Exploring part-of-speech frequencies in a sociohistorical corpus of English

This is the 'author accepted manuscript' of the following paper: Säily, Tanja, Turo Vartiainen & Harri Siirtola. 2017. Exploring part-of-speech frequencies in a sociohistorical corpus of English. In Tanja Säily, Arja Nurmi, Minna Palander-Collin & Anita Auer (eds.), *Exploring future paths for historical sociolinguistics* (Advances in Historical Sociolinguistics 7), 23–52. Amsterdam: John Benjamins. doi:10.1075/ahs.7.02sai

Tanja Säily[1], Turo Vartiainen[1] & Harri Siirtola[2]

[1]University of Helsinki, [2]University of Tampere

Abstract

We investigate the usefulness of part-of-speech (POS) annotation as a tool in the study of sociolinguistic variation and genre evolution. We analyse how POS ratios change over time in the *Parsed Corpus of Early English Correspondence* (c.1410–1681), which social groups lead the changes, and whether the changes can be connected to colloquialisation with regard to reduced complexity or an increasingly involved style. While we find gentry-led colloquialisation in terms of noun and verb frequencies as well as evidence for gendered styles, the results on structural complexity are more mixed. We argue that POS annotation can be a useful tool when complemented by a thorough textual analysis, but that more fine-grained categories are needed to reach firmer conclusions.

Keywords

part-of-speech frequencies, historical sociolinguistics, language change, genre evolution, colloquialisation, structural complexity, gendered styles, involved text production, exploratory data analysis, correspondence

1. Introduction

The composition of a corpus in terms of part-of-speech ratios crucially depends on the way in which the corpus is compiled. A particularly important factor affecting POS distribution is genre balance: different genres are associated with different communicative purposes (e.g. Biber & Conrad 2009: 6), and these purposes are linguistically communicated in different ways and expressed through different constructions and parts of speech (Biber et al. 2016: 643–644). For instance, if a corpus consists of texts from a highly interactive and involved genre, such as spoken conversation, the proportion of first- and second-person pronouns will be relatively high compared to the proportion of nouns. By contrast, if the corpus includes texts from a more information-oriented genre, such as academic prose, we expect to see a higher proportion of nouns and a lower proportion of pronouns.

Many corpus-based genre studies subscribe to the multi-dimensional model of genres that was first systematically discussed in Biber (1988).

Biber's key idea was that genre variation in English could be described by studying co-occurring linguistic features in texts, which are then automatically classified into dimensions reflecting their typical communicative functions by factor analysis. In short, in Biber's model an observation based on a single POS label (e.g. that a text has a high proportion of nouns) is interesting, and it may be suggestive of information orientation, for instance, but it is the co-occurrence of nouns with other linguistic markers, such as attributive adjectives and prepositional phrases, that leads to the conclusion that the text in question is in fact informationally oriented (Biber & Gray 2010). In addition to "Informational vs. Involved Production", the dimensions in Biber's model include, for example, "Narrative vs. Non-narrative Concerns" and "Explicit vs. Situation-dependent Reference". All these dimensions are described in terms of linguistic features that are weighted according to their relative importance within the dimension. However, as is also apparent in Biber's model, some features are clearly more important than others, which raises the methodological question of whether genres could also be studied in a way that would be technically less demanding than multi-dimensional factor analysis, yet informed by the insights of Biber's computational approach. More specifically, as modern linguistic corpora are increasingly often tagged for parts of speech, it would be especially interesting to see if it were possible to study variation and change in different genres simply by

observing changes in the proportion of POS labels without access to any kind of semantic or parsing information.

Our paper is intended to contribute to the discussion of the usefulness of POS ratios in the study of language variation and change and genre evolution. In what follows, we will discuss the genre of *personal correspondence* by making use of the data from the *Parsed Corpus of Early English Correspondence*. Because this corpus covers a long period of time (c.1410–1681) and it is annotated for many sociolinguistically relevant features, such as the gender of both the letter's author and the recipient, and their mutual relationship, we are able to study both the evolution of the genre and sociolinguistic variation in terms of POS ratios. Our main research questions, which will be discussed in more detail in Section 4, are listed below.

1. How does the distribution of POS tags change over time in the corpus?
2. Is there any evidence of the colloquialisation of the genre that can be measured by changes in the POS ratios?
3. Can we find any sociolinguistic variation and/or sociolinguistically conditioned change in the distribution of POS tags in the corpus?

The first question is a very general one, and we will discuss it from the perspective of how a simple analysis of changing POS proportions can inform linguistic research questions. The second question concerns

colloquialisation, that is, the gradual shift towards a more "oral" or colloquial style that has been observed in many written genres of English (see Section 2 below for a more detailed discussion). Furthermore, this question bears specifically on the results acquired by Biber & Finegan (1989), who found that in their data from *A Representative Corpus of Historical English Registers* (ARCHER) personal letters from the 17th century represented a more involved style of writing than letters from the 18th and 19th centuries, suggesting that personal letters had actually become *less colloquial*, as evidenced by the scarcity of such involved features as contractions, pronouns and hedges (Biber & Finegan 1989: 501). By extending the analysis to 15th and 16th century data, our study will shed more light on the early history of the genre. The final research question will be examined from the perspective of involved text production and sociolinguistically conditioned change: when we divide the data according to gender or social rank, or focus our attention on the roles within the nuclear family (e.g. letters written by husbands and wives), can we see differences in the proportions of POS tags? If we do, is there evidence of sociolinguistically conditioned change and colloquialisation?

The rest of this paper is organised as follows. Section 2 starts by introducing relevant POS-based research on sociolinguistic variation and diachronic change in English from the perspective of colloquialisation, gendered styles and structural complexity. Section 3 continues by discussing the data and the methodology used in the case studies. Section 4 focuses on

analysing the data from the perspective of the research questions outlined above, and Section 5 concludes the paper with a discussion of the main findings and suggestions for further research.

2. Background

2.1 POS ratios in the study of (sociolinguistic) variation

Variation and change in part-of-speech frequencies have previously been studied, for example, in Hudson (1994), Hardie (2007) and Mair et al. (2002). Hudson (1994) compared the LOB corpus of British English with the Brown corpus of American English from the same period (1961) and found that both corpora had a noun frequency of c. 37%, which he hypothesised to be a universal property of English. Hardie (2007) questioned this result by pointing out that Hudson's noun category, which not only included nouns but also pronouns and even other word classes, is so general that it is both controversial from a theoretical perspective and difficult to reproduce by comparing POS ratios in corpora that use different annotation schemes. The latter point is applicable to comparing POS ratios in general, as we shall see. Mair et al. (2002), on the other hand, compared the POS distribution in the LOB and the F-LOB corpora, the latter

representing British English usage in 1991. One of the hypotheses tested by the authors was whether their data supported earlier results obtained in multi-dimensional analyses according to which many written genres of English have gradually become more similar to spoken genres, that is, the genres have become *colloquialised* over time (Biber & Finegan 1989, 1997). However, contrary to expectations, Mair et al. found no evidence of a colloquialisation trend in their data; on the contrary, they found, for instance, that the proportion of nouns was actually higher in the more recent data (F-LOB). The finding was all the more puzzling because there was no corresponding decrease in verb frequencies (Mair et al. 2002: 257).[1]

