Variation in noun and pronoun frequencies in a sociohistorical corpus of English

Säily, Tanja

2011

Säily , T , Nevalainen , T & Siirtola , H 2011 , ' Variation in noun and pronoun frequencies in a sociohistorical corpus of English ' , Literary and Linguistic Computing , vol. 26 , no. 2 , pp. 167-188 . https://doi.org/10.1093/llc/fqr004

http://hdl.handle.net/10138/39576
https://doi.org/10.1093/llc/fqr004

acceptedVersion

Variation in noun and pronoun frequencies in a sociohistorical corpus of English

Note: This is a post-print version of the article published in 2011 in *Literary and Linguistic Computing* 26(2): 167–188, http://dx.doi.org/10.1093/llc/fqr004

Tanja Säily
University of Helsinki, Finland
Terttu Nevalainen
University of Helsinki, Finland
Harri Siirtola
University of Tampere, Finland

Correspondence:
Tanja Säily, P.O. Box 24, FI-00014 University of Helsinki, Finland
E-mail:
tanja.saily@helsinki.fi

Abstract

Many corpus linguists make the tacit assumption that part-of-speech frequencies remain constant during the period of observation. In this article, we will consider two related issues: (1) the reliability of part-of-speech tagging in a diachronic corpus, and (2) shifts in tag ratios over time. The purpose is both to serve the users of the corpus by making them aware of potential problems, and to obtain linguistically interesting results. We use noun and pronoun ratios as diagnostics indicative of opposing stylistic tendencies, but we are also interested in testing whether any observed variation in the ratios could be accounted for in sociolinguistic terms. The material for our study is provided by the Parsed Corpus of Early English Correspondence (PCEEC), which consists of 2.2 million running words covering the period 1415–1681. The part-of-speech tagging of the PCEEC has its problems, which we test by reannotating the corpus according to our own principles and comparing the two annotations. While there are quite a few changes, the mean percentage of change is very small for both nouns and pronouns. As for variation over time, the mean frequency of nouns declines somewhat, while the mean frequency of pronouns fluctuates with no clear diachronic trend. However, women consistently use more pronouns than men, while men use more nouns than women. More fine-grained distinctions are needed to uncover further regularities and possible reasons for this variation.

1. Introduction

In the last twenty years, English historical linguistics has developed a strong corpus linguistic orientation. Since the release of the Helsinki Corpus of English Texts (HC) in 1991, a number of diachronic corpora have been made available, and new ones are in the process of being compiled. Historical corpus linguists are typically interested in analysing language change, and comparing the frequencies of linguistic features over time in different genres and areal or regional varieties. Multigenre corpora such as the HC naturally strive for genre continuity, although long-term diachronic succession can only be achieved for a few prototypical genres.[1] The time period to be covered by English diachronic corpora, twelve hundred years of recorded history, creates challenges not only for genre continuity but also for grammatical annotation. Many of the corpora have been tagged for parts of speech, and some have also been parsed. To some extent, different annotation schemes are needed for Old English, an inflected language, and later periods. The material analysed in this study, the Parsed Corpus of Early English Correspondence (PCEEC; Santorini, 2010), uses the same basic scheme as the Penn-Helsinki Parsed Corpus of Middle English (PPCME2) and of Early Modern English (PPCEME).

Grammatically annotated diachronic corpora are relatively recent and have been used much less extensively than their original plain-text versions. Users of unannotated corpora, both present-day and historical, tend to share the 'null hypothesis' discussed in Mair *et al*. (2002, p. 248) that part-of-speech frequencies have remained constant within the period of observation. A more refined version of this null hypothesis is that part-of-speech frequencies have remained constant in the period of observation in the genre(s) studied. This is the basis for interpreting temporal changes in frequencies of a particular linguistic feature as an indication of a language change in progress. The increase in the frequency of the progressive, for example, has been discussed in a number of studies as an autonomous change in progress rather than a corollary of, say, the increased use of verbs in written genres over the last two or three centuries (for the frequency of progressives in various corpora, see e.g. Nesselhauf, 2007, pp. 193–194, 196).

However, historical linguists are the first to admit that genres change over time. While chronicles and homilies have long since lost their pre-eminence among written genres in English, newspapers and novels are relative newcomers. Moreover, research into genre-internal variation has revealed long-term shifts in the linguistic properties of written genres. Biber and Finegan (1989, 1997) found that drama, diaries, and fiction have become increasingly associated with speaker involvement over the last three centuries, displaying higher frequencies of features like private verbs, contractions, *that*-deletion, and personal, demonstrative, *wh*-, and indefinite pronouns. By contrast, science, news, and legal genres have diverged in the informational dimension, with increased frequencies of features such as nouns, long words, and a high type/token ratio.

In this article we will consider two related issues: (1) the reliability of part-of-speech (POS) tagging in a diachronic corpus, and (2) potential shifts in tag ratios over time. Reliable tagging is the prerequisite for the assessment of diachronic shifts in the distribution of word classes, which would undermine the 'null hypothesis' discussed above. We use noun and pronoun ratios as diagnostics indicative of opposing stylistic tendencies. The material for our study is provided by the Parsed Corpus of Early English Correspondence, which consists of 2.2 million running words and covers the period from the early 15th century to the last decades of the 17th century. In A

Representative Corpus of Historical English Registers (ARCHER, 1650–1990), studied by Biber and Finegan (1997), the letter genre was found to move in the more involved direction over time. We are interested in how stable the genre was in the preceding centuries.

However, we are also interested in testing whether any observed variation in POS ratios could be accounted for in sociolinguistic terms. The sociolinguistic variables encoded in the PCEEC metadata include the correspondents' gender. Noun and pronoun scores will be correlated with information on male/female writers and male/female addressees. This analysis was prompted by studies of Present-day English suggesting that noun and pronoun ratios systematically correlate with gender differences (Rayson *et al*., 1997; Argamon *et al*., 2003). Previous studies on the Corpus of Early English Correspondence support the idea of testing the gender variable, which shows systematic differences in processes of language change (e.g. Nevalainen and Raumolin-Brunberg, 2003).

We recognize the general methodological challenge of gaining insight into text corpora and the rich numerical data they can yield. Besides presenting information in a tabular form, we will use some information visualization methods (Spence, 2007) to explore and illustrate various aspects of the PCEEC. These methods are useful in both exploratory and confirmatory data analysis. For the former, we mainly use the Mondrian interactive data analysis tool (Theus and Urbanek, 2008), and for the latter, the statistical data language and environment R (R Development Core Team, 2010).

The rest of this article is organized as follows. Section 2 surveys previous work on POS ratios in English corpora and on annotating historical corpora, while Section 3 describes our material, the PCEEC. Section 4 analyses the reliability of POS tagging in the PCEEC, and Section 5 extends the analysis to variation in POS ratios. Section 6 discusses possible reasons for the variation, pointing out directions for future research. Finally, Section 7 summarizes our views on the reliability and stability of the corpus.

2. Background

2.1 Findings from Present-day English

Many studies of noun ratios in Present-day English corpora refute the claim made in Hudson (1994) that about 37% of all word-tokens in a corpus are nouns (such as Biber *et al*., 1999; Mair *et al*., 2002; Rayson *et al*., 2002; Hardie, 2007). The arguments against it arise from the problem of defining the category of nouns, and difficulty in comparing noun ratios across different annotation schemes. Hudson's noun category is a liberal one and comprises common and proper nouns as well as pronouns. Apart from combining parts of speech which are usually considered and tagged as separate word classes, the decision to include nouns and pronouns in one superordinate category would present problems for genre comparisons.

Not only are category combinations potentially problematic, but Hardie (2007) shows how analysing the superordinate, 'first-letter' categories of noun (N) and pronoun (P) can also conceal a good deal of variation, depending on the tag groups included. The status of interrogative adverbs, numerals, and determiners, in particular, can be debated. However, deciding on the content of a category is not always a matter of the researcher's choice but depends on the tagset of the annotation scheme used as well. In the Brown family of corpora, different CLAWS tagsets which differ with respect to some subcategories of these superordinate categories have been used over the years.

Changes have been introduced, for example, to the distinction between pronouns and determiners, and the interpretation of the possessive suffix, which has changed from being marked as an inflection to being treated as a clitic. This change means that it is tokenized as a word of its own, and thus affects the total word count of the corpus. Tokenization is, of course, influenced by the entire tagset, and the decisions taken on how to treat such things as multi-word units. As Hardie (2007, p. 71) points out, the comparison of POS-tag ratios in tagged corpora is complicated not only by the differences in their (sub)classification schemes but also the implications that diverse tagsets have for tokenization.

Despite these problems of comparing POS-tag ratios across data sets – and the impossibility of comparing their precise values – a number of studies indicate that the superordinate categories of nouns and pronouns pattern differently depending on genre. Biber *et al*. (1999, pp. 65, 92) analyse the Longman Corpus, showing that news is characterized by a higher proportion of nouns than conversation, fiction, and even academic prose, whereas conversation scores highest in pronoun frequency, and academic prose lowest. Rayson *et al*. (2002, pp. 301–302) report similar findings on the British National Corpus Sampler, nouns being more frequent in writing than in speech, and more frequent in informative writing than in imaginative writing. Pronouns, by contrast, are found to be more common in speech than in writing, especially in conversational as opposed to task-oriented speech, and more common in imaginative than in informative writing. Hudson (1994, pp. 332–335), too, reports a range of variation in his noun category according to genre and medium.

Comparing the frequencies of noun tags in the LOB and Freiburg-LOB corpora, Mair *et al*. (2002) find a systematic rise in the use of the noun category over time, matched by a drop in the overall frequency of pronouns. The four subdivisions of the corpora – press, general prose, learned texts, and fiction – all show a similar increase in the use of nouns. This trend is not paralleled by a corresponding decrease in the frequency of verbs. However, the fact remains that the overall increase observed does not mask genre differences, fiction showing a much lower frequency of nouns than non-fiction, and news in particular (Mair *et al*., 2002, p. 255).

Noun and pronoun frequencies have also been associated with gendered styles. In Present-day British English, as represented by the British National Corpus (BNC), women have been shown to use fewer nouns and more personal pronouns than men. Rayson *et al*. (1997) show this for conversation; even in formal written texts, however, women seem to use more personal pronouns, while men use more nouns and certain types of noun specifiers (Argamon *et al*., 2003). According to Argamon *et al*. (2003, p. 321), these findings lend support to the notion that men's style is more 'informational' and women's more 'involved'.

