestingly, Culpeper opted to utilise a modern edition of *Romeo and Juliet* (W. J. Craig's 1914 edition) as opposed to, for example, the First Folio from the Oxford Text Archive. His reasoning is that he wished to avoid as much spelling variation as possible, not least because "spelling variation is perhaps the greatest obstacle in the statistical manipulation of historical texts" (Culpeper 2002: 14).

The idea that multiple variant spellings within a text greatly hinder standard corpus linguistic methods (such as frequency profiling, concordancing and key word analysis) is commonly-held (e.g. Markus, 2002). Indeed, it is highlighted by Archer and Culpeper (forthcoming, 2009) in their key word (key part-of-speech and key domain) study of EmodE social dyads (taken from comedy plays and trial proceedings), and by Archer et al. (forthcoming) in their key word (and key domain) comparison of Shakespearean *love*-comedies and *love*-tragedies.[2] It is the belief that spelling variation adversely affects the

[1] See: http://www.lexically.net./wordsmith/corpus_linguistics_links/papers_using_wordsmith.htm).

[2] Not all researchers who have employed key word techniques on historical data explicitly raise the issue of multiple spelling variants. This should not be taken as a sign that they do not regard the latter as a problem. On the contrary, they may have sidestepped the problem of spelling variation altogether by opting for modern editions and/or they may work on texts that represent an historical period where spelling was relatively fixed

accuracy of the statistical manipulation of historical texts which also prompted a number of researchers to develop a variant detector that can detect and normalise spellings, using a variety of computational techniques (see, e.g., Archer et al. 2003; Rayson et al 2005).

Prior to this paper, however, no specific work had been undertaken to test the degree to which key word results are affected by multiple spelling variants (as far as we are aware). We seek to address this, here, by quantifying the effect of historical spelling variation on the lists of key words extracted from corpora.

3. Case study

3.1. The extent of spelling variation

The aim for the first part of the analysis presented here was to discover, quantitatively, the extent of spelling variation in the Early Modern English (EModE) period, not least because many researchers comment on the large amount of spelling variation within the period without explicitly quantifying it (see, e.g., Vallins and Scragg (1965); Görlach (1991)). One exception is Schneider (2002) who, in his attempts to develop a normalised version of the Zurich English Newspaper (ZEN) Corpus (1670-1799)[3], produced an overview of the spelling variations contained within. Schneider found that 3.99% of the tokens and 38.02% of the types within the corpus were unrecognised by the ENGCG tagger[4], and hence could be considered spelling variants. The corpus was also split into four time periods, 1670-1709, 1710-1739, 1740-1769 and 1770-1799. The percentage of unrecognized tokens and types reduced in each subsequent time period, from 4.66% tokens and 36.57% types in the 1670-1709 sub-corpus to 2.85% tokens and 26.06% types in the 1770-1799 sub-corpus.

As this paper will cover the entire EModE period[5], a more thorough quantitative study of the spelling variation within the period is required. To this end, six different corpora were analysed: The ARCHER corpus, Early English Books Online, the Innsbruck Letter corpus, the Lampeter corpus, a corpus of Early English medical writings, and a collection of Shakespeare's works. The ARCHER corpus (A Representative Corpus of Historical English Registers)[6] is a multi-purpose diachronic corpus covering from 1650 to the present day (only texts dated before 1800 were used in this study). It was built to facilitate the analysis of historical change in written and speech-based registers. Early English Books Online (EEBO)[7] is a collection of digital facsimiles of virtually every English printed work between 1473 and 1700; nearly 125,000 works. As digital images of texts are of no use in this study, we have been given access to 12,268 of the 25,000 works that are being transcribed into ASCII SGML

(see, e.g., Mahlberg's investigations of Dicken's works, for example, and McEnery's investigations of swearing, and the response to "bad language" use exhibited by political movements such as the Society for the Reformation of Bad Manners).

[3] See Fries and Schneider (2000) for more details.

[4] See Voutilainen and Heikkilä (1994) for details.

[5] The precise dating of the EModE period is a topic of some contention, see for example Görlach (1991: 9-11). Henry V's commitment to the vernacular in 1417 (Richardson, 1980: 727) could be considered the earliest date for the period, whilst 1776, the year of the American Declaration of Independence - "the notional birth of the first (non-insular) extraterritorial English" (Lass, 1999a: 1) could be considered the latest date.

[6] We used the ARCHER-3.x version of the corpus (1990–1993/2002/2007/2010). Compiled under the supervision of Douglas Biber and Edward Finegan at Northern Arizona University, University of Southern California, University of Freiburg, University of Heidelberg, University of Helsinki, Uppsala University, University of Michigan, University of Manchester, Lancaster University, University of Bamberg and University of Zurich.

[7] http://eebo.chadwyck.com/

texts as part of the EEBO Text Creation Partnership.[8] The Innsbruck Letter corpus, part of the Innsbruck Computer-Archive of Machine-Readable English Texts (ICAMET) corpus (Markus, 1999) is a collection of 469 complete letters dated between 1386 and 1688, a total of 182,000 words. The Lampeter corpus of Early Modern English Tracts (Schmied, 1994) is a collection of tracts and pamphlets published between 1640 and 1740. Each decade has two texts from each of the following six domains: religion, politics, economy & trade, science, law, and miscellaneous; resulting in a corpus of 120 texts and c. 1.1 millions words. The Early Modern English Medical Texts (EMEMT) corpus (Taavitsainen et al., forthcoming; Taavitsainen and Pahta, 1997 and forthcoming) is a collection of specifically medical texts built to study the evolution of medical writing. The portion of the corpus available to us covers 1525 to 1700. The collection of Shakespeare's works is a digitally-transcribed version of the first folio, which was printed in 1623. This can be sourced from the Oxford Text Archive[9]. Shakespeare's works were written between c. 1590 and c. 1613. A summary of the corpora used in this study is shown in Table 2[10].

Corpus	**Genre and Type**	**Years Eligible**[11]	**Texts Eligible**	**Tokens Eligible**
ARCHER	General / Mixed	1660-1799	364	632,639
EEBO	General / Mixed	1470-1709	12,265	535,910,150
Innsbruck	Letters	1410-1689	436	170,538
Lampeter	Religion, Politics, Economy & Trade, Science, Law, and Miscellaneous tracts and pamphlets	1640-1739	120	1,124,131
EMEMT	Medical texts	1540-1699	51	491,384
Shakespeare First Folio	All plays (Comedies, Histories and Tragedies) from the First Folio.	c1590-c1613[12]	36	821,123

Table 2 - Summary of corpora used in study

The total coverage of the corpora used in the study dates from 1410 to 1799; thus representing the entire EModE period. The corpora are all very different, covering various genres and text types. It is important to note that the corpora are never combined in our study and are always treated as separate entities.

The first stage of the study involved sampling each corpus at regular intervals in order to gain a fair representation of the corpus over time. A sample period of ten years was chosen, hence the texts were split into their relevant decade (e.g. 1410 – 1419). This level of sampling

[8] http://www.lib.umich.edu/tcp/eebo/

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

[10] We wish to thank Manfred Markus for allowing us to use the Innsbruck Letter Corpus and Irma Taavitsainen for providing us with a copy of the EMEMT corpus.

[11] The full decade range was not used from all corpora due to texts dating too far from the EModE period or a lack of texts and/or words from certain decades.