POS ratios in Present-day English corpora have also been studied from the perspective of gendered styles (e.g. Rayson et al. 1997, Argamon et al. 2003, Heylighen & Dewaele 2002, Newman et al. 2008, Bamman et al. 2014). Rayson et al. (1997) studied the demographically sampled spoken section of the *British National Corpus* and found that men tended to favour common nouns, while women favoured proper nouns, personal pronouns and verbs. Argamon et al.'s (2003) study of male and female writing in the BNC revealed that male writers favoured determiners and numerals, while female writers were characterised by their frequent use of personal

[1] We might point out here that there was a proportionate decrease in pronoun frequencies (Mair et al. 2002: 249), which suggests a trade-off between pronominal and lexical reference and is consistent with increased information orientation.

pronouns. In both studies, the results were argued to be indicative of involved vs. informational styles of writing, so that women's writing was generally more involved than men's. Heylighen & Dewaele (2002), on the other hand, found that women's language use tended to be more context-dependent than men's in terms of the frequency of pronouns, adverbs, inflected verbs and interjections. Newman et al. (2008) studied gender differences in 14,000 text samples through a multivariate analysis of a large number of features, including some that directly corresponded to POS labels. In their data, women tended to use more negations, pronouns and verbs in the present and past tenses, whereas men used more nouns and articles. Finally, Bamman et al. (2014) compared male and female language use on Twitter, finding that female gender markers included e.g. pronouns and male markers e.g. numerals.

Frequencies of parts of speech in the genre of personal correspondence have also been studied from a sociolinguistic perspective. For example, using the *Innsbruck Letter Corpus* (1386–1688), Markus (2001) found that women used more coordinators as well as certain kinds of subordinators, whereas men used more relative pronouns. Although Markus (2001: 196) emphasises the importance of further analyses, he suggests that the results might be explained by male literacy as opposed to female orality. Furthermore, similarly to the earlier studies on Present-day English, Säily et al. (2011: 179) discovered that women consistently used more pronouns than men in their letters, while men used more nouns than women (the

Parsed Corpus of Early English Correspondence (PCEEC), c.1410–1681). This result is consistent with the argument that women's style of writing is generally more involved than men's (cf. Biber & Burges 2000; Palander-Collin 1999, 2000). Säily et al. (2011: 177) also found that the proportion of nouns decreased slightly over time in the corpus, suggesting a small degree of colloquialisation. Vartiainen et al. (2013), on the other hand, refined the analysis of gender differences in pronoun frequencies by considering the influence of social roles within the nuclear family in the *Corpora of Early English Correspondence*. They discovered, for example, that for males the frequency of pronoun use varied depending on whether the men were writing as fathers, sons or husbands. They also found that the gender differences decreased in the 18th century. Finally, as mentioned above, Biber & Finegan (1989) found that in the ARCHER corpus personal letters showed fewer features associated with spoken interaction and high speaker involvement in the 18th and 19th century than in the 17th century data.

To summarise, variation and change in POS ratios have been studied both with Present-day English data and with historical data from various corpora. From a sociolinguistic perspective, these studies have revealed interesting gender differences, for example, in the use of pronouns (favoured by women) and nouns (favoured by men) that are remarkably consistent between genres and also over time. Other features typically associated with female usage include verbs, negations and interjections, whereas features like determiners and numerals are particularly frequent in

men's usage. These corpus-linguistic findings are also consistent with earlier sociolinguistic research on gendered discourse styles, where men's speech has been argued to be more information-oriented as opposed to women's more interactive style (Tannen 1991: 76–77). While these differences cut across genres, Newman et al. (2008) found that they were the most pronounced in informal conversation. Audience design (Bell 1984) also plays a role: Bamman et al. (2014) discovered that the use of gender-specific markers on Twitter intensified within same-sex networks. The majority of previous research, however, has focused on a limited number of parts of speech, often complemented by other features; moreover, social categories other than gender have mostly been ignored. Our goal therefore is to utilise the entire range of POS categories in our corpus and see how far they can take us along with metadata on gender, social rank and the relationship between the sender and recipient of the letter.

2.2 Complexity in the genre of personal correspondence

Our approach to complexity mainly corresponds with Rescher's (1998) definition of *structural complexity*. Structural complexity, and *hierarchical complexity* in particular, refers to the degree of embedding and modification on various structural levels (phrases, clauses, sentences), and we will measure the structural complexity of the texts in our data by examining the proportions of word classes that contribute to the complexity of

modification and complementation patterns in the corpus (e.g. prepositions, conjunctions, *wh*-words). While doing so, we acknowledge the results of earlier research which has shown that complexity is manifested in different ways in speech and writing. For instance, Biber & Gray (2011) found that speech is more complex than writing if measured in terms of the frequency of clausal embedding and subordination. By contrast, the complexity of noun phrases is typically higher in written genres than in spoken ones both in terms of the frequency of premodifying nouns and adjectives as well as postmodifying phrases and clauses (see below). In this study, we are mainly interested in complexity from the perspective of colloquialisation, but we will also study the data from a sociolinguistic perspective with the aim of finding out whether any potential changes in the data are led by a certain social group (e.g. women, the upper ranks). Furthermore, although we make no claims about the relationship between the degree of structural complexity and the relative ease of comprehension, for example, some of the research that will be presented below, and which has also informed our research, specifically argues to this effect (see also Karlsson 2008 for a comprehensive overview of complexity and how the term has been understood in linguistic research).

As a genre, personal correspondence has been found to resemble spoken interaction more than other written genres, such as academic prose or press reportage (see e.g. Biber & Finegan 1989, Biber 1992). Similarly to face-to-face conversation, private letters often focus on interpersonal

concerns, and this correlates with a high frequency of linguistic features that are typical of conversation and show high speaker involvement such as first- and second-person pronouns, private verbs (e.g. *think*, *know* and *suppose*) (Quirk et al. 1985: 1180–1182), and various kinds of stance markers (Biber 1995: 275–276). However, personal letters are also different from spoken discourse in terms of their production circumstances: contrary to spoken interaction, which is constrained by the demands of online text production, letters can be produced carefully and revised according to need (see e.g. Biber 1992: 139). Consequently, while personal letters are in many ways less complex in their structure, and less information-oriented, than other written genres, they nevertheless exhibit some of the complexity that is typically associated with written language (Biber 1992: 151, 159).

The high frequency of pronouns in personal correspondence in part explains why many of the features related to structural elaboration of reference are very rare in private letters. These features increase the structural complexity of noun phrases, and they include, for instance, attributive adjectives, postmodifying prepositional phrases, restrictive and non-restrictive relative clauses and complement *that*-clauses; that is, structures that are extremely rarely used with pronouns, and which are particularly typical of written genres (Biber & Gray 2011). Importantly for our purposes, most of these categories can be studied from the perspective of the distribution of POS labels, which provides us with a good opportunity to explore changes in structural complexity in our data. In other words, by

investigating the changing frequencies of nouns, adjectives, prepositions, *wh*-words and complementisers in the corpus, we hope to find evidence of increased or decreased complexity that could possibly be linked to colloquialisation, and perhaps also to the usage of certain social groups.