2.2 Issues with Historical Data

Automated and semi-automated tagging has reached a high level of precision in Present-day English corpora – Rayson *et al*. (2008: 32) note that the accuracy of the CLAWS tagger on standard English is around 97–98%. In historical corpora, however, the process is far more complex and error-prone. One reason for this is the enormous degree of spelling variation in pre-standard English, an example of which is provided by this (non-exhaustive) list of spellings of the word *tomorrow* in the PCEEC: *to marrow*, *to moroe*, *to moroughe*, *to morow*, *to morowe*, *to morroughe*, *to morrow*, *to morrowe*, *to*

morue, *to morwe*, *to-morow*, *to-morowe*, *to-morrow*, *to-morrowe*, *to-morw*, *to-morwe*, *tomorow*, *tomorowe*, *tomorrow*, *tomorrowe*, *too morrow*, *toomorrow*. A tagger will find it difficult to identify the part of speech of a non-standard spelling variant not found in its lexicon; furthermore, some of the variants are spelled as two words, which affects tokenization and further complicates the tagger's work. Some of the spellings also overlap with those of other words: the words *to* and *too* could be spelled inter-changeably.[2]

In some cases, variant spellings are due to a process of grammaticalization. For instance, the earliest (Old English) instances of the item *today* written separately were indeed two separate words, a preposition and a noun – the prepositional phrase was later grammaticalized into an adverb. Even if the grammaticalization process is over, the same word may occur in more than one part of speech: in both Modern English and Present-day English, *today* can be either an adverb or a noun (OED, s.v. *today*). These period- and context-dependent issues need to be addressed by the annotation scheme.

What kinds of annotation system have been used, then? According to Rayson *et al*. (2007, p. 4), labour-intensive seems to be the keyword here. Originally designed for present-day texts, the TreeTagger, the English Constraint Grammar Parser (ENGCG), and the Penn Treebank have all been applied to historical data with some modifications and much manual post-editing. Biber and Finegan (1989) are among the pioneers of annotating historical texts, using a tagger developed by Biber for annotating the LOB corpus before the official tagged version became available (1988: Appendix II). Durrell *et al*. (2007) adapt the TreeTagger for early German newspaper texts, achieving an accuracy of c. 80%, after which the tagging is checked manually.

Kytö and Voutilainen (1998) get good initial results with ENGCG in Early and Late Modern English correspondence by simply augmenting its lexicon and adding a few rules to its grammar. Their new lexical entries include several possible morphological descriptions for spelling variants such as *ther* (which could represent either *there* or *their*), abbreviations common in letters, as well as obsolete and nonce words. Kytö and Voutilainen (1998, p. 165) discover that the parser makes by far the most mistakes with the earliest period, but it remains unclear how much additional effort would be required to successfully tag a corpus of Early Modern English using ENGCG.

While the work described above was carried out on small pilot corpora, complete historical corpora have also been tagged and parsed. For instance, the Penn Historical Corpora, including the PPCME2, PPCEME, and the new Penn Parsed Corpus of Modern British English (PPCMBE), have all been annotated by adapting guidelines originally developed for the Penn Treebank Project (Santorini, 1990, 2010). The same applies to the York–Toronto–Helsinki Parsed Corpus of Old English Prose (YCOE), a sister corpus to the PPCME2; however, quite a few changes to the annotation scheme have been necessary 'due to the inflected nature of Old English' (Taylor, 2003, Introduction). Indeed, it is questionable whether a single scheme could even in principle be constructed that could be used for the entire history of the English language. Another member of the English Parsed Corpora Series is the PCEEC, which is described in more detail in the next section.

3. Material

In this study, we use the Parsed Corpus of Early English Correspondence (PCEEC). The PCEEC is a published version of the Corpus of Early English Correspondence (CEEC), which is described in Section 3.1. More information on the PCEEC is given in Section 3.2, while Section 3.3 provides a brief description of its annotation scheme.

3.1 Description of the CEEC

The CEEC, a corpus of personal letters written in English (as used in England), was compiled in the 1990s by the 'Sociolinguistics and Language History' project team at the University of Helsinki. It was designed with historical sociolinguistics in mind: the genre – which in certain respects can be regarded as close to spoken interaction – was kept constant, and the sampling unit was the individual letter writer. The aim was to include writers of both genders and all social ranks from each successive 20-year period covered by the corpus, which spans the years 1415–1681.[3] Regional coverage was also taken into account.

The people living during these centuries can be divided socially in various ways (see Nevalainen, 1996; Nevalainen and Raumolin-Brunberg, 2003, Ch. 7). Despite efforts to create a socially balanced corpus, the dominance of men from the upper ranks was unavoidable as they were the most literate group, were considered important enough that their letters were preserved, and their letters were later considered important enough to be published. Temporal coverage is likewise uneven, with more material from the later periods than the earlier ones.

The letters in the corpus were selected from printed editions, digitized, and proofread by the team; in a fair number of cases, it was possible to check the letters against the originals in various archives and libraries. Good original-spelling editions were preferred, but to achieve a better coverage of women and the lower ranks, a few less reliable editions had to be included. For more information on the CEEC, see Nevalainen and Raumolin-Brunberg (2003), Raumolin-Brunberg and Nevalainen (2007), Nurmi *et al*. (2009), and the entry for the CEEC in the Corpus Research Database (CoRD).[4]

3.2 PCEEC

Published in 2006, the PCEEC is a part-of-speech tagged and syntactically parsed version of those letter collections in the CEEC for which permission to publish could be obtained, amounting to over three-quarters of the original CEEC. The part-of-speech tagging was carried out by Arja Nurmi at the University of Helsinki, and the syntactic annotation by Ann Taylor at the University of York. The PCEEC consists of 4,969 letters written by c. 660 informants between 1415 and 1681, with a total word count of approximately 2.15 million words. The corpus comes in three different formats: plain text files, part-of-speech (POS) tagged files, and syntactically parsed files. For this study, we use the POS tagged files, together with external databases of sociolinguistic information on the letters and the correspondents. These databases are an extended version of the Associated Information File provided with the corpus (for the latter, see Taylor and Santorini, 2006). Because they are still under construction and contain many

parameters subject to interpretation, they are currently in use by members of the CEEC team only.

Being based on the original CEEC, the PCEEC is not a perfectly balanced corpus with respect to social and temporal factors. As can be seen from the statistics below, there is more data from men than women (Table 1), more data from the upper ranks than the middle and lower ranks (Table 2), and more data from the later periods than the earlier periods (Table 3). This presents challenges in comparing the effects of these factors, and the small amounts of data from some groups may prevent a thorough analysis of their language use.

Table 1 Proportion of men vs. women in the PCEEC split according to gender of letter sender and recipient

Gender/ sender	**Gender/ recipient**	**Words**		**Letters**		**Informants**	
Men	Men	1,503,359	(83%)	3,421	(83%)	465	
	Women	299,634	(17%)	701	(17%)	124	
	Unknown	1,186	(0%)	5	(0%)	4	
Total		1,804,179	(84%)	4,126	(83%)	516	(78%)
Women	Men	286,224	(81%)	647	(77%)	114	
	Women	66,148	(19%)	192	(23%)	46	
	Unknown	1,022	(0%)	3	(0%)	2	
Total		353,394	(16%)	842	(17%)	143	(22%)
Grand total		2,157,573		4,969		659	

Table 2 Proportion of different ranks in the PCEEC

Rank	**Words**		**Letters**		**Informants**	
Royalty	74,404	(3%)	295	(6%)	17	(2%)
Nobility	431,362	(20%)	1,004	(20%)	103	(16%)
Gentry	1,018,614	(47%)	2,249	(45%)	281	(43%)
Clergy	273,472	(13%)	618	(13%)	95	(14%)
Professionals	231,215	(11%)	460	(9%)	78	(12%)
Merchants	86,124	(4%)	229	(5%)	33	(5%)
Other	35,951	(2%)	98	(2%)	44	(7%)
Unknown	6,431	(0%)	16	(0%)	8	(1%)
Total	2,157,573		4,969		659	

Table 3 Proportions of time periods in the PCEEC. Informants who have letters from more than one period are counted each time they occur.

Period	**Words**		**Letters**		**Informants**	
1415–1453	74,645	(3%)	212	(4%)	40	(5%)
1454–1491	295,717	(14%)	745	(15%)	133	(18%)
1492–1529	95,710	(4%)	236	(5%)	63	(9%)
1530–1567	226,374	(11%)	461	(9%)	97	(13%)
1568–1605	426,610	(20%)	914	(19%)	143	(20%)
1606–1643	559,578	(26%)	1,390	(28%)	160	(22%)
1644–1681	478,939	(22%)	1,011	(20%)	93	(13%)
Total	2,157,573		4,969		729	

A diachronic corpus can be split into periods on many grounds (see Gries and Hilpert, 2008). For this paper, we chose to use fixed-length periods and determined the length of a period, 38 years, as a result of experimentation. This periodization is a trade-off between the 20-year one commonly used in sociolinguistics and the above-mentioned imbalance of the PCEEC. Figure 1 is a mosaic plot (Hartigan and Kleiner, 1984), or an area-dividing visualization, that illustrates the issues in the PCEEC: the large number of note-like, short text samples in the first period, the relatively low number of samples from the first and third periods, and the overall imbalance of genders.

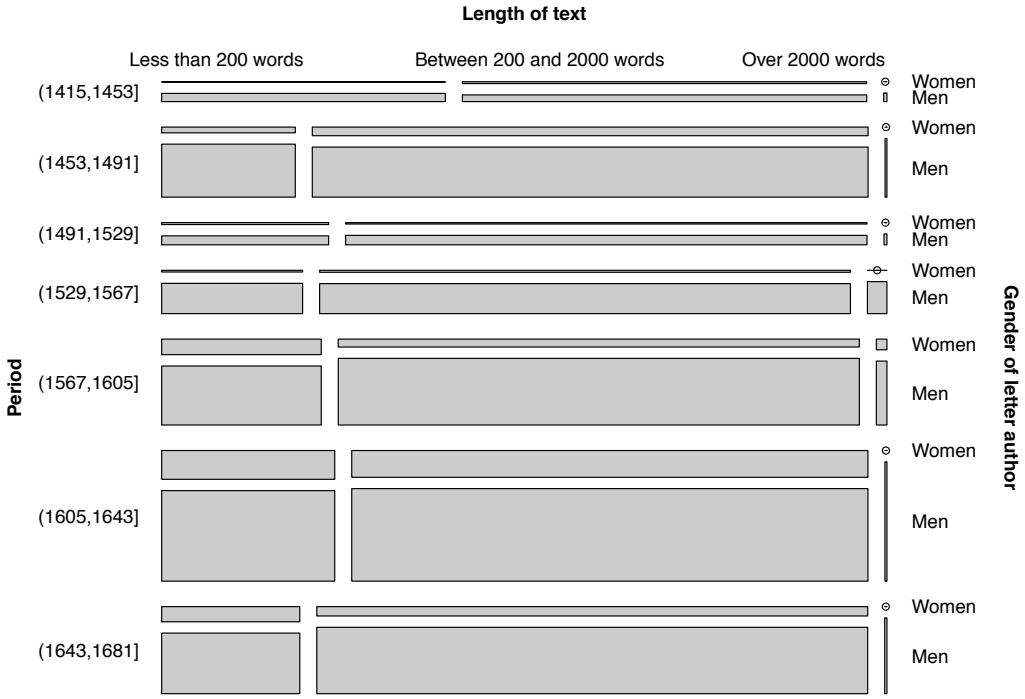

Fig. 1 Mosaic display of PCEEC texts split according to time period, length of text, and gender of letter writer

3.3 Annotation Scheme

The annotation scheme for the PCEEC is the same as that used by the Penn-Helsinki Parsed Corpus of Middle English (2nd edition) and the Penn-Helsinki Parsed Corpus of Early Modern English, with slight changes documented in the manual (Taylor and Santorini, 2006). All of these corpora are part of the English Parsed Corpora Series developed for the use of historical syntacticians (Taylor, 2007). The developers' own interests have further shaped the annotation scheme, so that the focus is on sentential syntax; POS tagging is mainly seen as a necessary step before parsing, and lemmatization or normalization of the lexis is not even considered. Nevertheless, this series of corpora is an unparalleled resource for historical linguists interested in the long diachrony of English.