[12] It should be noted that the dates given for Shakespeare's plays are estimates as there is considerable debate in respect to precise dating. In any case, the First Folio was printed in 1623 and it is difficult to know exactly the extent to which the editors adhered to the original source of each play. The Shakespeare plays cover only a small section of the EModE period and are included here to show any contrast between them and other corpora from the same time period.

did mean a small number of decades were omitted in certain corpora due to a lack of texts and/or words. The smaller EMEMT corpus could not be sampled in this way due to many decades containing only one or two files, or a small number of words; therefore the decision was made to include everything from the EMEMT corpus with a minimum of two files per decade. All results were normalised to a percentage in order to compare corpora with different sample sizes. The sampling sizes for each corpus are shown in Table 3.

Corpus	**Decade Sample Size**	**Minimum Texts**	**Decades <u>not</u> included due to a lack of texts and/or words**
ARCHER	4,000	10	1740
EEBO	80,000	10	
Innsbruck	1,200	4	1420, 1430, 1490, 1590
Lampeter	40,000	10	
EMEMT	Total Possible	2	1620, 1640
Shakespeare First Folio	60,000	4	

Table 3 - Corpus sample sizes.

For the more general corpora (ARCHER, EEBO and Lampeter), a minimum of ten texts per decade were required to ensure that one text did not account for more than 10% of a decade's sample. Elsewhere, a smaller number of texts were sufficient due to the fact that the specialised form of the corpora resulted in less variety of text. Samples were chosen from randomly selected texts from each decade, with the sample from each text beginning at a randomly selected index (word count) within the text.

In order to discover the extent of spelling variation per corpus and per decade, each word in a given historical sample was compared to a modern word list derived from the Spell Checking Oriented Word Lists (SCOWL)[13] and a word list containing words with a frequency greater than 5 in the British National Corpus (BNC) (Leech et al., 2001). If a word was not found in the modern word lists it was classed as a spelling variant. This analysis provided a percentage of variant types and tokens per corpus and per decade sample. The variant type percentages are plotted in Figure 1 and the variant token percentages are plotted in Figure 2. An average variant percentage over all the available corpora for each decade was also calculated; this is shown for types in Figure 3 and for tokens in Figure 4. The general trend line is shown with a dotted line in all four graphs.

[13] http://wordlist.sourceforge.net/scowl-readme

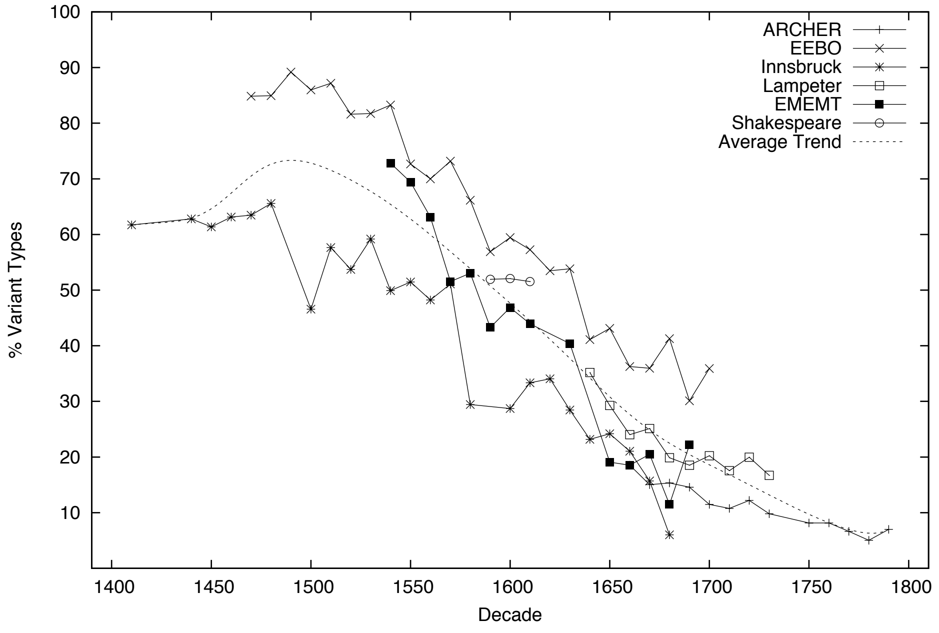


Figure 1 - Graph showing variant types % in all corpora over time.

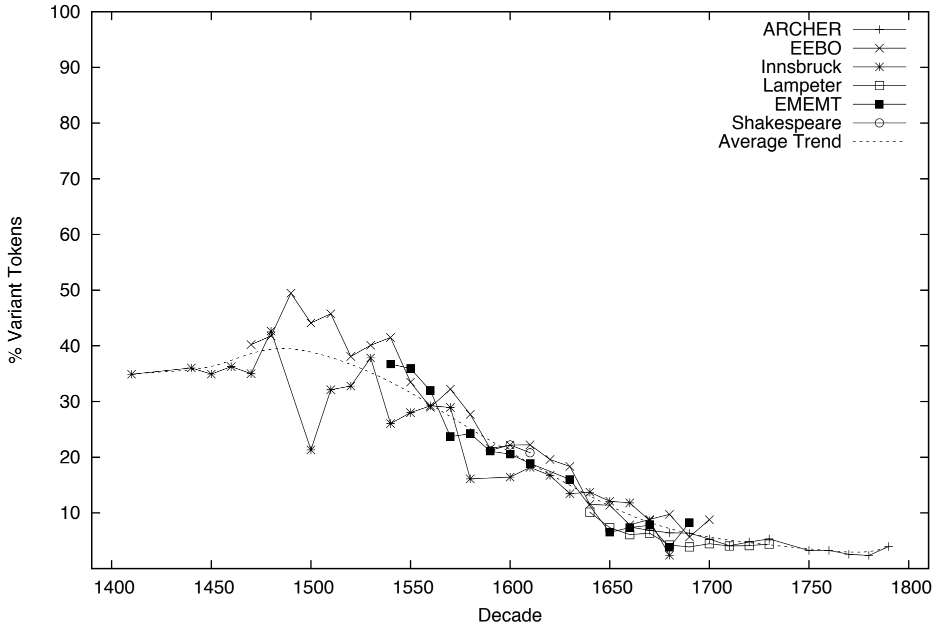


Figure 2 - Graph showing variant tokens % in all corpora over time.

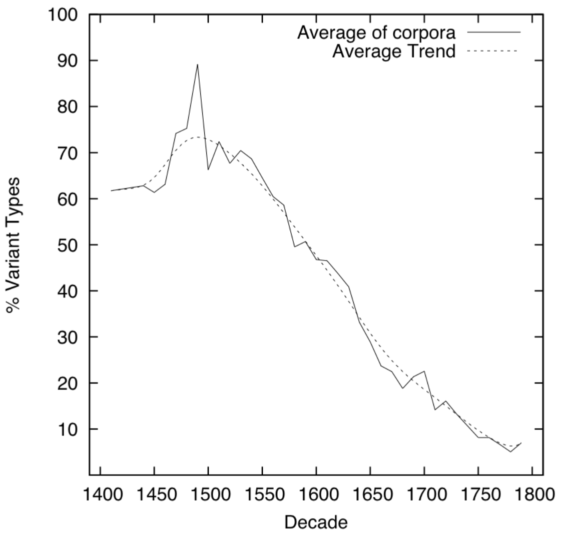


Figure 3 - Average variant types % over corpora available for each decade.