Word classes like prepositions, complementisers, (attributive) adjectives and relative pronouns have generally been considered to contribute to increased complexity in previous literature, but we will also include an additional, and somewhat more controversial, category in our discussion of complexity and colloquialisation: coordinating conjunctions. On the one hand, coordinators have been regarded as markers of reduced complexity in previous literature (Chafe 1982; Biber 1992: 140) because they represent a structurally simpler alternative to more complex forms of expression, such as nominalisations, participles and subordinate clauses. On the other hand, in her study of the complexity of statutes in the history of English, Lehto (2015: 16, 139) argued that coordinators may actually *increase* the overall complexity of texts by making the sentences longer and thus imposing a higher cognitive load on working memory. Adopting a more pragmatic view of complexity that was particularly designed for the study of texts from the Early Modern period, Lehto (2015: 140) also argued that punctuation should be considered a complexity feature in historical genres: texts with scarce or no punctuation at all are more difficult to understand (and thus more complex) than texts where clause and sentence boundaries are marked with punctuation. Bearing in mind that our data may

have been affected by editorial practices, we will also discuss the use of punctuation in the letters from the perspective of complexity in Section 4.1 below.

3. Material and method

3.1 PCEEC and ReCEEC

The *Parsed Corpus of Early English Correspondence* (PCEEC) is the published version of the *Corpus of Early English Correspondence* (CEEC), which was compiled in the 1990s by the Sociolinguistics and Language History project team at the University of Helsinki for the purposes of historical sociolinguistics. Based on published editions of letters, the CEEC consists of 2.6 million words of personal correspondence from c.1410–1681, along with metadata on the letters, writers and recipients. The metadata include social categories such as gender, social rank, social mobility, place of birth, domicile, migration history and the relationship between the sender and recipient of the letter, making the corpus an excellent resource for historical sociolinguistic research.

In this paper, we will study our research questions from the perspective of gender, social rank and the relationship between the sender and recipient of the letter. The *gender* category is binary, male vs. female, as

this was and remains the basic social division of gender. *Social rank* can be divided into royalty, nobility, upper gentry, lower gentry, upper clergy, lower clergy, professionals, merchants and other non-gentry (Nevalainen & Raumolin-Brunberg 2003: 136). As the amount of data does not permit us to use such a fine-grained division, we use a dichotomous model of gentry (royalty, nobility, upper gentry, lower gentry, upper clergy) and non-gentry (lower clergy, professionals, merchants, other non-gentry). This model, too, is theoretically motivated as it can be argued that the most basic division in the society of the time was between gentry and non-gentry (Laslett 1965: 26). As royalty is such a special case in terms of language use in both official and family letters, we have chosen to exclude them from our analysis (cf. Vartiainen et al. 2013). The categories of the *relationship between the sender and recipient* of the letter include nuclear family, other family, family servants, close friends and other acquaintances. In our study of gendered styles (Section 4.2), we focus on the nuclear family as this is the only category where we have enough data from women. To analyse specific social roles within the nuclear family, we further zoom in on spousal correspondence.

The PCEEC (2006) comprises those collections of the original CEEC for which permission to re-publish could be obtained (c. 2.2 million words). The corpus comes in three versions: plain text, POS tagged and syntactically parsed. The annotated versions were produced in collaboration between the universities of York and Helsinki. The POS tagging was

performed using the Brill tagger, with extensive manual post-editing (Arja Nurmi, p.c.). The corpus comes with an Associated Information File, which contains part of the metadata from the original CEEC; as the original (as yet unpublished) metadata is more complete and fine-grained, we will use it in our analysis.

The PCEEC is part of the English Parsed Corpora series, which was developed for the use of historical syntacticians (Taylor 2007). The focus of the annotation has been on sentential syntax, with POS tagging seen as a necessary step before parsing, and the lexis has not been normalised or lemmatised. To be applicable to the entire history of English, the POS annotation is very conservative with respect to e.g. adverbs that have grammaticalised during the recorded history of the language: for instance, *likewise* is tagged as a combination of an adjective and a noun (ADJ+N) rather than as an adverb (ADV; see further Säily et al. 2011). Moreover, the annotation scheme follows the *Cambridge Grammar of the English Language* (Huddleston & Pullum 2002) in that most subordinators are tagged as prepositions (Taylor & Santorini 2006). According to Huddleston & Pullum (2002: 598–601), prepositions may take both phrasal and clausal complements. This means that many "prepositions" in our data are words that in more traditional models of grammar are categorised as subordinators. This leads to severe problems when analysing the data in terms of complexity and colloquialisation, as we shall see below.

Säily et al. (2011) produced a new version of the PCEEC called the ReCEEC, which reclassified some of the items tagged as nouns into more appropriate categories. Combination tags such as ADJ+N were collapsed into the final tag except when reclassified: for example, gentleman_ADJ+N was collapsed into N (default case), but likewise_ADJ+N was turned into ADV (exception). Some retokenisation was also involved, such as separating articles and nouns written together (Säily et al. 2011: 174). Even though the changes are proportionally small, the present study utilises the ReCEEC because it provides a better description of the stages of English used in the corpus and is a better match to Present-day English corpora than the original PCEEC, making our study more comparable with e.g. Mair et al. (2002). To produce the final POS labels, the individual tags have been collapsed into somewhat larger categories loosely following Santorini (2016: POS annotation): adjectives, adverbial particles, adverbs, articles (which in this annotation scheme also include demonstrative determiners), BE verbs, complementisers, coordinating conjunctions, DO verbs, existential *there*, foreign words, HAVE verbs, modals, negations, nouns, (cardinal) numbers, other verbs, prepositions, (personal) pronouns and *wh*-words. As a heterogeneous and somewhat disputed category, 'quantifiers' has been left out, as have some very small categories (such as the words *one* and *other*) and some erroneous tags that do not belong to the tagset. The complete list of tags included in each of the categories is given in Appendix 1.

3.2 Visualisation

The PCEEC, with its metadata, is a complex dataset to understand. It spans over two and a half centuries and contains heterogeneous and unevenly distributed samples, which poses a challenge for many confirmatory statistical methods. Our approach in this paper is exploratory data analysis – we quantify and visualise the aspects of the PCEEC we are interested in, and use the pattern recognition capabilities of human vision to gain insight (cf. Siirtola et al. 2011).