Initial POS annotation was added to the PCEEC automatically using the Brill tagger, the accuracy of which was c. 80–90% (Arja Nurmi, private communication).

The latter figure was approached as the tagger was trained, but the results still required extensive manual post-editing. In the POS tagged version of the corpus, each word is suffixed with an underscore followed by a code for part of speech, as in (1). As mentioned above, the words are not normalized or lemmatized. Punctuation is separated from the words and given its own tags. Editors' comments, header information, and other non-text materials are given the tag CODE.

(1) Yow_PRO say_VBP som_Q were_BED in_P dignity_N at_P home_N ;_, to_P whome_WPRO I_PRO promis_VBP that_C their_PRO$ lyving_VAG here_ADV shall_MD be_BE as_ADVR correspondent_ADJ <P_9>_CODE to_P their_PRO$ quality_N and_CONJ degree_N in_P England_NPR ,_, as_P that_D that_C they_PRO have_HVP in_P Loven_NPR {ED:Louvain}_CODE ;_.

According to Taylor and Santorini (2006), the 'primary goal has been to create an annotation system that facilitates automated searches, not to give a correct linguistic analysis of each sentence'. Furthermore, the designers 'have avoided making decisions that would be controversial, whether with regard to text interpretation or to linguistic theory' (Corpus annotation: General introduction). The implications of this cautious attitude for the categorization of nouns and pronouns are discussed in the next section.

4. Analysis of the Reliability of POS Tagging in the PCEEC

4.1 Nouns

The PCEEC manual lists the following noun tags: N (common noun, singular), N$ (common noun, singular, possessive), NPR (proper noun, singular), NPR$ (proper noun, singular, possessive), NPRS (proper noun, plural), NPRS$ (proper noun, plural, possessive), NS (common noun, plural), and NS$ (common noun, plural, possessive). Only two tags, PRO and PRO$ (possessive), were listed for personal pronouns, with a note that reflexive pronouns had been tagged using the compound tags PRO+N or PRO$+N. For nouns, we found a number of other compound tags in the corpus; see Appendix 1 for a list of all noun tags. Our initial classification of nouns included the simplex tags listed above as well as all compound tags whose last component was one of the simplex noun tags, as the last component usually determines the part of speech in English compounds (Plag, 2003, p. 135).

We found that many adverbs in the PCEEC had been conservatively tagged as nouns, and that the word 'self' in reflexive pronouns had been tagged as a noun when the compound was written as two words, as often happened in the early part of the corpus. To resolve these and other issues of categorization and tokenization, we created our own version of the corpus, ReCEEC. The rules for producing the revised version of the corpus were eventually boiled down to a relatively simple Python script. The most difficult task turned out to be pruning the class of nouns; hence, it is described in detail in Appendix 2. As most of the compound tags were relatively rare, there was a large number of minor changes; the discussion in the appendix focuses on the most frequent types of change (the frequency of each type being given in parentheses). Changes to the tokenization of the corpus are discussed further in Section 4.3.

4.2 Categorization

Figure 2 shows the most frequent changes made to the categorization of nouns and pronouns in the ReCEEC. As discussed in Appendix 2, the major changes are from nouns (N, N$, NS) to pronouns (PRO, PRO$), adverbs (ADV), 'quantifiers' (Q), and subordinating conjunctions (P). A number of changes also involve tokenization: nouns written as two words are combined (N N → N, e.g. *lord ship*), articles and nouns written together are separated (N → D N, e.g. *th'enquest*), etc. Even though thousands of changes have been made, the mean change is only -0.19% to the class of nouns and +0.15% to pronouns. The changes to other parts of speech are generally even smaller. Nevertheless, as the ReCEEC is a somewhat better match to Present-day English corpora than the original PCEEC, it is used for the analyses in this paper.

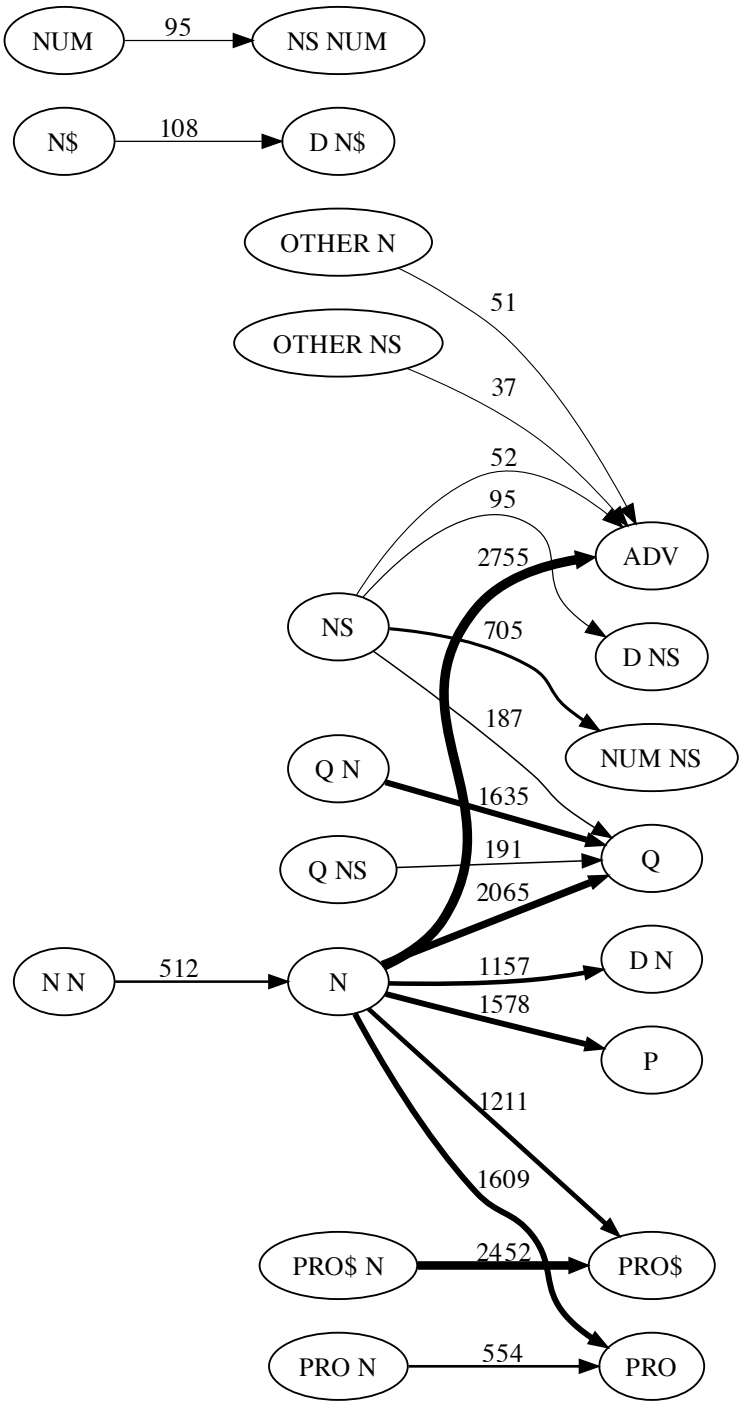

Fig. 2 The most frequent changes involving nouns and pronouns in the ReCEEC. Compound tags are collapsed into their last component (e.g. N includes both the tag N and all compound tags ending in N).

4.3 Tokenization

Even though the ReCEEC was edited extensively, its tokenization still does not match that of Present-day English corpora (which do not always agree among themselves, either). To ensure PDE tokenization, it would have been necessary to read the corpus word by word to spot all compounds written separately and all separate words written together. As this was obviously impossible, we decided to take a manageable sample to evaluate how much the tokenization-related spelling in the PCEEC differed from PDE spelling, and to compare the results with the ReCEEC.

The sample consisted of 20 letters from each of these three 20-year periods (60 letters in all): 1450–1470, 1570–1590, and 1660–1680. The periods were selected to represent maximally different phases in English spelling, keeping in mind the amount of data in the corpus from each subperiod. For each period, we randomly sampled five letters from women and fifteen from men. This was done to ensure that women had sufficient representation in each period, because it was to be expected that, being in general less well educated than men, they might use more variant spellings. The 60 letters were read by a research assistant who noted down all instances of compounds written separately and separate words written together. The word counts for each period and gender were then adjusted and compared with the original counts to determine the amount of change. The results of this analysis are shown in Table 4. As the number of instances was small, calculating statistical significance was not attempted.

Table 4 Effect of spelling variation on word counts: samples from three centuries

Period	**Gender**	**Word count**	**Compounds written separately**		**Separate words written together**		**Corrected word count**	**Change**
1450–1470	Men	7,338	28	(0.38%)	9	(0.12%)	7,319	-0.26%
	Women	2,981	11	(0.37%)	3	(0.10%)	2,973	-0.27%
	Total	10,319	39	(0.38%)	12	(0.12%)	10,292	-0.26%
1570–1590	Men	7,462	45	(0.60%)	7	(0.09%)	7,424	-0.51%
	Women	1,304	14	(1.07%)	0	(0.00%)	1,290	-1.07%
	Total	8,766	59	(0.67%)	7	(0.08%)	8,714	-0.59%
1660–1680	Men	8,725	18	(0.21%)	0	(0.00%)	8,707	-0.21%
	Women	1,908	4	(0.21%)	0	(0.00%)	1,904	-0.21%
	Total	10,633	22	(0.21%)	0	(0.00%)	10,611	-0.21%
Total	**Men**	23,525	91	(0.39%)	16	(0.07%)	23,450	-0.32%
	Women	6,193	29	(0.47%)	3	(0.05%)	6,167	-0.42%
	Grand total	29,718	120	(0.40%)	19	(0.06%)	29,617	-0.34%

As is evident from Table 4, the changes are generally very small, < 1%. There seem to be some differences between the periods: 16th-century data undergoes the most change, followed by the 15th and 17th centuries. There is not much difference between men and women, save perhaps for the middle period, in which women's letters seem to contain more compounds written separately. During the first period, women's letters were usually written by male scribes, which explains the lack of difference there. Overall, compounds written separately are more common than separate words written together. There are no separate words written together in the 17th-century data, which may reflect the stabilization of spelling in that respect.