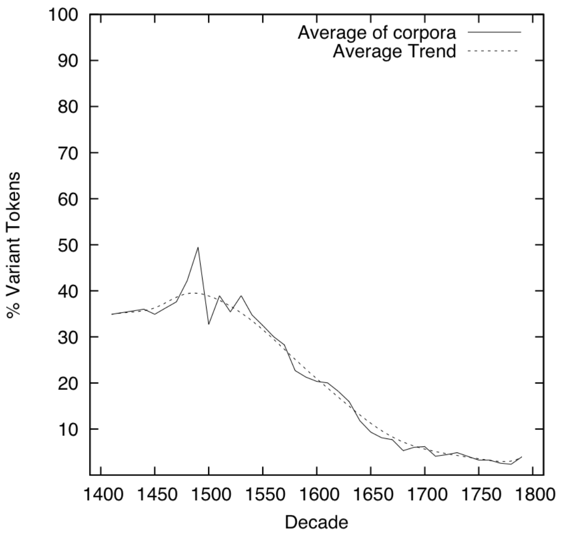


Figure 4 - Average variant tokens % over corpora available for each decade.

Figures 1-4 all show a definite downwards trend in respect to the amount of spelling variation occurring throughout the EModE period. This not only corroborates Schneider's (2002) quantitative analysis of the ZEN Corpus for the latter part of the EModE period (1670-1799), but also quantifies the trend over the entire EModE period, verifying many scholar's claims that the language was under significant change throughout the period (see, e.g., Görlach, 1991:8-9; Lass, 1999b: 56, Rissanen, 1999: 187). Another point to note is that the rate of reduction in variation slows from around 1700; this is particularly noticeable in the graphs representing tokens (Figures 2 and 4). This backs up Görlach's (1991: 11) claim that, by 1700, the language had achieved "considerable homogeneity," with regional (written) dialect differences no longer present and "the period of co-existing variants, so typical of all levels of EModE, being over."

It should be noted that the variant percentages shown in Figures 1-4 do not represent absolutely precise variant rates; they are all approximate values. It is extremely difficult to precisely calculate variant rates for large samples of text due to the problems involved when computing which words are actually variants – automatically processing large samples is necessary due to the time required for manual normalisation. First, so called 'real-word errors' are a concern; these are undetectable when comparing to a modern word list (as in this study), contextual knowledge is required to distinguish between variants which happen to match another modern word and words which are spelt in the 'standardised' modern form, e.g. "be" for "bee". An analysis of two small manually standardised samples (one from the Lampeter corpus, the other from Shakespeare's First Folio) used in a previous study (see Rayson et al., 2007) indicated the real-word error rates shown in Table 4. These figures are relatively low compared to real-word error rates in Modern English. By way of illustration, Peterson (1986) found that between 2% and 16% of typing errors would be undetected depending on the size of the word list used. Mitton (1987) found much larger rates; 40% of the spelling errors found in his study were real-word errors. In addition, when we opted to replicate the procedure we have used to analyse the Lampeter and Shakespeare samples (see above) on a manually-processed corpus of child language spelling errors we found that 24.07% of the variant types identified and 18.31% of variant tokens identified were real-word errors.

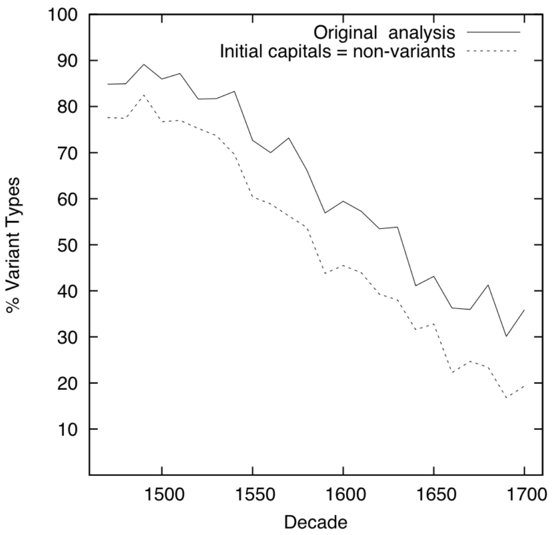
Sample	**Total words**		**% of words which required normalisation (i.e. variants)**		**% of variants which are real-word errors**		**% of words erroneously marked as variants**[14]	
	Types	**Tokens**	**Types**	**Tokens**	**Types**	**Tokens**	**Types**	**Tokens**
Lampeter	839	2,726	19.19%	9.61%	4.35%	2.67%	12.04%	4.37%
Shakespeare	897	3,991	63.88%	24.03%	8.55%	5.11%	7.80%	3.38%

Table 4 – Analysis of variants found in manually standardised EModE samples.

Table 4 indicates another problem when detecting spelling variants automatically - words incorrectly marked as variants. These may include proper nouns, encoded words (e.g. with Unicode entity values), words in languages other than English (e.g. Latin and French) and words which are simply not in the modern word list but are perfectly valid (e.g. archaic and obsolete words such as *betwixt* and *howbeit*). All of the problems listed occur in some of the corpora used in this study. Whilst a large amount of time was spent "cleaning" the texts, it is impossible to remove all imperfections. EEBO, for example, contains many Unicode entities for which there is no obvious ASCII replacement, and any word containing one (or more) of these values will be counted as a variant by our detector. Lampeter, ARCHER, Innsbruck and EEBO are known to contain sections of Latin and, in some cases, French passages; some of these passages will no doubt have been passed into the corpora samples. Aside from the odd exception all words in these foreign passages will be counted as variants.

Proper nouns invariably cause problems when detecting spelling errors, whether in historical texts or in modern spell checkers. Due to the potentially large number of proper nouns which could be found within any text, it is not sensible to try and list them all (although adding more frequent proper nouns is a sensible first step). A common-sense approach to the problem would be to exploit the rule that proper nouns always begin with a capital letter in Modern English; this, however, does not work in all cases as a capital letter is also used to signify the start of a sentence. The problem is even worse in EModE, particularly in later EModE texts. Osselton (1998) describes how between 1550 and 1750 there was a distinct climb in the use of a capital letter to begin nouns where one would not be present in Modern English. The effect of this proper noun "problem" is evaluated in Figures 5 and 6 where the EEBO corpus samples are analysed as above and also by counting all words beginning with a capital as non-variants. As can be seen, variant counts are consistently lower if words with initial capitals are not considered as variants. However, the general downward trend remains the same with the lines following almost parallel paths. Marking all initial capital words as non-variants will no doubt lead to an increase in real-word errors due to "abnormal" capitalization of words which are also variants, sentence initial variants and inconsistently spelt proper nouns.

[14] These are words which are “detected” as variants after the text has been normalised.



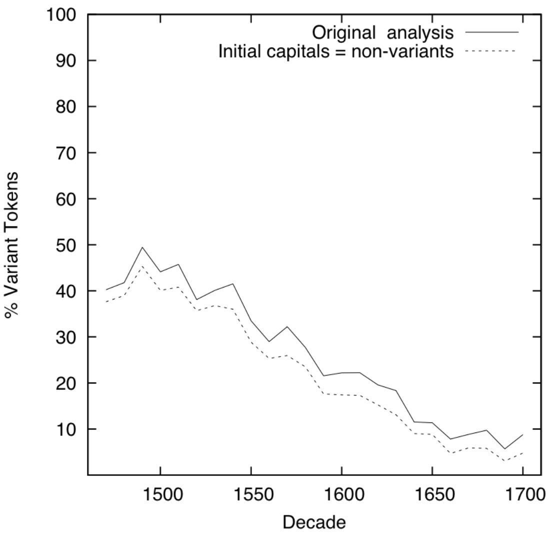


Figure 5 - Comparison of variant type counts in EEBO corpus samples with (=original) and without initial capital words.