Computationally, we subscribe to the *tidyverse* approach[2] developed by Hadley Wickham. We use *Statistical System R* (R Core Team 2016) packages *tidyr*, *dplyr*, and *stringr* to manipulate the data, and the package *ggplot2* to construct the visualisations. The computations are constructed from simple R operations and functions glued together with the pipe operator. In our analysis, we use what Hinneburg et al. (2007: 140) call "averaging the averages": we divide the corpus into samples, calculate the average frequency of the feature in question in each sample, and calculate the average of these averages for each time period of interest. Typically, each sample consists of a person's letters from a 20-year period, which enables us to account for variability both across and within individuals. On the one hand, our method ensures that the number of samples per person is

[2] http://tidyverse.org

low enough that each person has a similar impact on the results, regardless of the amount of data they have produced, so that an individual outlier cannot easily skew the results. On the other hand, the method takes into consideration possible change in the person's language use over time.

The most common visualisation type showing change over time is a scatter plot, with time on the *x*-axis and the measurement of interest on the *y*-axis. These plots are then divided into facets per measurement, and colour-coding is used to encode additional metadata. Uncertainty is indicated by 95% confidence intervals of the regression line in some plots.

4. Analysis

4.1 Complexity in the Parsed Corpus of Early English Correspondence

The changes in the frequencies of POS labels in the corpus are described in Figure 1. Based on the graphs, we can see some relatively clear trends that are relevant to the structural complexity of the texts. First of all, there are many categories that imply a decrease in the overall complexity of the genre over time: the proportions of nouns, complementisers and prepositions all show a downward trend. Taken together, these results may suggest a reduction in NP complexity in the most recent periods in particular, although the decreased proportion of complementisers and prepositions (a

category which also includes subordinators) may also imply that clausal complexity has been reduced to some extent. As discussed above, reduced NP complexity would suggest colloquialization, whereas a decrease in clausal complexity would not. There are also other trends worth noting that are difficult to interpret. For instance, *wh*-words show a slight increase over time, and there is a moderate increase in the proportion of adjectives in the data. As we are interested in establishing what kind of information POS ratios can provide to the study of complexity without parsing information, the results are inconclusive: there is no way, for example, to show that the increase of adjectives is connected to attribution (which would imply increased complexity and also provide counterevidence to the colloquialization hypothesis) instead of predication (which would not), and we likewise have no way of knowing whether the change in the frequency of *wh*-words is related to complexity-increasing structures and decreased orality (such as adverbial connectives or relative clauses; see Biber & Gray 2011: 18), or structures that are neutral with respect to complexity (such as interrogatives in main clauses; e.g. *Where is he?*). Nevertheless, the noticeable decrease in the high-frequency classes (nouns and prepositions) provides some support to the idea that NP complexity may have become reduced and the genre more speech-like in the period studied.

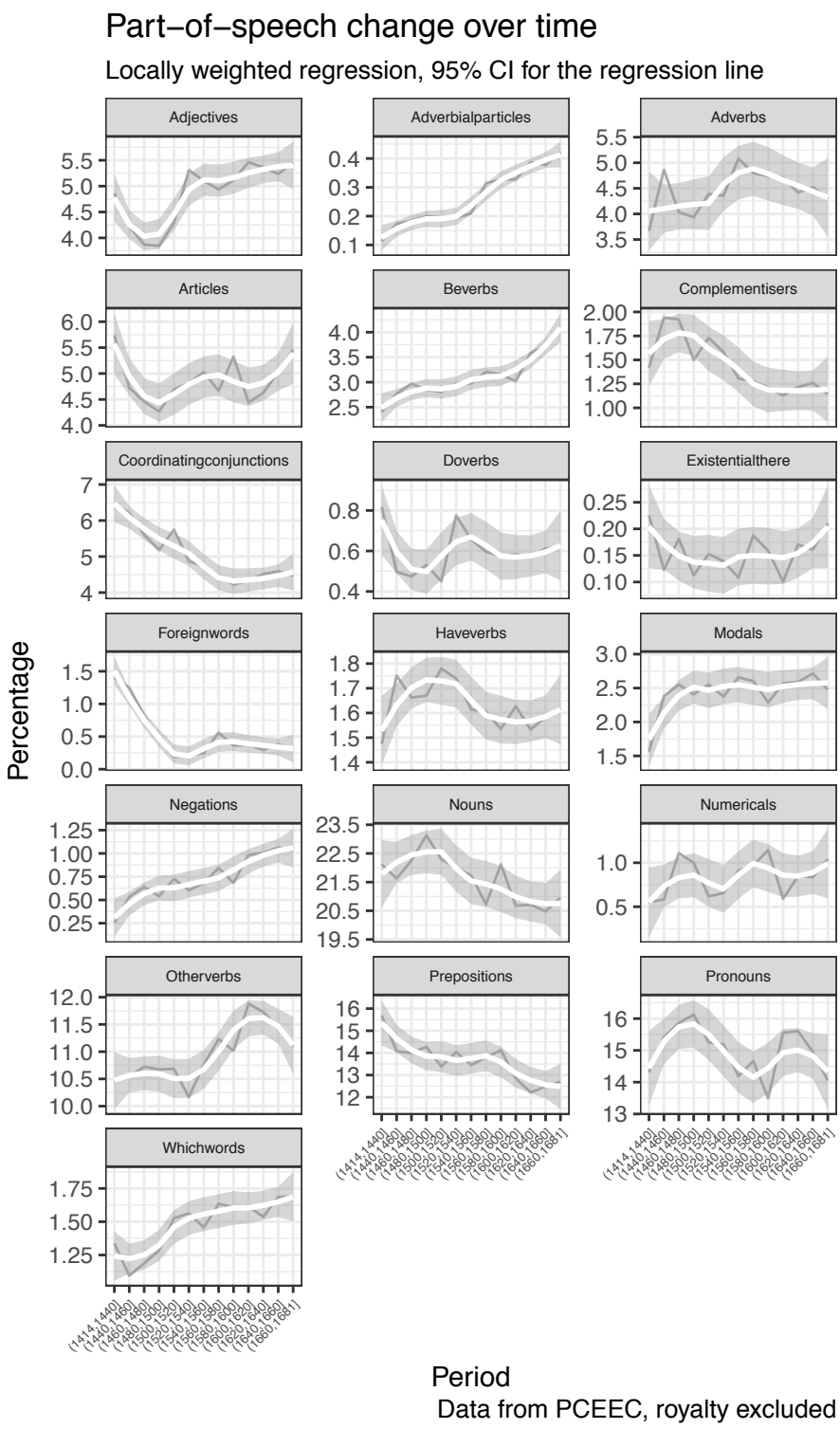

Figure 1. The proportions of parts of speech in the corpus over time.

Focusing on the more controversial markers of complexity, we can see that there is a noticeable decrease in the proportion of coordinating conjunctions over time.[3] As discussed above, in Biber's model (1992: 140) coordinators are considered a marker of low complexity, and from this perspective our result could imply that the genre has become more complex in the period studied. On the other hand, if we accept Lehto's suggestion (2015: 16, 139) that coordination may actually increase the complexity of texts, we should argue to the contrary. Bearing in mind that Lehto mainly based her arguments on the role of coordinators on sentence length, it is instructive to see how the texts in our data change according to this parameter. Figure 2 shows that sentence length has actually stayed roughly the same in the entire period studied. In short, there is no correlation between sentence length and the proportion of coordinators in our data, which suggests that the decreased frequency of coordinators is not a reflection of increased complexity (but neither does it suggest decreased complexity).[4]

[3] Kohnen's (2007) study of connectives in 15th and 16th century religious sermons revealed a similar decrease in the frequency of coordinators.