The most common 15th-century word written separately is *methinks*, which – like many of these words – was still in the process of grammaticalization at the time. Other common compounds include *therein* and *nothing*. Frequent 16th-century words written separately include *myself*, *anything*, and *whereunto*, and in the 17th-century data *anything* and reflexive pronouns continue to dominate. The most common separate words written together in the first two periods include *asmuch*, *aswell*, and *nomore*, as well as articles written together with their head, such as *th'enquest*.

The tokenization of many of the above-mentioned items is normalized in the ReCEEC, bringing the word count of the 60 letters to 29,647, which is somewhere in between the original and the manually corrected count. Some of the compounds written separately in the PCEEC are marked as belonging together in its tagging (see Taylor and Santorini, 2006, 'Items treated as unitary'). If only these items are combined, the word count becomes 29,678, which is already an improvement on the original. Therefore, PCEEC users who intend to compare their results with PDE data may wish to retokenize the corpus by combining these items, which can be done automatically.

5. Analysis of Shifts in Tag Ratios

5.1 Overview

Leech and Smith (2005, p. 89) suggest that genre evolution is especially important within the twentieth century, because it was a period characterized by rapid social and stylistic change. We are now in a position to assess the degree to which this was also the case with the earlier centuries in personal letters.

Both tables and visual representations are used to show the variation in noun and pronoun proportions and distributions over the time span of the PCEEC. We use *beanplots* (Kampstra, 2008) for visual summaries and to facilitate comparisons between time periods and data groups. A beanplot is a combination of a mirrored density trace and an embedded 1-dimensional scatter plot, also known as a strip plot. A small line presents each observation in a group of data, or a batch, and multiple observations with the same value are over-plotted to appear darker. Overall average (dashed line) and the average of a batch (solid line extending from density plot) are drawn, allowing comparisons between the batches.

First, the proportions of nouns and pronouns are characterized by beanplots in Figs. 3 and 4, without background variables. The numeric values of means, medians, and quintiles of noun distributions can be found in Tables 5 and 6 in Section 5.2. The imbalances in the data shown in Fig. 1 can also be detected in Fig. 3: the number of observations (each representing the noun ratio of a letter) is much smaller in the first and third periods than in the others, evidenced by the sparseness of the small lines in the scatter plot. Hence, the results from these periods may be less reliable, which should be kept in mind throughout the analysis. There was also an outlier in the last period whose letters skewed the results so much that her data was removed from the data set. This was Dorothy Osborne, a gentlewoman writing to her future husband in the 1650s, who has proved to be an outlier in terms of nearly every change studied in the CEEC.

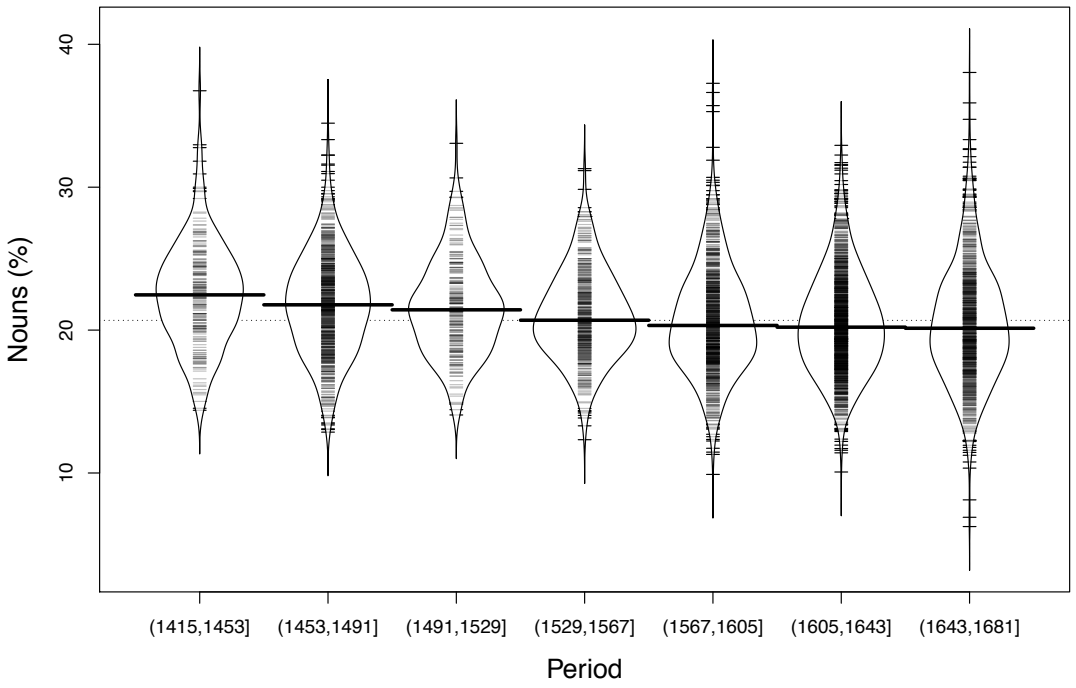

Fig. 3 Variation in noun ratio over time

Figure 3 suggests that the proportion of nouns either declines or remains the same towards the end of the era, without exception. Post-hoc pair-wise comparison of noun percentages (Tukey contrasts with corrections, package *multcomp* in R) shows that whenever two time periods have at least two periods between them, the difference is statistically significant ($p < .001$) except for the (1643,1681] period. This is a compelling demonstration that there indeed is a solid declining trend in the proportion of nouns within this time period. In the first period, the mean and median proportion of nouns is 22.5%, which steadily declines to the value of 20.4% observed in the last period (Table 5).

Variation in pronoun ratio seen in Fig. 4 is more complex than the variation in nouns. Again, post-hoc pair-wise comparison with Tukey contrasts and corrections reveals differences between adjacent time periods that are statistically significant except for the periods (1529,1567] – (1567,1605]. However, there is no clear trend in the variation. Even if the unreliable first and third periods (the latter of which is now bimodal) are ignored, the picture does not become any clearer at this general level of analysis.

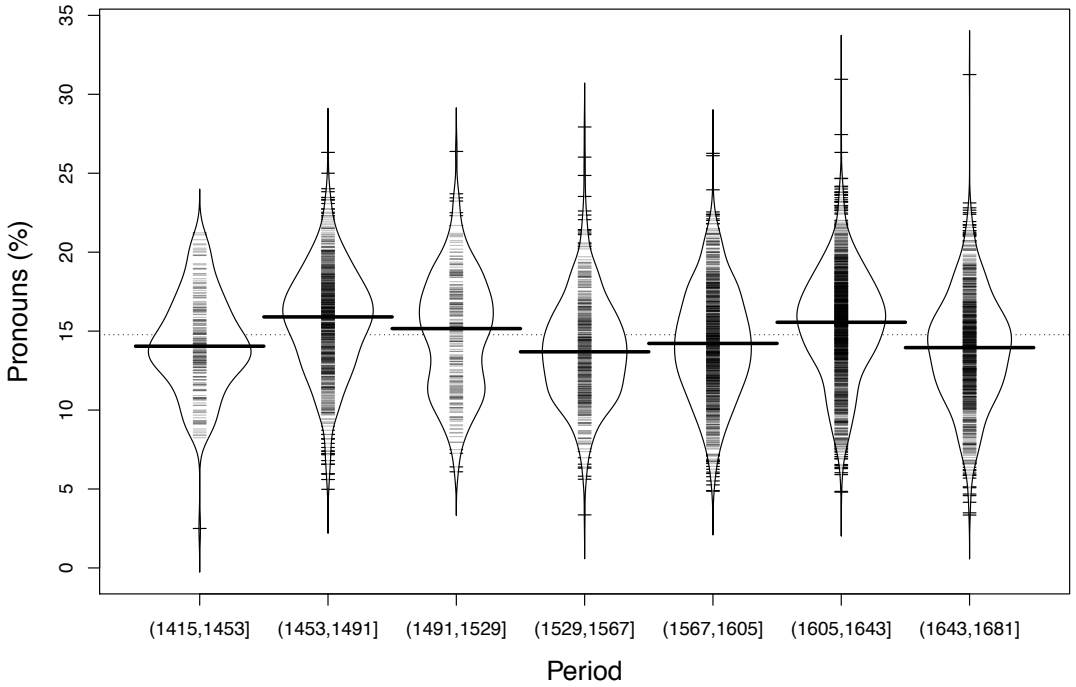

Fig. 4 Variation in pronoun ratio over time

5.2 Sociolinguistic Variation

As noted in Section 2.1, women use fewer nouns and more personal pronouns than men in Present-day British English (Rayson *et al*., 1997; Argamon *et al*., 2003). Figures 5 and 6 show that the same seems to hold for historical English, since in every subperiod of our corpus, women's letters contain more pronouns and fewer nouns than men's. The difference in the proportion of nouns is statistically significant ($p < .05$) for all periods except (1529,1567], where the difference of means is 0.5%. Although the data is unevenly distributed across periods and genders, with especially sparse and peculiarly distributed observations from women in the third and fourth periods, the tendency is remarkably consistent. In the case of the pronoun proportion, there is even more variation in the distributions, but the differences between men and women are statistically significant ($p < .005$) across the board.