Figure 6 - Comparison of variant tokens counts in EEBO corpus samples with (=original) and without initial capital words.

It is clear that the level of variation displayed in Figures 1-6 are approximations.[15] However, it is reasonable to assume that the level of "noise" leading to inaccuracies is relatively uniform throughout corpus samples and thus the general trend of spelling variation reducing over time throughout the Early Modern English period is maintained.

3.2. The effect of spelling variation on keyword analysis

The second part of our case study analyzes the effect caused by the levels of spelling variation described in the previous section. The focus of our analysis will be the effect on key word lists, as described in section 2.1. In order to discover any effect caused by spelling variation, key word lists need to be formulated before and after spelling variation is removed, thus, any change in the key word list rankings will indicate an effect of spelling variation.

Producing versions of texts or corpora with spelling variation removed is no simple task; except for very small samples, manually standardising texts is an exceedingly time-consuming process. Fortunately, one of the corpora in our study, the Innsbruck Letter Corpus, has been standardised and manually checked. The standardised corpus contains parallel line pairs, the first line in each pair contains the original text, the second line contains a standardised version of the first line with any spelling variants replaced with modern English word equivalents. The corpus was split into two parts, one containing just the original text lines (this was sampled in section 3.1), the other containing the standardised equivalent lines. This resulted in two separate corpora on which a key word analysis could be completed, and the differences between the lists analysed.

For this particular part of our study, log-likelihood was used to identify key words, and Wmatrix (Rayson, 2007) was used to produce key word lists. The BNC Written

[15] The average Shakespeare decade sample variant rates for types and tokens respectively were 51.84% and 21.41%, compared to 63.08% and 23.04% in the manually processed sample. For Lampeter the average decade sample variant rates were (types/tokens) 22.64%/5.50% and for the manually processed sample: 19.19%/9.61%.

Sampler[16] was used as a reference corpus. Any word with a log-likelihood greater than or equal to 6.63 ($p < 0.01$ for 1 d.f.) was considered key, and any word with a frequency less than 5 in either the Innsbruck Letter Corpus (before or after standardisation) or the BNC sample was removed from the key word list. We included both overused and underused words that were considered key. After this filtering process, two key word lists remained, one representing the original corpus and the other representing the standardised corpus, each containing the same list of words along with their log-likelihood value representing each word's keyness in its parent (original or standardised) corpus. It was important that both lists contained the same list of words, as we wanted to analyse the effect on key word list ranks, not the number of extra variants appearing in the original list. Our hypothesis was that whilst there will be some similarity between the key word list rankings from the original corpus and the standardised corpus due to them originating from essentially the same texts, we expect a large deviation in the rankings; therefore showing a degradation in accuracy due to spelling variation. We wished to both prove this hypothesis and quantify the amount of deviation.

In order to calculate the difference between the two key word lists, rank correlation was used. Rank correlation measures the correspondence between two different rankings on the same set of items and returns a value between -1 and 1; -1 is returned if one ranking is the exact reverse of the other, 0 is returned if the rankings are completely independent and 1 is returned if the two rankings are exactly the same. For this study, two rank correlation statistics were used: Spearman's Rank Correlation Coefficient (Spearman, 1904) and Kendall's Tau Rank Correlation Coefficient (Kendall, 1938).

The first stage was to produce a set of log-likelihood observation pairs, these were created by performing a look-up of the log-likelihood values from both lists for each word. Both rank correlation statistics convert the log-likelihood values into ranks; that is every word will have a rank associated to it representing where the word appears in each list sorted descending by log-likelihood. For Spearman's Rank Correlation Coefficient the differences (d_i) between each word's ranks are calculated, then the coefficient (ρ) is given by:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

where n is the number of words. Kendall's Tau Rank Correlation Coefficient works slightly differently in that it looks at the difference between each possible pairing in one list, if the sign of this difference (whether it is greater than, equal to, or less than 0) is equal to the sign of the difference between the same pair in the other list a concordant pair is counted (n_c), otherwise a discordant pair is counted (n_d). The coefficient (τ) is then calculated with:

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n - 1)}$$

with n again representing the number of words.

Both rank correlation statistics were calculated on the paired log-likelihoods as described above and the results are shown in Table 5. Both figures show that whilst there is some correlation between the two key word lists there is a definite difference between the rankings of the standardised version's key word list and the original version's key word list. We can therefore confirm our original hypothesis (i.e. a deviation in the rankings of some key

[16] Although clearly not the best match as a comparable corpus since it is from a different time period and design to the historical corpora, this effect will be minimised since we are using the same reference corpus for both before and after standardisation corpus comparisons. For more details about the BNC Sampler, see: http://www.natcorp.ox.ac.uk/corpus/index.xml.ID=products#sampler

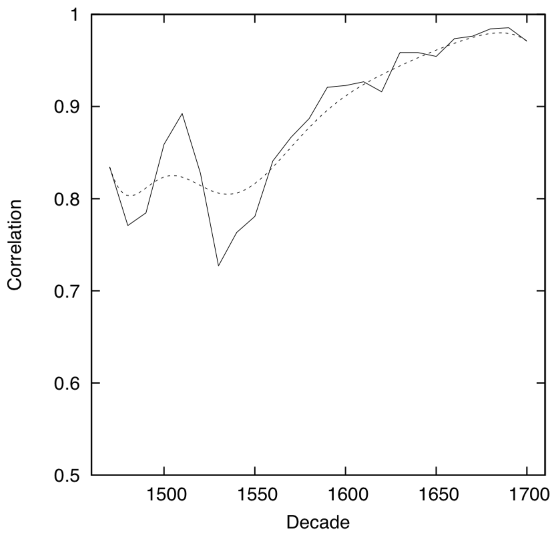
words) and conclude that spelling variation does have an effect on key word analysis of the Innsbruck Letter Corpus.

Rank Correlation Method	Score
Spearman's Rank Correlation Coefficient	0.7045437
Kendall's Tau Rank Correlation Coefficient	0.5304464

Table 5 - Rank correlations found when comparing the original and standardised versions of the Innsbruck Letter Corpus.

In order to further show the effect of spelling variation on key word analysis, we wished to analyse key word lists before and after standardisation of samples from different time periods. Our hypothesis was that there would be more differentiation between the key word lists for samples that represent the earlier centuries of the EModE period, due to the greater level(s) of spelling variation evidenced at that time (as shown in section 3.1). As with the key words analysis of the Innsbruck Letter Corpus, we required both original and standardised versions of a corpus, this time sampled at regular intervals throughout the EModE period. Due to the significant amount of time required to manually standardise large samples, automatically (partly) standardised samples were deemed sufficient to detect a trend. A tool, named VARD (Rayson et al, forthcoming; Baron and Rayson, 2008), has been developed which can perform automatic standardisation of historical texts. The tool inserts modern equivalents alongside any historical spelling variants where the probability of a match is greater than a threshold set by the user. The tool does not successfully replace all spelling variants in a given text automatically, however a large amount of spelling variants can be dealt with before the user manually processes the remaining variants.