[4] We also studied POS ratios and sentence length from the perspective of social rank but found no clear results.

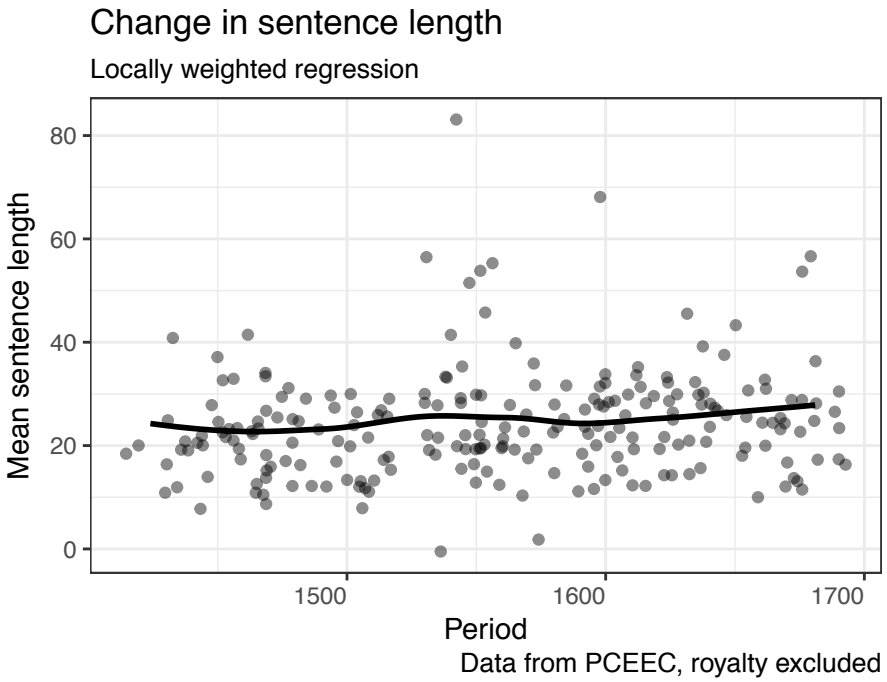

Figure 2. The mean sentence length in the PCEEC over time.

Considering the decrease in the proportion of coordinating conjunctions, and also in the proportion of prepositions, the fact that there is no change in sentence length over time is surprising. However, a closer look at the data shows that the decrease in the frequency of coordinators is at least in part due to a development which may not be very relevant from the perspective of complexity. Figure 3 shows that the decrease can be explained by the fact that the use of the sentence-initial coordinator *and* has plummeted in the period studied: while in the early 15th century c. eleven per cent of all sentences started with the coordinator *and*, the corresponding proportion in the 17th century data is c. four per cent.

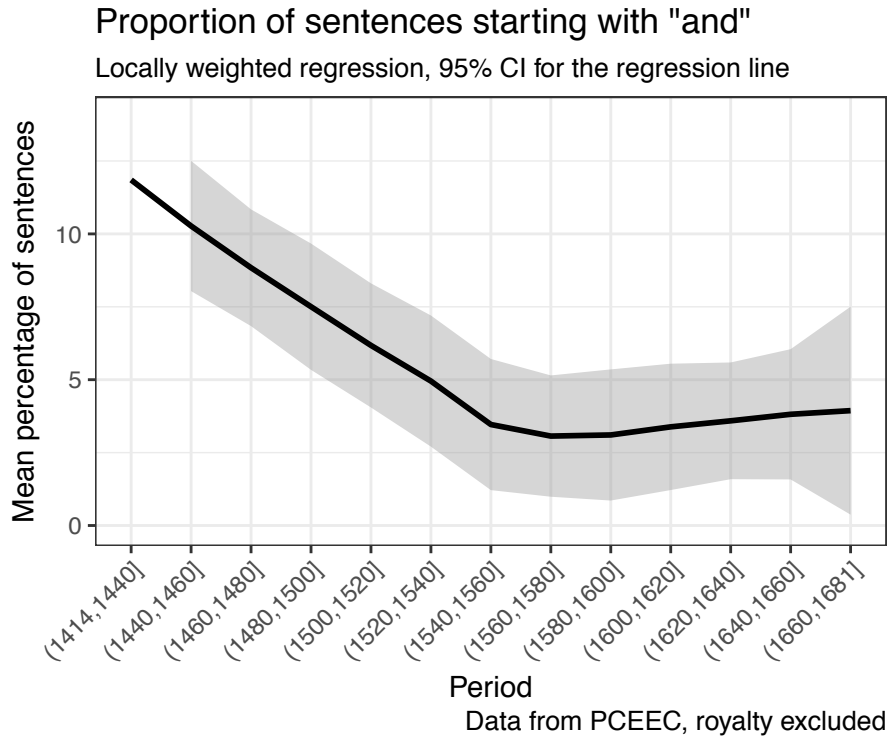

Figure 3. The frequency of sentence-initial *and* in the corpus over time.[5]

Sentence-initial use of *and* has in previous literature been linked to a text-organising function (e.g. Halliday & Hasan 1976: 235, 244) that is particularly typical of spoken language (Schiffrin 1987). Lehto (2015: 186) also found that sentence-initial *and* was used to indicate topic shifts in the 16th and 17th century legal texts, a function that in later periods was

[5] In our definition, "sentence" is an orthographic unit that ends in the following set of punctuation marks: {. ! ? : ;} When measuring sentence length, we improved the accuracy of the query by removing the most common abbreviations like *Mr.* from the results.

increasingly fulfilled by punctuation (colons and semi-colons in particular). In Lehto's data, the decreased frequency of sentence-initial *and* correlates relatively well with the increased frequency of punctuation, and in our data there is also a clear trend. However, although sentence-initial *and*, colons and semi-colons are used to indicate topic shifts in our data, they are more often used to organise the texts more subtly, indicating changes within the same topic. Figure 4 shows that as the frequency of sentence-initial *and* decreases, the frequencies of the colon and the semi-colon increase. Examples (1), (2) and (3), on the other hand, illustrate how sentence-initial *and*, colons and semi-colons are used to organise the texts and indicate shifts in the topics and sub-topics.

(1) Also my lady Clyfforde is sore syk of the ague and dropsey and is not lyke to lyve long as this berer will shewe your good lordshippe with oder thynges more at large. **And** thus our lorde Jh[es]u have your good lordshippe yn his blyssed kepyng, at London on Seynt Lukys day.