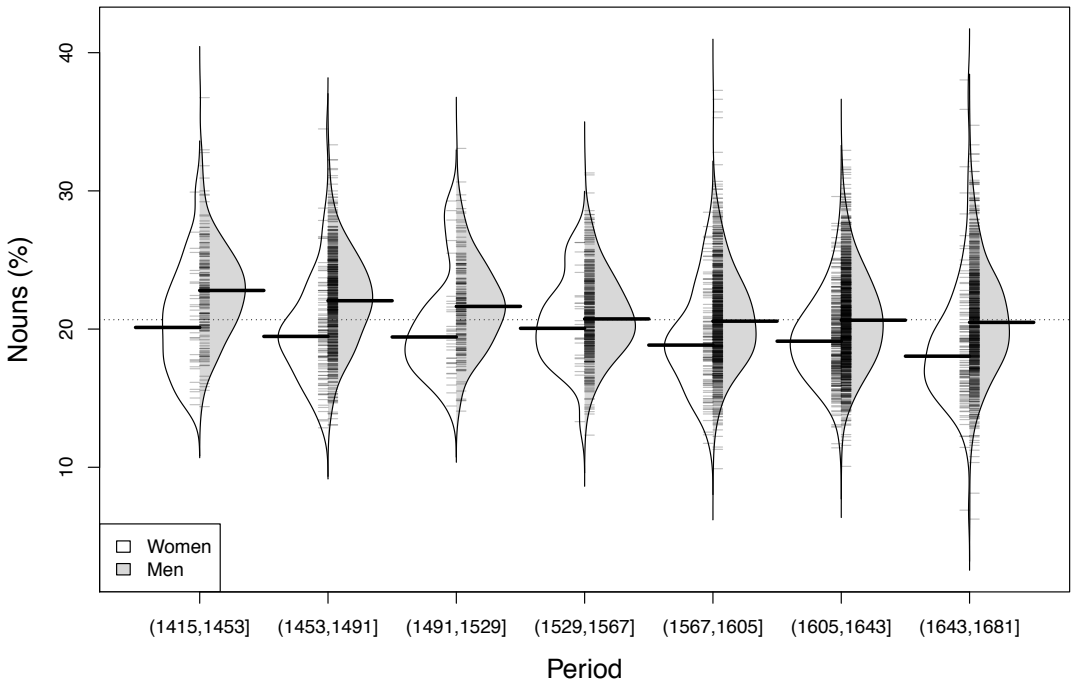

Fig. 5 Sociolinguistic variation in noun ratio over time: gender

Table 5 Proportion of nouns in different time periods split according to gender of letter writer

			Quintiles				
Period	**Gender**	**Mean**	**0%**	**25%**	**50%**	**75%**	**100%**
1415–1453	Women	20.5%	14.5%	17.7%	20.1%	22.9%	29.9%
	Men	22.8%	14.4%	20.3%	22.8%	24.8%	36.7%
	Total	22.5%	14.4%	20.1%	22.5%	24.6%	36.7%
1454–1491	Women	19.7%	12.9%	17.4%	19.5%	21.4%	34.5%
	Men	22.1%	13.0%	19.7%	22.1%	24.4%	33.3%
	Total	21.8%	12.9%	19.3%	21.8%	24.2%	34.5%
1492–1529	Women	20.2%	14.4%	17.5%	19.4%	21.6%	29.3%
	Men	21.8%	14.1%	19.6%	21.6%	23.8%	33.1%
	Total	21.6%	14.1%	19.2%	21.4%	23.6%	33.1%
1530–1567	Women	20.5%	13.3%	18.0%	20.1%	22.1%	26.3%
	Men	21.0%	12.3%	18.9%	20.7%	23.0%	31.3%
	Total	21.0%	12.3%	18.9%	20.7%	22.9%	31.3%
1568–1605	Women	19.1%	11.7%	16.7%	18.8%	20.9%	28.5%
	Men	20.9%	9.9%	18.3%	20.6%	23.0%	37.3%
	Total	20.7%	9.9%	18.1%	20.3%	22.9%	37.3%
1606–1643	Women	19.3%	11.4%	17.2%	19.1%	21.0%	29.6%
	Men	20.9%	10.1%	18.2%	20.6%	23.2%	32.9%
	Total	20.5%	10.1%	17.9%	20.2%	22.7%	32.9%

1644–1681	Women	16.2%	6.9%	16.0%	18.0%	20.2%	38.0%
	Men	20.8%	6.2%	18.0%	20.5%	23.1%	34.7%
	Total	20.4%	6.2%	17.7%	20.1%	22.9%	38.0%
Total	**Women**	19.3%	6.9%	17.0%	19.1%	21.2%	38.0%
	Men	21.2%	6.2%	18.6%	21.0%	23.4%	37.3%
	Grand total	20.9%	6.3%	18.3%	20.7%	23.2%	38.0%

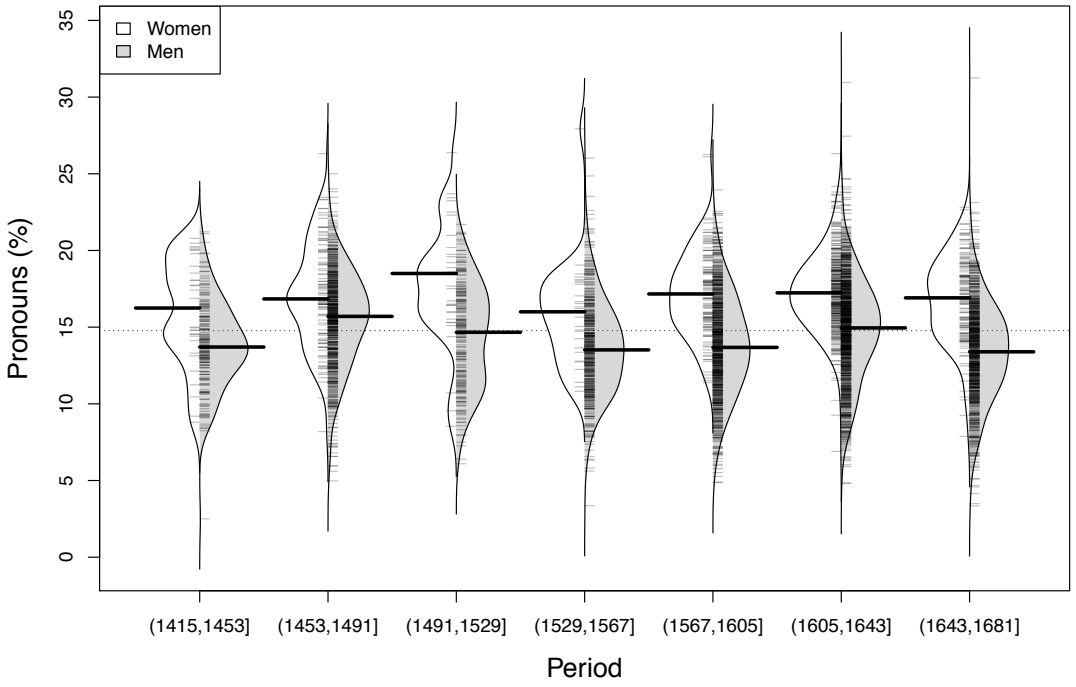

Fig. 6 Sociolinguistic variation in pronoun ratio over time: gender

Table 6 Proportion of pronouns in different time periods split according to gender of letter writer

			Quintiles				
Period	**Gender**	**Mean**	**0%**	**25%**	**50%**	**75%**	**100%**
1415–1453	Women	16.0%	8.8%	14.3%	16.2%	18.7%	20.8%
	Men	13.8%	2.5%	12.1%	13.7%	15.7%	21.2%
	Total	14.2%	2.5%	12.4%	14.0%	16.2%	21.2%
1454–1491	Women	17.1%	8.2%	14.6%	16.8%	19.4%	26.3%
	Men	15.4%	5.0%	13.3%	15.7%	17.7%	25.0%
	Total	15.7%	5.0%	13.5%	15.9%	17.8%	26.3%
1492–1529	Women	17.9%	8.5%	16.0%	18.5%	19.7%	26.4%
	Men	14.4%	6.1%	11.5%	14.7%	17.1%	21.7%
	Total	14.9%	6.1%	11.8%	15.2%	17.5%	26.4%
1530–1567	Women	15.8%	10.8%	13.5%	16.0%	17.9%	27.9%
	Men	13.7%	3.4%	11.3%	13.5%	15.6%	26.0%
	Total	13.8%	3.4%	11.4%	13.7%	15.8%	27.9%

1568–1605	Women	17.1%	11.4%	15.3%	17.2%	18.8%	26.3%
	Men	13.8%	4.9%	11.5%	13.7%	16.2%	24.0%
	Total	14.2%	4.9%	11.9%	14.2%	16.6%	26.3%
1606–1643	Women	17.3%	6.9%	15.7%	17.2%	18.8%	26.3%
	Men	14.8%	4.8%	12.6%	15.0%	17.0%	31.0%
	Total	15.3%	4.8%	13.4%	15.6%	17.6%	31.0%
1644–1681	Women	16.8%	7.9%	14.8%	16.9%	18.9%	22.8%
	Men	13.3%	3.3%	11.2%	13.4%	15.5%	31.2%
	Total	13.8%	3.3%	11.5%	14.0%	16.0%	31.2%
Total	**Women**	17.1%	6.9%	15.2%	17.0%	18.9%	27.9%
	Men	14.2%	2.5%	11.9%	14.3%	16.5%	31.2%
	Grand total	14.7%	2.5%	12.3%	14.8%	17.0%	31.2%

Another aspect worth exploring is the variation in noun and pronoun ratios according to the gender of both the sender and the recipient of each letter. Unfortunately, the data is too unbalanced for statistical analysis (Table 7), and the differences in Figs. 7–10 are suggestive only. The only period with a reasonable number of observations in each category is 1606–1643, where the differences are very slight. In letters sent by men, there are a few other periods with a good amount of data, and we may observe that men seem to use more nouns when writing to men than to women (Fig. 7, except for 1606–1643), and more pronouns when writing to women than to men (Fig. 9). In letters sent by women, there is no clear trend, but looking at the best period in terms of the amount of data, 1606–1643, it seems that women use more of both nouns and pronouns when writing to men than when writing to women (Figs. 8 and 10). It is probable that there are other factors at play here besides gender.[5]

Table 7 Number of letters in different time periods split by gender of letter sender and recipient

Period	**Woman to woman**	**Woman to man**	**Man to woman**	**Man to man**
1415–1453	0	31	0	179
1454–1491	3	99	66	575
1492–1529	1	32	13	190
1530–1567	6	20	34	400
1568–1605	24	95	24	771
1606–1643	132	185	454	618
1644–1681	26	185	110	688