For this study we decided to use the EEBO corpus as it covers the EModE period and has enough texts available per decade to build a large sample. The same decade samples used in section 3.1 were processed by VARD, producing partly standardised matching samples. As with the Innsbruck corpus, both versions of the samples were then processed with Wmatrix to produce two key word lists per sample. These lists were then filtered exactly as before, after which Spearman's Rank Correlation Coefficient and Kendall's Tau Rank Correlation Coefficient were calculated for each decade sample. The two coefficients are plotted in Figure 7 and Figure 8, with the dotted line showing the average trend.



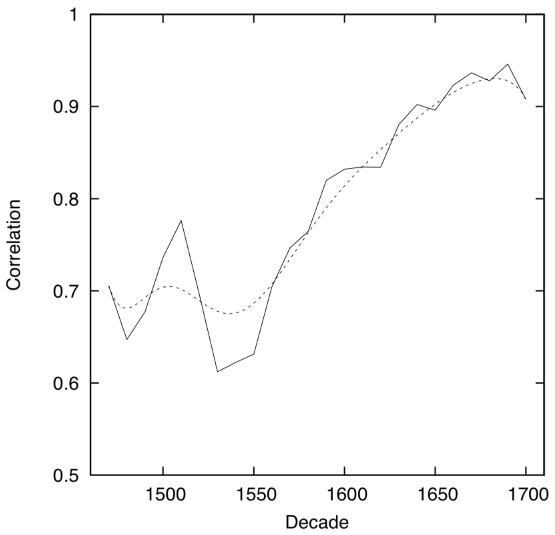
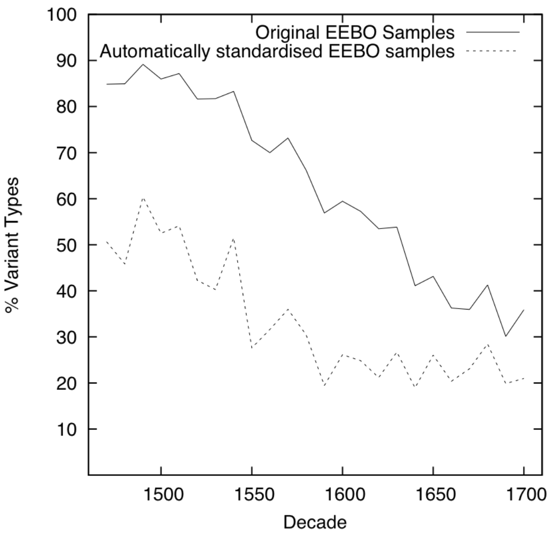


Figure 7 - Graph showing Spearman's Rank Correlation Coefficients comparing EEBO decade samples' key word lists before and after automatic standardisation.

Figure 8 - Graph showing Kendall's Tau Rank Correlation Coefficients comparing EEBO decade samples' key word lists before and after automatic standardisation.

The two graphs show erratic results for the earliest decade samples. This is mirrored (although not to the same extent) in the variant rates shown in Figures 1-6. This can be explained by examining the samples, especially that for 1510-19, a local maximum in Figures 7 and 8. The sample for 1510-19 contains a large section of foreign translations, containing many different languages. It is not possible to standardise this section, and so the standardised version will be more similar to the original version. This is shown in Figures 9 and 10, where the amount of spelling variation remaining after automatic standardisation is both higher and more erratic for the earlier decade samples.



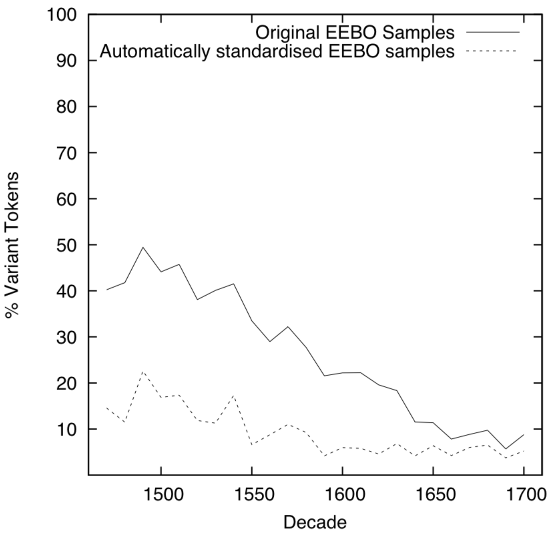


Figure 9 - Graph showing the frequency of spelling variant types in the EEBO samples before and after automatic standardisation.

Figure 10 - Graph showing the frequency of spelling variant tokens in the EEBO samples before and after automatic standardisation.

Noise in corpora of this nature is unavoidable and will have an influence on the results, also the effect of spelling variation is underestimated due to spelling variation still remaining (shown in Figures 9 and 10). However, the general upwards trend can be clearly seen for both coefficients, indicating an increase in correlation between the two key word lists the later the decade of the sample. We can conclude that a reduction in spelling variation over time produces less effect on key word analysis, thus proving our hypothesis.

4. Conclusion and future work

In this paper, we have given an overview of the use of word frequency profiles and key words in corpus linguistics. We began with a review of the various statistics used when comparing word frequencies between corpora. We also noted that the studies that have exploited the key words technique on historical data have tended to use modernised versions of their datasets in order to sidestep the issue of spelling variation. In the case study presented in this paper, we carried out a quantitative analysis of spelling variation in a set of well-known historical corpora. The trends identified match the expected rapid decline in spelling variation until around 1700. For the first time, we have been able to quantify the extent of spelling variation in these corpora. The second part of the case study showed the effect of this variation on the key words procedure. We were able to demonstrate how the key words lists were affected by comparing the lists produced from original historical data with that of a standardised version of the corpora. We also showed that the reduction in spelling variation over time has a knock-on effect on key word accuracy, with samples from later decades suffering less of an effect from spelling variation.

We will continue to refine our techniques for detecting historical spelling variants, including, for example, contextual clues to detect so called 'real-word errors' as described in section 3. However, the quantitative trends presented here are already clear enough. Researchers using frequency-based techniques on non-standardised historical datasets should be wary of spelling variation and need to exercise caution when interpreting key words analyses carried out on such data. Where standardised versions of corpora are available, the results obtained from them can be considered more robust. However, where it is unfeasible to carry out manual standardisation, for example on the vast digitised textual resources such as EEBO, there is a need for a tool which can detect historical variants and automatically standardise them in a pre-processing step for the application of key words and other modern corpus linguistic procedures.

A prototype for such a tool, VARD, was discussed in section 3.2. In future work, we plan to further develop the VARD tool. Currently, VARD employs the following procedures as a means of detecting variants, and mapping them to their 'modern' equivalents: a manually produced list of variants, SoundEx phonetic matching, edit distance and letter replacement heuristics. But these procedures are merely dealing with the surface forms of words. We will therefore be attempting semantic disambiguation in the near future so that we can also begin to distinguish the underlying meanings of words and their variants. This is important in respect to variants such as 'peece', which have more than one potential modern form (i.e. 'peace' and 'piece'). It is worth noting that, by adding a semantic component to the VARD, we have come full circle in our research endeavour as the VARD initially grew out of attempts to develop an historical version of the Wmatrix tool so that we could semantically annotate historical texts automatically (see Archer et al 2003). A related issue is the problem of 'real-word errors', as previously mentioned in section 3.1. Variants such as 'bee' for 'be' or 'then' for 'than' are impossible to detect with a dictionary check alone. Therefore, future work will involve using part-of-speech and semantic information to detect potential spelling variants of this type in order to achieve more accurate automatic standardisation. However, it is important to be aware that this is a circular issue in that spelling variation will have an effect on part-of-speech and semantic tagging accuracy as shown by Rayson et al (2007) and Archer et al (2003) respectively. One solution may be to incorporate the part-of-speech tagger in the variant detection process; this has been partially explored by Atwell and Elliot (1987).