Sir Thomas Clifford to the first earl of Cumberland, 1526

(CLIFFO_024; Clifford, 72)

(2) This brother […] I brought up at school, the universities, and after maintained him in the warrs, so as he is risen to what he hath in lyvelihood by my means, and the tytle he hath, I also purchased for

him, besyds many other beniffits**:** this ungratefull man demanded a legacy of 300^{li} of me […]

John Holles to Lord Norris, 1617

(HOLLES_052; Holles, I, 164)

(3) They will themselves testify theire thanckfull myndes**;** I shall ever thincke my selfe beholdinge unto you, and rest readye to deserve your courtesyes, as good occasion shalbe offered.

Anne, countess dowager of Arundel, to Sir Thomas Edmondes, 1614

(ARUNDEL_014; Arundel, 87)

In (1), Sir Thomas Clifford proceeds from recounting the condition of Lady Clifford to the closing formula of the letter by using a sentence-initial *and*. In (2), on the other hand, the colon indicates a shift in perspective: first John Holles describes how he has provided financial support to a brother, then how the brother keeps asking him for more money. Finally, in (3) we see a shift from third-person reference to first-person reference, also indicating a change in perspective.

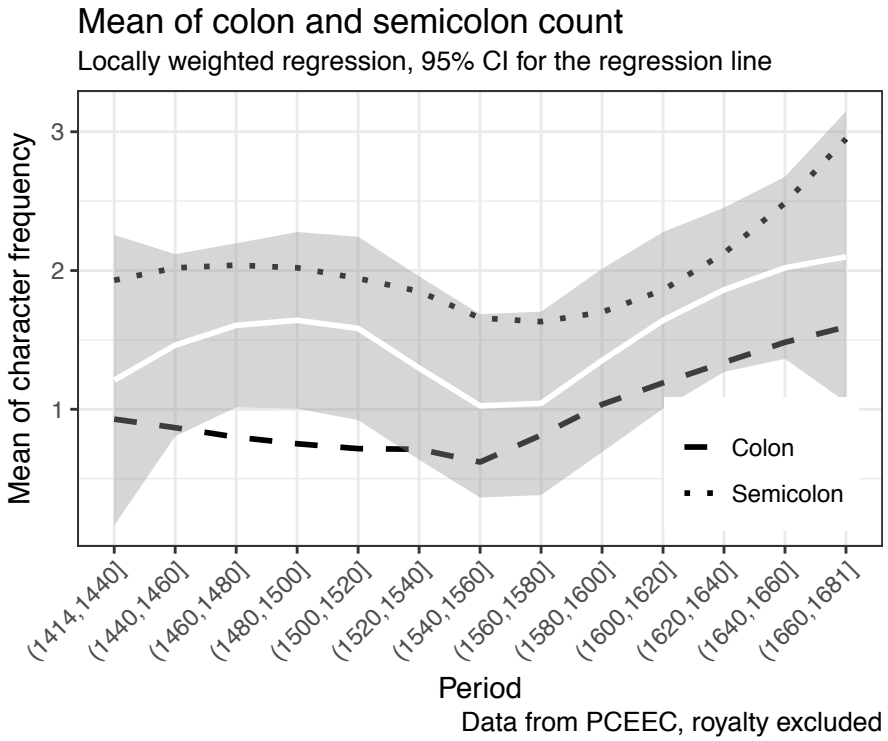

Figure 4. Change in the frequencies of the colon and the semi-colon in the corpus measured by the mean of character frequency.

We would therefore argue that the function of *and* was largely taken over by punctuation in the course of the period studied, and we interpret this result as having no effect on the overall complexity of the texts or the genre. As we have seen, the use of sentence-initial (or utterance-initial; see Meurman-Solin 2011) *and* has been regarded as a spoken feature and a marker of reduced complexity (Biber 1992). On the other hand, Lehto (2015) has argued that increased punctuation decreases complexity. Although Lehto and Biber disagree on the role of coordinators in structural complexity, their combined insights could explain the development seen in our data: if both

sentence-initial *and* and punctuation (colons and semi-colons) are regarded as markers of low complexity, it could be argued that the overall complexity of the genre remains unchanged by the development described above: one marker of low complexity has in part taken over the functional load that was previously associated with another marker of low complexity. Indeed, although this explanation rests on two different views of complexity, it should be pointed out that it is consistent with the fact that sentence length has remained unchanged in the period studied (Figure 2).

Based on the evidence, we conclude that using POS ratios to study changes in the structural complexity of private correspondence is not without problems. Although several POS labels suggested reduced complexity over time and a drift towards a more speech-like genre, a closer look at the data did not provide unequivocal support to this conclusion. Most significantly, given the decrease in the proportion of prepositions (which in this annotation scheme also includes subordinators), complementisers and coordinators, we would have expected to see a decrease in sentence length in the period studied as all the classes in question introduce new phrases and clauses, thus making the text more complex. However, this expectation was not borne out by the data. What we found instead was a change in the way in which the letters were structured: the overt marking of textual organisation and topic shifts with the coordinator *and* was gradually replaced by the increased use of punctuation, which we argue to be a neutral phenomenon in terms of the overall

complexity of the genre. As for the other markers of high complexity, we found that while the frequency of some parts of speech that are associated with increased complexity decreased (prepositions, complementisers, nouns), the frequency of others increased (*wh*-words, adjectives). Here, the interpretation of the data greatly suffers from the conflation of subordinators with prepositions in the tagset, on the one hand, and from the lack of parsing information, on the other. Therefore, we conclude that although POS ratios can shed some light on the development of the genre from the perspective of complexity, and they may suggest a certain degree of colloquialisation, the results remain largely inconclusive.

4.2 Colloquialisation and gendered styles

Returning to the overall picture of change in POS ratios (Figure 1), let us now focus on colloquialisation from the perspective of features indicating high involvement. As was already observed in Säily et al. (2011), there is a decrease in the proportion of nouns over time. Looking at the full inventory of parts of speech, we can also see a corresponding increase in the proportion of lexical verbs ('Otherverbs' in the figure) and BE verbs. A high frequency of verbs in general has been regarded as a feature of a more oral style (e.g. Mair et al. 2002: 247), and the increase in BE verbs might also be connected to the rise of the progressive aspect, which has been argued to indicate colloquialisation in previous literature (e.g. Smitterberg 2008).

While the other classes of verbs exhibit a more complex pattern, the overall situation seems to imply that a certain degree of colloquialisation in the sense of Mair et al. (2002) does take place in the corpus over time. Although the proportion of personal pronouns fluctuates with no clear trend, this category is probably too inclusive for our purposes: it is only the frequencies of first- and second-person pronouns that we would expect to increase in colloquialisation.

Are we able to detect which social groups lead this change? Figure 5 shows the data separated by social rank (gentry vs. non-gentry). For both nouns and lexical verbs, it is the gentry who are consistently in the lead. This makes sense: letters by the non-gentry, especially in the earlier periods, are perhaps more likely to deal with business issues and transmission of information, whereas the gentry could increasingly afford to write simply to keep in touch with friends and family, for which a more oral, involved style would be in order. Colloquialisation may also be seen in Nevala's (2004) study of address terms in family letters the CEEC: she found that the terms became increasingly intimate over time, especially in the 17th century.