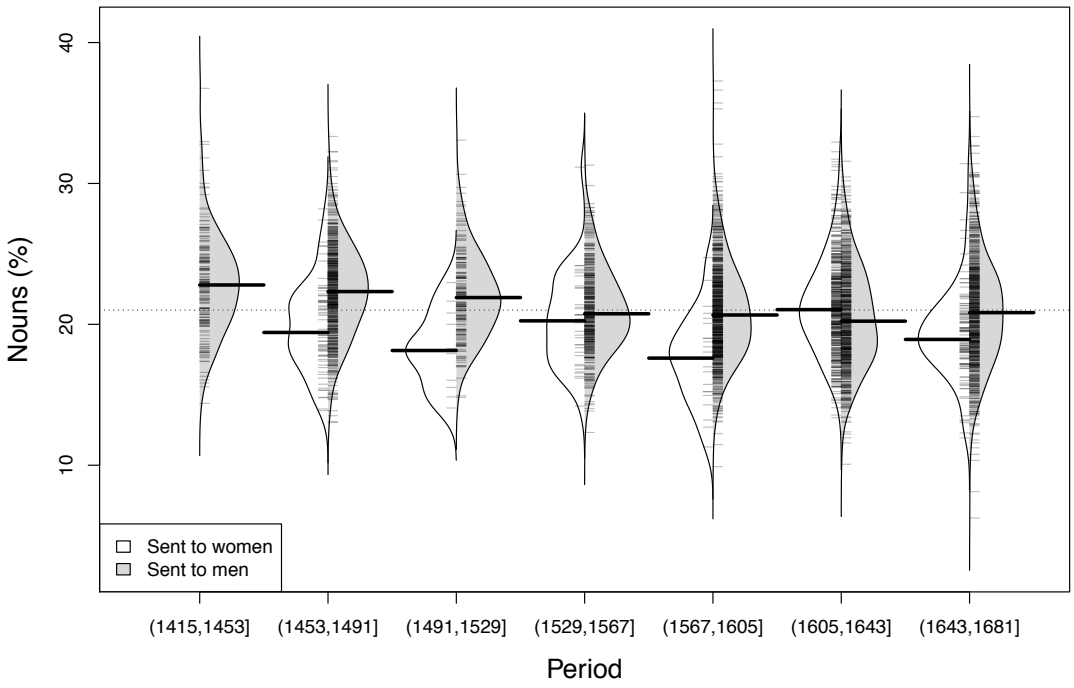

Fig. 7 Variation in noun ratio in letters sent by men according to recipient's gender

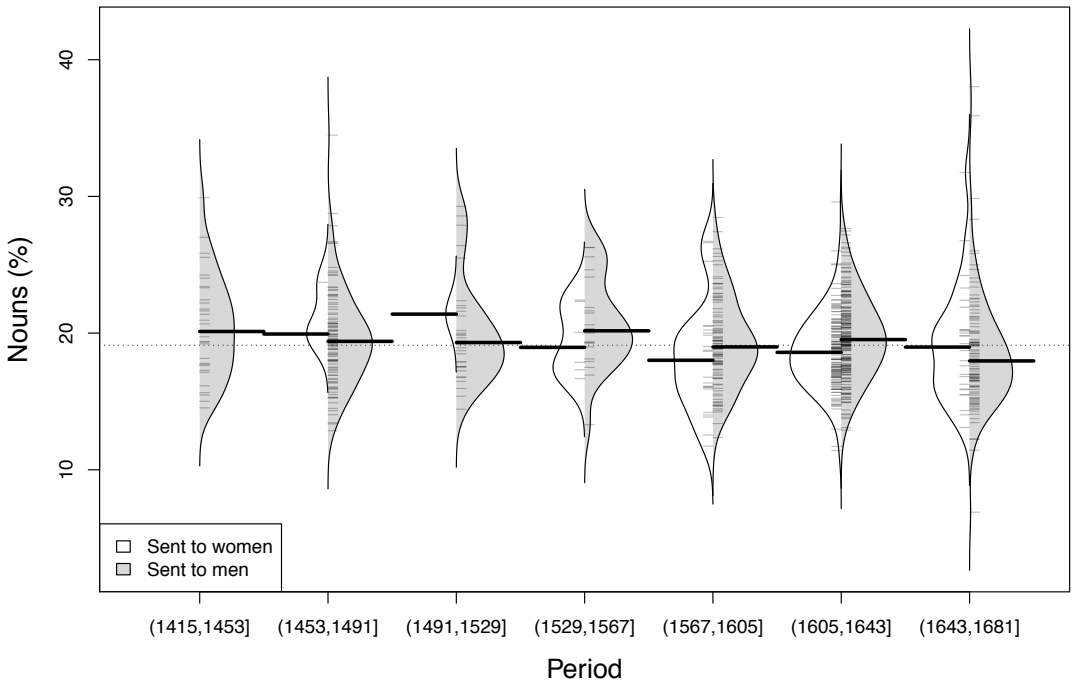

Fig. 8 Variation in noun ratio in letters sent by women according to recipient's gender

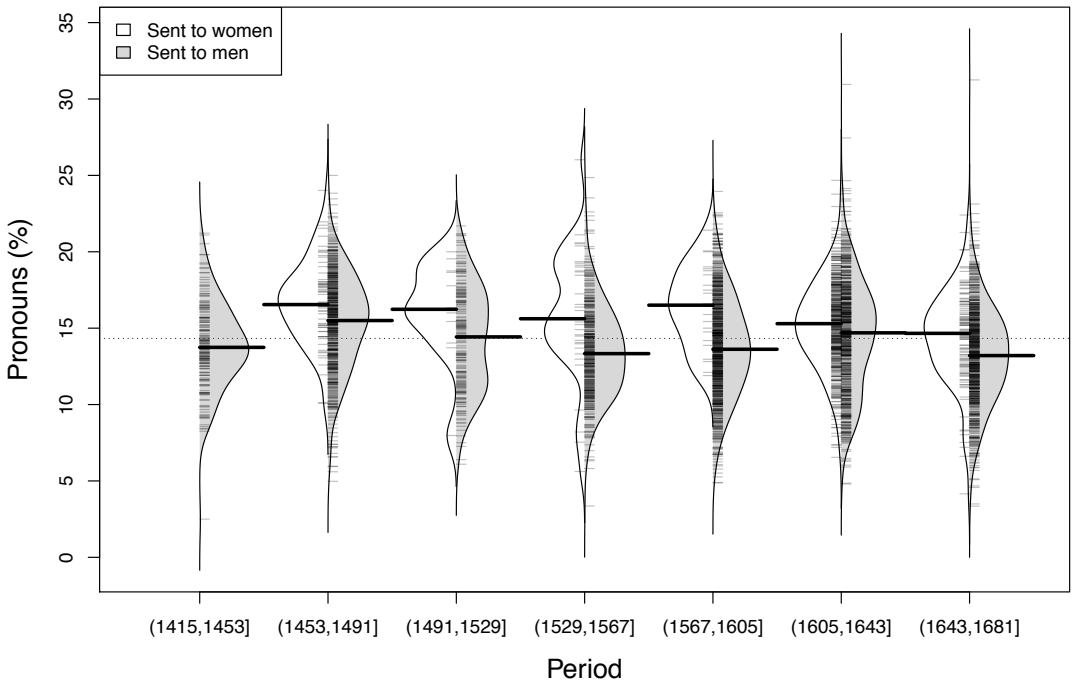

Fig. 9 Variation in pronoun ratio in letters sent by men according to recipient's gender

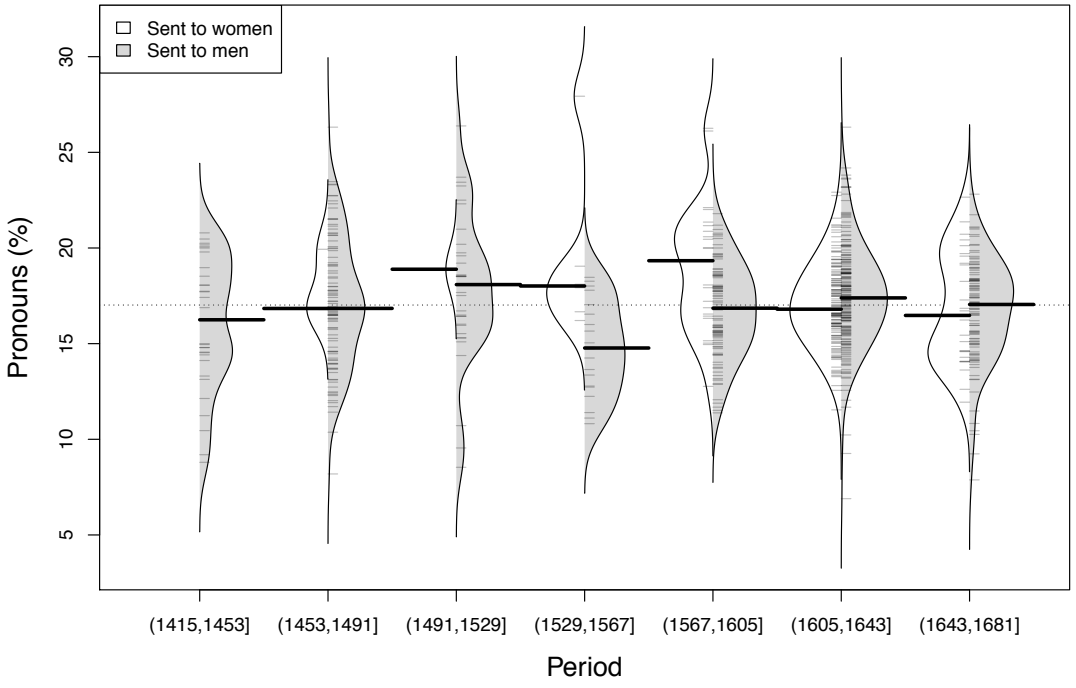

Fig. 10 Variation in pronoun ratio in letters sent by women according to recipient's gender

6. Discussion

We set out to discover whether there was variation in part-of-speech ratios over time in the PCEEC. We have indeed found such variation: the mean frequency of nouns declines from 22.5% to 20.4%, while the mean frequency of pronouns fluctuates between 13.8% and 15.7%, with no clear diachronic trend. Our findings would thus seem to confirm the observations in Biber and Finegan (1997) and Mair *et al*. (2002) that there may be genre evolution in terms of part-of-speech frequencies. Our 38-year periodization is fairly comparable to Mair *et al*.'s 30 years; unlike Mair *et al*., however, our pronoun ratios do not mirror noun ratios, there being no rise in pronouns corresponding to the fall in nouns. Thus, we cannot straightforwardly state, like Biber and Finegan (1997), that the letter-writing genre becomes more 'involved' over time in our corpus.

In future work, it would be of interest to break down the category of pronouns by person and number, which might reveal regularities in the variation. Alternatively, the explanation could depend on the use of verbs. Perhaps the frequency of verbs increases over time, which would make for less 'nouniness' but might not affect pronouns so much. This could also make the sentences shorter and less complex, with noun phrases appearing in fewer functions, which could be explored using the parsed version of the corpus.