4.1. Investigating spelling from a diachronic perspective

Although our main aim in this paper has been determining the effect that spelling variation has on (the meaningfulness of) keyness results, we effectively provide a means of quantifying the ongoing process of standardisation of written English throughout the EModE period, as witnessed by the decreasing levels of spelling variation. Moreover, we do so by exploring written texts that are both representative of the different centuries (and decades within) that make up the EModE period and also representative of different genres (i.e. plays, letters, medical texts, etc.). To our knowledge, we are the first to do this systematically on such a large scale: prior to Schneider (2002), who looks at the Zen corpus, this study and an earlier study by Archer and Rayson (2004) (for details of which, see below), most studies that have explored spelling from a diachronic (i.e. historical) perspective have tended to be qualitative in focus, that is, they have attended to the most obvious spelling patterns for a given period. Smith (2005: 222), for example, comments on the following patterns for Shakespearean English: the inter-changeability of <u> / <v> (depending on their initial/medial positioning), the use of <i> to represent <j> and the use of <vv> for <w> (see also Blake 1996; Scragg 1974). This focus is not surprising, given that these are the patterns that will strike the consciousness of the researcher as they read through texts. But it means that patterns below the level of consciousness – patterns that, for example, might be more subtle or only emerge across many texts – go unnoticed. The VARD tool therefore affords us with the opportunity to begin exploring spelling variability more subtly and systematically, whilst also determining the point(s) at which standardisation occurred (depending on the genre(s) under investigation).

In respect to our own future work, we plan to assess the extent to which *genre* plays a part in which variants are used as well as the extent to which they are used. This work will build on Archer and Rayson's (2004) study of 3,823 spelling variants in a variety of text-types representative of the 17th, 18th and 19th centuries, which appears to suggest that levels of spelling variation differed quite substantially across individual genres. For example, they examined a seventeenth century *Newsbook Corpus*, which effectively contained 296 occurrences per million words (of the 3,823 forms identified by them) – a frequency that seems very low, when compared to the 2,247 occurrences (per million words) found in (the seventeenth century component of) the *Lampeter* dataset. As Culpeper and Archer (forthcoming) highlight, the latter effectively contains genres - 'science', 'religion', 'politics', 'law' and 'economy' - which are regarded as having some of the very factors that are meant to provide a motivating force for standardisation (i.e. prestige and power). Yet, Archer and Rayson's study suggests that the more broad-based, popular genre of newsbooks was in the vanguard instead in the seventeenth century. Such a (surprising) result merits the type of systematic diachronic comparison of genres that the current study affords.

References

Agresti, Alan. *Categorical data analysis*. New York: Wiley, 1990.

Archer, Dawn. "Tracing the development of 'advocacy' in two nineteenth century English trials." *Diachronic Perspectives on Domain-Specific English*. Eds. Marina Dossena and Irma Taavitsainen. Bern: Peter Lang; Linguistic Insights series, 2006.

Archer, Dawn and Jonathan Culpeper. "Identifying *key* sociophilological usage in plays and trial proceedings (1640-1760): An empirical approach via corpus annotation." *Journal of Historical Pragmatics* Sociopragmatics Special Issue. Ed. Jonathan Culpeper. Forthcoming 2009.

Archer, Dawn and Paul Rayson. "Using an historical semantic tagger as a diagnostic tool for variation in spelling." *Thirteenth International Conference on English Historical Linguistics (ICEHL 13)*. Vienna, Austria: University of Vienna. (23-29 Aug. 2004).

Archer, Dawn, Tony McEnery, Paul Rayson and Andrew Hardie. "Developing an automated semantic analysis system for Early Modern English." *Proceedings of the Corpus Linguistics 2003 conference. UCREL technical paper number 16*. Eds. Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery. (2003): 22-31.

Archer, Dawn, Jonathan Culpeper, and Paul Rayson. "Love – 'a familiar or a devil'? An Exploration of Key Domains in Shakespeare's Comedies and Tragedies." *What's in a word-list? Investigating word frequency and keyword extraction*. Ed. Dawn Archer. Ashgate, forthcoming.

Atwell, Eric and Stephen Elliot. "Dealing with ill-formed English text." *The Computational Analysis of English. A corpus-based approach.* Eds. Roger Garside, Geoffrey Leech and Geoffrey Sampson. London/New York: Longman, 1987.

Barnett, Stephen and Timothy M. Cronin. *Mathematical formulae for engineering and science students*. 4th ed. London: Longman, 1986.

Baron, Alistair and Paul Rayson. "VARD 2: A tool for dealing with spelling variation in historical corpora." *Proceedings of the Postgraduate Conference in Corpus Linguistics*. Birmingham: Aston University (22 May 2008).

Blake, Norman F. *Shakespeare's Language: An Introduction*. 2nd ed. Basingstoke: Macmillan, 1996.

Butler, Christopher. *Statistics in linguistics*. Oxford: Blackwell, 1985.

Church, Kenneth W. and William A. Gale. "Poisson mixtures." *Natural Language Engineering* 1.2. Cambridge: Cambridge University Press, 1995. 163-190.

Cochran, William G. "Some methods for strengthening the common χ^2 tests." *Biometrics* 10 (1954): 417-451.

Copeck, Terry, Ken Barker, Sylvain Delisle and Stan Szpakowicz. "More Alike than not - An Analysis of Word Frequencies in Four General-purpose Text Corpora." *Proceedings of the 1999 Pacific Association for Computational Linguistics Conference (PACLING 99).* Ontario, Canada: Waterloo (25-28 Aug. 1999): 282-287.

Cressie, Noel and Timothy R. C. Read. "Multinomial Goodness-of-Fit Tests." *Journal of the Royal Statistical Society. Series B (Methodological)* 46.3 (1984): 440-464.

Cressie, Noel and Timothy R. C. Read. "Pearson's X^2 and the Log Likelihood Ratio Statistic G^2: A comparative review." *International Statistical Review* 57.1. Belfast: Belfast University Press, 1989. 19-43.

Culpeper, Jonathan. "Computers, language and characterisation: An analysis of six characters in Romeo and Juliet." *Conversation in Life and in Literature: Papers from the ASLA Symposium* 15. Eds. Ulla Merlander-Marttala, Carin Ostman and Merja Kytö. Uppsala : Universitetstryckeriet, 2002. 11-30.

Culpeper, Jonathan and Dawn Archer. "The History of English Spelling." *English Language and Linguistics.* Eds. Jonathan Culpeper, Francis Katamba, Paul Kerswill, Ruth Wodak and Tony McEnery. Basingstoke, UK: Palgrave Macmillan, forthcoming.

Davies, Mark. *A frequency dictionary of Spanish*. London: Routledge, 2006.

Davies, Mark. and Ana Maria Raposo Preto-Bay. *A frequency dictionary of Portuguese*. London: Routledge, 2008.

De Cock, Sylvie. "A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English." *International Journal of Corpus Linguistics* 3.1. Amsterdam: John Benjamins, 1998. 59-80.