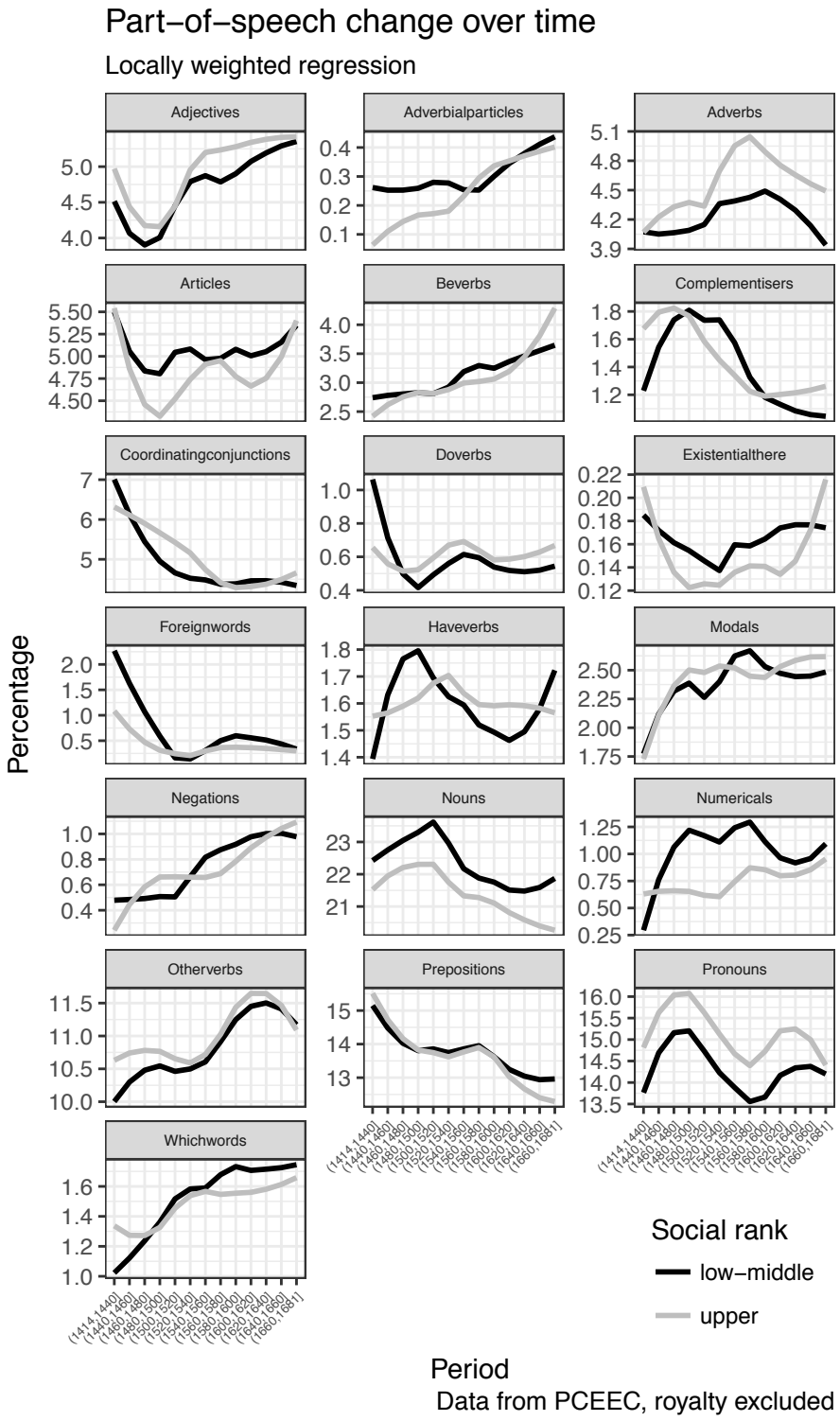

Figure 5. Change in POS ratios over time by social rank (low-middle = non-gentry, upper = gentry).

What about gender? Figure 6 shows that women appear to lead the change in the proportion of nouns, but there is no clear pattern in verbs. The only other category with a consistent pattern of variation, if not change, is that of personal pronouns, women consistently using them more than men, as already observed by Säily et al. (2011). However, the data are very heterogeneous. To make the data more comparable, we should account for audience design, or the relationship between the sender and recipient of the letter. Most of the women's letters are written to close family members, whereas men also have a fair number of letters to e.g. other acquaintances. Moreover, these proportions change over time in the corpus.