Interestingly, we have also discovered gender-based variation in part-of-speech frequencies that persists over time, women consistently using more pronouns than men, while men use more nouns than women. These findings correspond to the PDE results in Rayson *et al*. (1997) and Argamon *et al*. (2003), which would seem to suggest that gendered styles may remain stable throughout the centuries (cf. Labov, 1982, p. 38; Labov, 1990, pp. 206–207; Nevalainen, 2002, pp. 191–194). There are, of course, other factors besides writer gender at play here, including the social status of the writers, the kind of education they had, and the role of the recipients, as well as the length and topics of the letters. We took the first step in this direction by comparing letters written by men to men, men to women, women to men and women to women, but this was not enough to reveal clear underlying patterns, partly due to lack of data from women in particular.

Among other things, future work should consider the relationship between the writer and the recipient, which is encoded in the header of each letter in the PCEEC: FN for nuclear family, FO for other family, FS for family servants, TC for close friends, and T for other acquaintances. Palander-Collin (2009, p. 269) shows that the frequency of the first-person pronoun *I* varies with both gender and family relationship in early English letters: while women use more *I* than men, male members of the gentry do vary their usage in that they employ *I* when writing to family more frequently than when writing to non-family, the frequency increasing over time in both of these contexts. Therefore, it seems possible that relationships could play a role in pronoun use in general. In the PCEEC, the proportions of the various relationships do not remain stable over time, which could in part explain the fluctuation in pronoun ratios. The relatively high ratios in the second and penultimate periods in Fig. 4 seem to be matched by peaks in the proportion of letters written to nuclear family, while the low ratio of the middle period matches a peak in the proportion of letters written to acquaintances. More data is available from later centuries, as well as from other genres such as trial proceedings and drama texts (e.g. Biber and Burges, 2000).

7. Conclusion

Could our findings be an artefact of annotation? The part-of-speech tagging of the PCEEC has its problems, which we tested by reannotating the corpus according to our own principles and comparing the two annotations. While there were quite a few changes, the mean change was only -0.19% in nouns and +0.15% in pronouns, which is smaller than the difference observed by Hardie (2007, p. 67) between the noun ratios in different tokenizations of the Brown Corpus (0.36%). In any case, the results presented here were obtained using the reannotated version of the corpus (ReCEEC), which should reduce the possibility of annotation skewing the results while increasing their comparability with PDE findings. As noted in Section 4.3, users of the PCEEC can easily make the tokenization of the corpus more comparable with PDE by combining the items written separately but tagged as unitary (see Taylor and Santorini, 2006, 'Items treated as unitary'). There are, however, still problems with comparisons between exact percentages across corpora annotated using different schemes.

In summary, how stable is the PCEEC over time? The data is unevenly distributed, yet we were able to observe diachronic change in noun ratios. The mean change from the first period to the last was around two percentage points; whether this matters to the users of the corpus depends on what they are using it for. In addition to the uneven distribution of data with respect to time and sociolinguistic categories, there are outliers in the data: individuals whose language use differs significantly from their peers. For this study, we ended up removing one such individual from the data set. Again, the decision to keep or remove outliers depends on the individual user's purposes. Our aim here has been to provide users with reliable information on which to base their decisions.

Acknowledgements

We would like to thank Jukka Suomela for coding the Python script used to produce the ReCEEC, and our research assistant Mikko Hakala for doing the groundwork for the tokenization study in Section 4.3 (among other things). We are also grateful to Arja Nurmi and the members of the DAMMOC project for comments and discussions, and to anonymous reviewers for helpful feedback.

Notes

1. For more information on the HC and its structure, see the Corpus Resource Database (CoRD) at: http://www.helsinki.fi/varieng/CoRD/corpora/HelsinkiCorpus/
2. A promising solution for the future is offered by the Variant Detector (VARD; Baron and Rayson, 2009), which normalizes historical texts so that they can be analysed and tagged using tools developed for Present-day English material. Version 2 of the program can to some extent cope with a number of issues, such as extra word-final *e*, variation between *u–v*, *i–j*, *ie–y*, *'–e*, and *ee–ea*, fused forms such as *'tis*, morphological variants such as *didst*, hyphenated compounds such as *on-foot*, missing letters as in *hardning*, doubling of consonants as in *comming*, as well as combinations of these (Rayson *et al*., 2008, p. 41). Nevertheless, some problems remain – for instance, VARD 2 does not detect open compounds that are now closed, such as *it self*, or real-word variants such as *then* for *than* (Rayson *et al*., 2008, pp. 33, 39).
3. The date of the first letter in the corpus is unknown, but it is estimated to have been written in the 1410s. Elsewhere, we have given the year as '1410?', but for the purposes of this article, we use 1415.
4. http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/
5. In a historical corpus such as the PCEEC, the various social categories are necessarily unevenly distributed. At the suggestion of an anonymous reviewer, we fit an analysis of variance model to the data using R. As Section 5.1 showed no chronological trend for pronouns (Fig. 4) but a clear trend for nouns (Fig. 3), we used a linear model with the relative frequency of nouns as the dependent variable and the following independent variables: time (continuous, i.e. year of the letter); writer gender; recipient gender; time and writer gender; time and recipient gender; writer gender and recipient gender; time, writer gender, and recipient gender. There was an overall main effect of time ($p < 0.001$) such that the percentage of nouns decreased, a main effect of writer gender ($p < 0.001$) such that men used more nouns, and a main effect of recipient gender ($p < 0.01$) such that letters sent to men had more nouns. Furthermore, there seemed to be an interaction between time and recipient gender ($p < 0.001$), but no significant interaction between time and writer gender.

Funding

This work was supported by Langnet, the Finnish Graduate School in Language Studies [T. S.]; and the Academy of Finland [129282 to T. N. and H. S.].

References

Argamon, S., Koppel, M., Fine, J., and Shimoni, A. R. (2003). Gender, genre, and writing style in formal written texts. *Text*, **23**(3): 321–346.

Baron, A. and Rayson, P. (2009). Automatic standardisation of texts containing spelling variation: How much training data do you need? In Mahlberg, M., González-Díaz, V., and Smith, C. (eds), *Proceedings of the Corpus Linguistics Conference: CL2009, University of Liverpool, UK, 20–23 July 2009*. http://ucrel.lancs.ac.uk/publications/cl2009/ (article #314, accessed 1 July 2010).

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. and Burges, J. (2000). Historical change in the language use of women and men: Gender differences in dramatic dialogue. *Journal of English Linguistics*, **28**(1): 21–37.

Biber, D. and Finegan, E. (1989). Drift and the evolution of English style: A history of three genres. *Language*, **65**(3): 487–517.

Biber, D. and Finegan, E. (1997). Diachronic relations among speech-based and written registers in English. In Nevalainen, T. and Kahlas-Tarkka, L. (eds), *To Explain the Present: Studies in the Changing English Language in Honour of Matti Rissanen*. Mémoires de la Société Néophilologique de Helsinki, 52. Helsinki: Société Néophilologique, pp. 253–275.

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education.

CEEC = *Corpus of Early English Correspondence*. (1998). Compiled by T. Nevalainen, H. Raumolin-Brunberg, J. Keränen, M. Nevala, A. Nurmi, and M. Palander-Collin at the Department of English, University of Helsinki.

Durrell, M., Ensslin, A., and Bennett, P. (2007). GerManC: A historical corpus of German 1650–1800. *Sprache und Datenverarbeitung*, **31**: 71–80.

Gries, S. Th. and Hilpert, M. (2008). The identification of stages in diachronic data: Variability-based neighbour clustering. *Corpora*, **3**(1): 59–81.

Hardie, A. (2007). Part-of-speech ratios in English corpora. *International Journal of Corpus Linguistics*, **12**(1): 55–81.

Hartigan, J. A. and Kleiner, B. (2008). A mosaic of television ratings. *The American Statistician*, **38**: 32–35.

Hothorn, T., Bretz, F., and Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, **50**(3): 346–363.

Hudson, R. (1994). About 37% of word-tokens are nouns. *Language*, **70**(2): 331–339.

Kampstra, P. (2008). Beanplot: A boxplot alternative for visual comparison of distributions. *Journal of Statistical Software*, **28**: Code Snippet 1.

Kytö, M. and Voutilainen, A. (1998). Backdating the English Constraint Grammar Parser for the analysis of English historical texts. In Hogg, R. M. and van Bergen, L. (eds), *Historical Linguistics 1995, Volume 2: Germanic Linguistics. Selected Papers from the 12th International Conference on Historical Linguistics, Manchester, August 1995*. Current Issues in Linguistic Theory, 162. Amsterdam: John Benjamins, pp. 149–166.

Labov, W. (1982). Building on empirical foundations. In Lehmann, W. P. and Malkiel, Y. (eds), *Perspectives on Historical Linguistics*. Current Issues in Linguistic Theory, 24. Amsterdam: John Benjamins, pp. 17–92.

Labov, W. (1990). The intersection of sex and social class in the course of linguistic change. *Language Variation and Change*, **2**: 205–254.

Leech, G. and Smith, N. (2005). Extending the possibilities of corpus-based research on English in the twentieth century: A prequel to LOB and FLOB. *ICAME Journal*, **29**: 83–98.

Mair, C., Hundt, M., Leech, G., and Smith, N. (2002). Short term diachronic shifts in part-of-speech frequencies: A comparison of the tagged LOB and F-LOB corpora. *International Journal of Corpus Linguistics*, **7**(2): 245–264.

Nesselhauf, N. (2007). The spread of the progressive and its 'future' use. *English Language and Linguistics*, **11**(1): 191–207.

Nevalainen, T. (1996). Social stratification. In Nevalainen, T. and Raumolin-Brunberg, H. (eds), *Sociolinguistics and Language History: Studies Based on the Corpus of Early English Correspondence*. Amsterdam: Rodopi, pp. 57–76.

Nevalainen, T. (2002). Language and woman's place in earlier English. *Journal of English Linguistics*, **30**(2): 181–199.

Nevalainen, T. and Raumolin-Brunberg, H. (2003). *Historical Sociolinguistics: Language Change in Tudor and Stuart England*. Longman Linguistics Library. London: Pearson Education.

Nurmi, A., Nevala, M., and Palander-Collin, M. (eds). (2009). *The Language of Daily Life in England (1400–1800)*. Pragmatics and Beyond NS, 183. Amsterdam: John Benjamins.

OED = *The Oxford English Dictionary*, 2nd edition. (1989). OED Online. Oxford: Oxford University Press. http://dictionary.oed.com (accessed 1 July 2010).

Palander-Collin, M. (2009). Variation and change in patterns of self-reference in early English correspondence. *Journal of Historical Pragmatics*, **10**(2): 260–285.

PCEEC = *Parsed Corpus of Early English Correspondence*, tagged version. (2006). Annotated by A. Nurmi, A. Taylor, A. Warner, S. Pintzuk, and T. Nevalainen. Compiled by the CEEC Project Team. York: University of York and Helsinki: University of Helsinki. Distributed through the Oxford Text Archive.