Dunning, Ted. "Accurate Methods for the Statistics of Surprise and Coincidence." *Computational Linguistics* 19.1. MIT Press, March 1993. 61-74.

Everitt, Brian S. *The analysis of contingency tables*. 2nd ed. London: Chapman and Hall, 1992.

Francis, W. Nelson. and Henry Kučera. *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston: Houghton Mifflin, 1982.

Fries, Charles C. and A. Aileen Traver. *English word lists: a study of their adaptability for instruction.* Ann Arbor, Michigan: George Wahr Publishing Company, 1950.

Fries, Udo and Peter Schneider. "Zen: preparing the Zurich English Newspaper Corpus." *English media texts – past and present. Language and textual structure*. Ed. Friedrich Ungerer. Amsterdam / Philadelphia: John Benjamins, 2000. 3-24.

Görlach, Manfred. *Introduction to Early Modern English*. Cambridge: Cambridge University Press, 1991.

Granger, Sylviane. "The computer learner corpus: a versatile new source of data for SLA research." *Learner English on Computer*. Ed. Sylviane Granger. London: Longman, 1998. 3-18.

Gries, Stefan Th. "Exploring variability within and between corpora: some methodological considerations." *Corpora* 1.2 (2006): 109-151.

Hoffmann, Sebastian, Stefan Evert, Nicholas Smith, David Lee and Ylva Berglund Prytz. *Corpus Linguistics with BNCweb - a Practical Guide*. Frankfurt am Main, Germany: Peter Lang, 2008.

Hofland, Knut and Stig Johansson. *Word frequencies in British and American English*. Bergen, Norway: The Norwegian Computing Centre for the Humanities, 1982.

Jones, Randall and Erwin Tschirner. *A frequency dictionary of German*. London: Routledge, 2006.

Jucker, Andreas H., Gerd Fritz and Franz Lebsanft, eds. *Historical Dialogue Analysis*. Amsterdam: John Benjamins, 1999.

Juilland, Alphonse and Eugenio Chang-Rodríguez. *Frequency dictionary of Spanish words*. The Hague: Mouton & Co., 1964.

Juilland, Alphonse, P. Maximilian H. Edwards, and Ileana Juilland. *Frequency dictionary of Rumanian words*. The Hague: Mouton & Co., 1965.

Juilland, Alphonse, Dorothy Brodin, and Catherine Davidovitch. *Frequency dictionary of French words*. Paris: Mouton & Co., 1970.

Kendall, Maurice G. "A New Measure of Rank Correlation." *Biometrika* 30 (1938): 81-89.

Kilgarriff, Adam. "Which words are particularly characteristic of a text? A survey of statistical approaches." *Language Engineering for Document Analysis and Recognition (LEDAR), AISB 96 Workshop proceedings*. Eds. Lindsay J. Evett, and Tony G. Rose. Brighton, UK (April 1996a): 33-40.

Kilgarriff, Adam. "Why chi-square doesn't work, and an improved LOB-Brown comparison." *ALLC-ACH Conference*. Bergen, Norway (June 1996b).

Kilgarriff, Adam. "Using word frequency lists to measure corpus homogeneity and similarity between corpora." *Proceedings 5th ACL workshop on very large corpora*. Beijing and Hong Kong (1997): 231-245.

Kilgarriff, Adam. "Comparing Corpora." *International Journal of Corpus Linguistics* 6.1. Amsterdam: John Benjamins, 2001. 97-133.

Krenn, Brigitte and Christer Samuelsson. *The Linguist's Guide to Statistics: Don't Panic*. 19 Dec. 1997 <http://nlp.stanford.edu/fsnlp/dontpanic.pdf>.

Kretzschmar, William A., Charles F. Meyer and Dominique Ingegneri. "Uses of inferential statistics in corpus studies." *Corpus-based studies in English: papers from the seventeenth International Conference on English language research on computerized corpora (ICAME 17), Stockholm, May 15-19, 1996*. Ed. Magnus Ljung. Amsterdam: Rodopi, 1997. 167-177.

Lass, Roger. "Introduction." *The Cambridge History of the English Language: Volume III, 1476-1776*. Ed. Roger Lass. Cambridge: Cambridge University Press, 1999a.

Lass, Roger. "Phonology and Morphology." *The Cambridge History of the English Language: Volume III, 1476-1776*. Ed. Roger Lass. Cambridge: Cambridge University Press, 1999b.

Leech, Geoffrey and Roger Fallon. "Computer corpora - what do they tell us about culture?" *ICAME Journal* 16. Bergen, Norway: Norwegian Computing Centre for the Humanities, 1992. 29-50.

Leech, Geoffrey, Paul Rayson and Andrew Wilson. *Word Frequencies in Written and Spoken English: based on the British National Corpus*. London: Longman, 2001.

Lyne, Anthony A. *The vocabulary of French business correspondence*. Geneva: Slatkine, 1985.

Mahlberg, Michaela. "Clusters, key clusters and local textual functions in Dickens." *Corpora* 2.1 (2007a): 1-31.

Mahlberg, Michaela. "Corpora and translation studies: textual functions of lexis in Bleak House and in a translation of the novel into German." *La Traduzione. Lo Stato dell'Arte. Translation. The State of the Art*. Ravenna: Longo, 2007b. 115-135.

Mahlberg, Michaela. "A corpus stylistic perspective on Dickens' Great Expectations." *Contemporary Stylistics*. Eds. Marina Lambrou and Peter Stockwell. London: Continuum, forthcoming.

Markus, Manfred. "Manual of ICAMET (Innsbruck Computer-Archive of Machine-Readable English Texts)." *Innsbrucker Beitraege zur Kulturwissenschaft, Anglistische Reihe* 7. Innsbruck: Leopold-Franzens-Universitaet Innsbruck, Institut fuer Anglistik, 1999.

Markus, Manfred. "Towards an analysis of pragmatic and stylistic features in 15th and 17th century English letters." *New frontiers of corpus research: papers from the 21st International Conference on English Language Research on Computerized Corpora, Sydney, 2000*. Eds. Pam Peters, Peter Collins and Adam Smith. Amsterdam: Rodopi, 2002.

McEnery, Tony. *Swearing in English. Bad Language, Purity and Power from 1586 to the Present*. London: Routledge, 2005.

McEnery, Tony. "Keywords and Moral Panics: Mary Whitehouse and Media Censorship." *What's in a word-list? Investigating word frequency and keyword extraction*. Ed. Dawn Archer. Ashgate, forthcoming.

Mitton, Roger. "Spelling Checkers, Spelling Correctors and the Misspelling of Poor Spellers." *Information Processing & Management* 23.5 (1987): 495-505.

Mosteller, Frederick and Robert E. K. Rourke. *Sturdy statistics*. Reading, Massachusetts: Addison-Wesley, 1973.

Nelson, Gerald, Sean Wallis and Bas Aarts. *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Amsterdam: John Benjamins, 2002.

Oakes, Michael P. *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press, 1998.

Osselton, Noel E. "Spelling-Book Rules and the Capitalization of Nouns in the Seventeenth and Eighteenth Centuries." *A Reader in Early Modern English*. Eds. Mats Rydén, Ingrid Tiegen-Boon van Ostade and Merja Kytö. Frankfurt am Main, Germany: Peter Lang, 1998.

Pearson, Karl. "On the theory of contingency and its relation to association and normal correlation." *Biometric Series* 1. London: Drapers' Co. Memoirs, 1904.

Peterson, James L. "A Note on Undetected Typing Errors." *Communications of the ACM* 29.7 (1986): 633-637.

Rayson, Paul. *Wmatrix: a web-based corpus processing environment*. Computing Department, Lancaster University. 2007 <http://ucrel.lancs.ac.uk/wmatrix/>.

Rayson, Paul, Dawn Archer and Nicholas Smith. "VARD versus Word: A comparison of the UCREL variant detector and modern spell checkers on English historical corpora." *Proceedings of the Corpus Linguistics 2005 conference*. Birmingham, UK (14-17 July 2005).

Rayson, Paul, Dawn Archer, Alistair Baron, Jonathan Culpeper and Nicholas Smith. "Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora." *Proceedings of Corpus Linguistics 2007*. University of Birmingham, UK (27-30 July 2007).

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

Rayson, Paul. "From key words to key semantic domains." *International Journal of Corpus Linguistics* (forthcoming).

Read, Timothy R. C. and Noel A. C. Cressie. "Goodness-of-fit statistics for discrete multivariate data." *Springer series in statistics*. New York: Springer-Verlag, 1988.

Richardson, Malcolm. "Henry V, the English Chancery, and Chancery English." *Speculum* 55.4 (1980): 726-750.

Rissanen, Matti. "Syntax." *The Cambridge History of the English Language: Volume III, 1476-1776*. Ed. Roger Lass. Cambridge: Cambridge University Press, 1999.

Rissanen, Matti, Merja Kytö and Minna Palander-Collin, eds. *Early English in the Computer Age. Explorations in the Helsinki Corpus*. Berlin/New York: Mouton de Gruyter (Topics in English Linguistics), 1993.

Roland, Douglas, Daniel Jurafsky, Lise Menn, Susanne Gahl, Elizabeth Elder and Chris Riddoch. "Verb Subcategorization Frequency Differences between Business-News and Balanced Corpora: the role of verb sense." *Proceedings of the workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000)*. Hong Kong (1-8 Oct. 2000): 28-34.

Schmied, Josef. "The Lampeter Corpus of Early Modern English Tracts." *Corpora across the Centuries: Proceedings of the First International Colloquium on English Diachronic Corpora*. Cambridge, March 1993. Eds. Merja Kytö, Matti Rissanen, Susan Wright. Amsterdam: Rodopi, 1994.

Schneider, Peter. "Computer Assisted Spelling Normalization of 18th Century English." *New Frontiers of Corpus Research: Papers from the 21st International Conference on English language Research on Computerized Corpora, Sydney, 2000*. Eds. Pam Peters, Peter Collins and Adam Smith. Amsterdam: Rodopi, 2002. 199-211.

Scott, Mike. "PC analysis of key words – and key key words." *System* 25.2. Amsterdam: Elsevier, 1997. 233-245.

Scott, Mike. "Focusing on the text and its key words." *Rethinking language pedagogy from a corpus perspective: papers from the third international conference on teaching and language corpora*. Eds. Lou Burnard and Tony McEnery. Frankfurt: Peter Lang, 2000. 104-121.

Scott, Mike. "Comparing corpora and identifying key words, collocations, and frequency distributions through the WordSmith Tools suite of computer programs" *Small corpus studies and ELT: theory and practice*. Eds. Mohsen Ghadessy, Alex Henry and Robert L. Roseberry. Amsterdam: John Benjamins, 2001a. 47-67.

Scott, Mike. "Mapping key words to problem and solution." *Patterns of Text: in honour of Michael Hoey*. Eds. Mike Scott and Geoff Thompson. Amsterdam: John Benjamins, 2001b. 109-127.

Scott, Mike and Christopher Tribble. *Textual Patterns: keyword and corpus analysis in language education*. Amsterdam: John Benjamins, 2006.

Scragg, Donald C. *English Spelling*. Manchester: Manchester University Press, 1974.

Sinclair, John. *Corpus, concordance, collocation*. Oxford: Oxford University Press, 1991.

Sinclair, John. "A way with common words." *Out of corpora: studies in honour of Stig Johansson*. Eds. Hilde Hasselgård and Signe Oksefjell. Amsterdam: Rodopi, 1999. 157-179.

Smith, Jeremy J. *Essentials of Early English: An Introduction to Old, Middle and Early Modern English.* London: Routledge, 2005.

Spearman, Charles. "The proof and measurement of association between two things." *American Journal of Psychology* 15 (1904): 72-101.

Stubbs, Michael. *Text and corpus analysis: computer-assisted studies of language and culture*. Oxford: Blackwell, 1996.

Taavitsainen, Irma and Päivi Pahta. "Corpus of Early English medical writing 1375–1750." *ICAME Journal* 21 (1997): 71–81.

Taavitsainen, Irma and Päivi Pahta, eds. *Medical Writing in Early Modern English.* Cambridge: Cambridge University Press, forthcoming.

Taavitsainen, Irma, Päivi Pahta, Turo Hiltunen, Martti Mäkinen, Ville Marttila, Maura Ratia, Carla Suhr and Jukka Tyrkkö. *Early Modern Medical Texts*. forthcoming.

Tribble, Christopher. "Genres, keywords, teaching: towards a pedagogic account of the language of project proposals." *Rethinking language pedagogy from a corpus perspective: papers from the third international conference on teaching and language corpora*. Eds. Lou Burnard and Tony McEnery. Frankfurt: Peter Lang, 2000. 75-90.

Tribble, Christopher and Glyn Jones. *Concordances in the classroom*. Houston, Texas: Athelstan, 1997.

Vallins, George H. and Donald G. Scragg. *Spelling*. London: André Deutsch, 1965.

Virtanen, Tuija. "The progressive in NS and NNS student compositions: evidence from the International Corpus of Learner English." *Corpus-based studies in English: papers from the seventeenth International Conference on English language research on computerized corpora (ICAME 17), Stockholm, May 15-19, 1996*. Ed. Magnus Ljung. Amsterdam: Rodopi, 1997. 299-309.

Voutilainen, Atro and Juha Heikkilä. "An English Constraint Grammar (ENGCG) a surface-syntactic parser of English." *Creating and Using English Language Corpora. Papers from the 14th International Conference on English Language Research on Computerized Corpora*, *Zürich, 1993*. Eds. Udo Fries, Gunnel Tottie and Peter Schneider. Amsterdam: Rodopi, 1994. 189-199.

Williams, Raymond. *Keywords: a vocabulary of culture and society.* 2nd ed. London: Fontana Press, 1983.

Wikberg, Kay. "The style marker as if (though): a corpus study." *Out of corpora: studies in honour of Stig Johansson*. Eds. Hilde Hasselgård and Signe Oksefjell. Amsterdam: Rodopi, 1999. 93-105.

Woods, Anthony, Paul Fletcher, and Arthur Hughes. *Statistics in language studies*. Cambridge: Cambridge University Press, 1986.

Yates, Frank. "Contingency tables involving small numbers and the chi-squared test." *Journal of the Royal Statistical Society Supplement* 1 (1934): 217-235.

Yates, Frank. "Tests of significance for 2 × 2 contingency tables." *Journal of the Royal Statistical Society, A* 147.3 (1984): 426-463.

Yule, George. *The Statistical Study of Literary Vocabulary*. Cambridge: Cambridge University Press, 1944.