Plag, I. (2003). *Word-Formation in English*. Cambridge Textbooks in Linguistics. Cambridge: Cambridge University Press.

R Development Core Team. (2010). R: A Language and Environment for Statistical Computing. http://www.R-project.org (accessed 1 July 2010).

Raumolin-Brunberg, H. and Nevalainen, T. (2007). Historical sociolinguistics: The Corpus of Early English Correspondence. In Beal, J. C., Corrigan, K. P., and Moisl, H. L. (eds), *Creating and Digitizing Language Corpora* (Vol. 2). Houndsmills: Palgrave Macmillan, pp. 148–171.

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 Davies, M., Rayson, P., Hunston, S., and Danielsson, P. (eds), *Proceedings of Corpus Linguistics 2007, 27–30 July, University of Birmingham, UK*. http://www.corpus.bham.ac.uk/conference/proceedings.shtml (article #192, accessed 1 July 2010).

Rayson, P., Archer, D., Baron, A., and Smith, N. (2008). Travelling through time with corpus annotation software. In Lewandowska-Tomaszcyk, B. (ed.), *Corpus Linguistics, Computer Tools, and Applications – State of the Art: PALC 2007*. Łódź Studies in Language, 17. Frankfurt am Main: Peter Lang, pp. 29–46.

Rayson, P., Leech, G., and Hodges, M. (1997). Social differentiation in the use of English vocabulary: Some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics*, **2**(1): 133–152.

Santorini, B. (1990). *Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision)*. University of Pennsylvania Department of Computer and Information Science Technical Report No. MS-CIS-90-47. http://repository.upenn.edu/cis_reports/570/ (accessed 1 July 2010).

Santorini, B. (2010). *Annotation Manual for the Penn Historical Corpora and the PCEEC*. http://www.ling.upenn.edu/hist-corpora/annotation/ (accessed 1 July 2010).

Spence, R. (2007). *Information Visualization: Design for Interaction*. Harlow: Pearson Education / Prentice Hall.

Taylor, A. (2003). *YCOE Lite: A Beginner's Guide to the York Corpus of Old English*. http://www-users.york.ac.uk/~lang22/YCOE/doc/annotation/YcoeLiteToc.htm (accessed 1 July 2010).

Taylor, A. (2007). The York–Toronto–Helsinki Parsed Corpus of Old English Prose. In Beal, J. C., Corrigan, K. P., and Moisl, H. L. (eds), *Creating and Digitizing Language Corpora* (Vol. 2). Houndsmills: Palgrave Macmillan, pp. 196–227.

Taylor, A. and Santorini, B. (2006). *The Parsed Corpus of Early English Correspondence*. http://www-users.york.ac.uk/~lang22/PCEEC-manual/ (accessed 1 July 2010).

Theus, M. and Urbanek, S. (2008). *Interactive Graphics for Data Analysis: Principles and Examples*. Computer Science and Data Analysis. Boca Raton, FL: CRC / Chapman & Hall.

Appendix 1 List of N tags in the PCEEC

The following table lists all N tags in the PCEEC. These include the noun tags given in the PCEEC manual (N, N$, NPR, NPR$, NPRS, NPRS$, NS, NS$), all compound tags ending in any of those tags, and four other compound tags that were included in the class of nouns in our analysis (N+ADJ, NS+ADJ, NS+NUM, NS+RP). Tags ending in numbers mark unitary items (e.g. *non*_N21 *payment*_N22); only the first component is listed here (e.g. N21). 'Included?' refers to whether or not the tag was included in the class of nouns in our analysis. Retokenization is not shown (e.g. D+N$ was split into D and N$, the first component classified as an article and the second as a noun).

Tag	Frequency	Included?	Examples
ONE+N	1	No	*oneselfe*
OTHER+N	675	No	*otherwise*
OTHER+NS	44	No	*otherways*
P+N	2,148	No	*because*, *indeed*
P+N21	1	No	*a forehand*
P+NS	1	No	*apeeces*
P+P+N	3	No	*forbecause*
PRO$+N	1,211	No	*myself*
PRO$+N$	1	No	*yourselfe* (*'for yourselfe sake'*)
PRO+N	1,609	No	*himself*
Q+N	2,065	No	*anyway*, *nothing*
Q+N$	2	No	*nobodys*
Q+NS	187	No	*noways*, *sometimes*
Q+OTHER+N	1	No	*nodyrwyse*
QS+N	2	No	*leastwise*
ADJ+N	1,199	Part	*gentleman*, *likewise*
ADJS+N	2	Part	*nextyme*, *leastwise*
ADV+N	11	Part	*well-wisher*, *before-hand*
ADV+NS	24	Part	*wellwishers*, *oftentimes*
D+N	1,175	Part	*th'end*, *awhile*
N	260,926	Part	*man*, (*her*) *self*, *tomorrow*
N$	6,264	Part	*womans*, (*no*) *bodys*
N21	512	Part	*non payment*, *to morrow*
NS	58,503	Part	*abilities*, (*some*) *times*
ADJ+N$	23	Yes	*gentlemans*
ADJ+N+N	1	Yes	*gentlemanusher*
ADJ+N+NS	1	Yes	*lyghthorsemen*
ADJ+NS	396	Yes	*gentlemen*
ADJ+NS$	7	Yes	*gentlemens*
ADJ+P+N	1	Yes	*nere-a-kin*
D+N$	108	Yes	*themperours*
D+N+N	1	Yes	*th'Esteborder*
D+NPR	9	Yes	*thenglish*
D+NPR$	1	Yes	*th'Empnals*
D+NPRS	6	Yes	*thevangelikes*
D+NS	95	Yes	*thenstructions*

D+NS$	1	Yes	*th'offendors*
D+OTHER+N	2	Yes	*toþersyde*
N$+N	11	Yes	*hogshead*
N$+NS	4	Yes	*townsmen*
N$21	2	Yes	*Em prethorys*
N+ADJ	24	Yes	*governor-general*
N+CONJ+N	1	Yes	*time-and-place*
N+N	1,104	Yes	*horseback*
N+N$	17	Yes	*iremongers*
N+N+N	6	Yes	*checquer-chamber-case*
N+N+NS	1	Yes	*alehouskeepers*
N+NS	463	Yes	*countrymen*
N+NS$	6	Yes	*Marchant-taylers*
N+P+N	36	Yes	*father-in-law*
N31	1	Yes	*mare a ge* (i.e. *marriage*)
NPR	105,042	Yes	*England*
NPR$	3,272	Yes	*Cheretons*
NPR$+N	13	Yes	*newyeares-guift*
NPR$+NPR	1	Yes	*Grays-Inne*
NPR$21	1	Yes	*Saue wayes* (i.e. *Savoy's*?)
NPR+N	71	Yes	*godmother*
NPR+N$	4	Yes	*godfathers*
NPR+NPR	1	Yes	*M^r^ Edmondes*
NPR+NS	6	Yes	*Christmas-games*
NPR21	14	Yes	*portes mouthe*
NPRS	1,350	Yes	*Jesuits*
NPRS$	55	Yes	*Scots*
NS$	618	Yes	*mens*
NS$+N	2	Yes	*Haberdashers-hall*
NS$21	2	Yes	*ill wishers*
NS+ADJ	2	Yes	*states-generall*
NS+N	1	Yes	*almesfolk*
NS+NS	5	Yes	*merchantes-adventurers*
NS+NUM	95	Yes	*£14*
NS+RP	1	Yes	*standers-by*
NS21	26	Yes	*well wishers*
NUM+N	114	Yes	*sixpence*
NUM+NS	705	Yes	*2^s^*
NUM+NS$	2	Yes	*12-ms*
NUM+NS+NUM+NS	10	Yes	*6^s^-8^d^*
OTHER+NS$	1	Yes	*othermens*
P+ADJ+N	2	Yes	*inlikewise*
P+D+N	1	Yes	*inthende*
P+NPR	3	Yes	*withbrowne*
VAN+N	1	Yes	*saydbyll*

Appendix 2 Material omitted from nouns (most frequent changes)

Pronouns (5,826)

In addition to PRO+N (e.g. *himself*) and PRO$+N (e.g. *myself*), there were reflexive pronouns written as two words, the first of which was tagged PRO or PRO$ and the second N (e.g. *him*_PRO *self*_N). These instances of the word *self* in all its variant spellings were identified and combined with the preceding word, which was then classified as a pronoun.

Adverbs (2,369)

One of the largest categories of adverbs tagged as nouns consisted of words ending in *-wise* (e.g. *likewise*_ADJ+N, *otherwise*_OTHER+N). These were identified and reclassified as adverbs. Sometimes they were written as two words (e.g. *other*_OTHER *wise*_N); these instances of *wise*_N were combined with the previous word, which was then classified as an adverb. In some cases, *wise* appeared in phrases such as *in like wise*; however, as the majority of instances of *like wise* were adverbial uses without the preposition, they were all combined and classified as adverbs. There was clearly a cline of grammaticalization here, but the scales had already tipped in favour of the adverb.

The other major adverbial category pruned out of nouns was time adverbs ending in *day*, *night*, *morrow*, *even* (e.g. *today*_N, *yesterday*_N), written both together and as two words. A minority of these was used as nouns, but we followed the BNC in classifying all instances as adverbs.

'Quantifiers' (4,084)

There seemed to be an entire paradigm of what the PCEEC manual calls 'quantifiers', such as *all*, *any*, *every*, *many*, *no*, *some*, combined with the nouns *way*(*s*) and *time*(*s*) (e.g. *anyways*_Q+NS, *sometimes*_Q+NS, *anyway*_Q+N, *sometime*_Q+N, but in many more combinations than in present-day standard English). Therefore, we treated them all in the same way. They were also sometimes written as two words (e.g. *some*_Q *times*_NS), which we combined. Some appeared in prepositional phrases (e.g. *in any way* or *in anyway*), but we went by the majority, which usually meant pruning them out of nouns; if a certain item appeared with prepositions in the majority of cases, it was left in. As the class of quantifiers is a disputed one, we put the pruned items in a class called 'unused'. An argument could be made for including these in adverbs, but the main point is that they are not nouns, due to a process of grammaticalization.

The above also applies to combinations of quantifiers with the nouns *thing* and *body* (e.g. *anything*_Q+N, *any*_Q *thing*_N, *anybody*_Q+N, *any*_Q *body*_N), with the exception that these are more like indefinite pronouns.

Subordinating Conjunctions (1,578)

Finally, the subordinating conjunctions *because*_P+N and *instead*_P+N were removed from nouns and placed in the class of prepositions and subordinating conjunctions. The rest of the words tagged P+N consisted mainly of adverbs (e.g. *asleep*, *indeed*) and were classified as such.