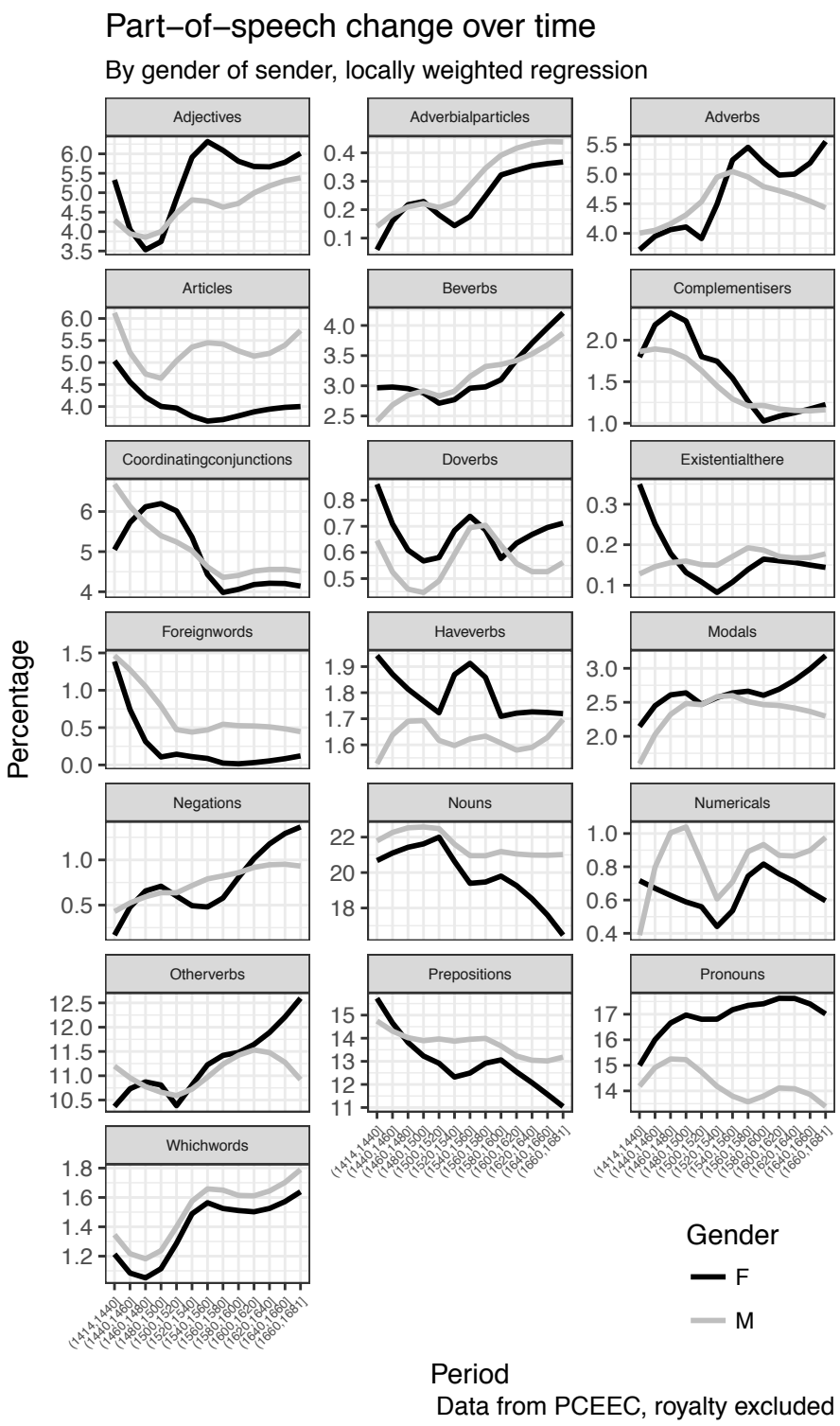

Figure 6. Variation and change in POS ratios over time by gender.

As we hypothesise that intimacy between the sender and recipient of the letter may have facilitated a colloquial style of writing and colloquialisation, let us take a more comparable sample of letters. We shall restrict the relationship between the sender and recipient to nuclear family only, and zoom in on the 17th century, from which we have more data from women. As noted in Section 3.1, royalty have been left out. Figure 7 shows the results. It is difficult to discern any clear changes, but several categories display consistent gender variation. Men tend to use more nouns, articles, prepositions, numerals and foreign words, while women use more personal pronouns, lexical verbs, BE verbs, DO verbs, modals and negations. These results are very similar to earlier findings regarding gendered styles in both historical and Present-day English (see Section 2.1). They are also a good match to several features along Biber's (1988) informational vs. involved dimension: nouns and prepositions belong to the informational pole, while (some) personal pronouns, verbs, modals and negations can be found on the involved pole. These results would then seem to lend strong support to the idea of relatively stable gendered styles that may span centuries (cf. Labov 1982: 38; 1990: 206–207; Nevalainen 2002: 191–194; Säily et al. 2011: 182; but see Vartiainen et al. 2013).

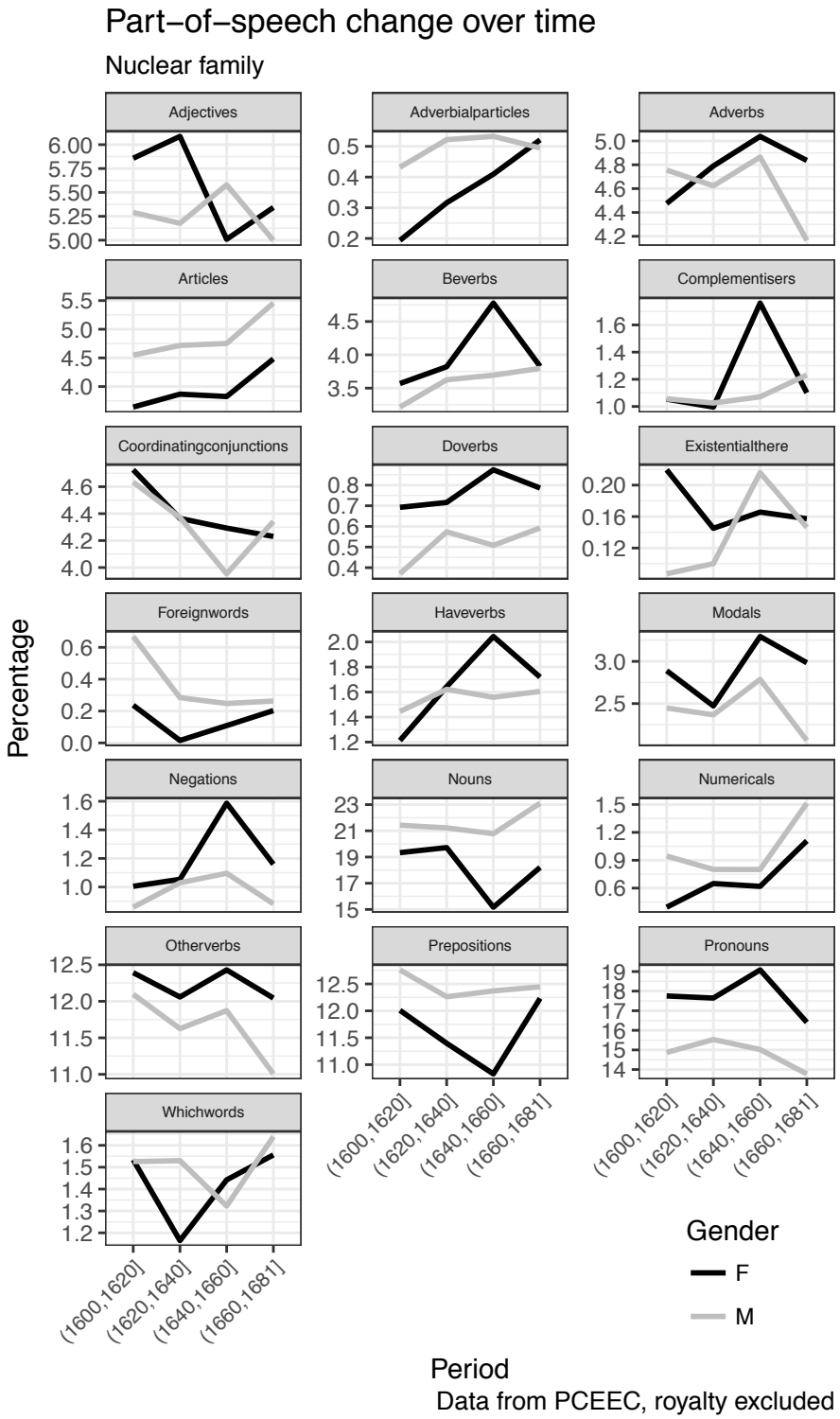

Figure 7. Variation and change in POS ratios by gender in 17th-century letters written to the nuclear family.

These data, however, are still not quite comparable across genders. Even within the nuclear family, we have multiple social roles – parents, children, spouses and siblings – and the proportions of these in the corpus change over time (see Vartiainen et al. 2013: 237–238). As we have the most data from women writing as wives, let us restrict our analysis further to spousal correspondence only. Owing to the relatively small amount of data, we need to use longer, 40-year time periods. The results can be seen in Figure 8. Here gender variation remains stable in some categories but is mixed or even reversed in others. Husbands tend to use more nouns, articles and prepositions, but there is a crossover in the category of numbers, and wives in fact use more foreign words than husbands in the first period. Wives, on the other hand, tend to use more personal pronouns, BE verbs, DO verbs, modals and negations, but husbands use slightly more lexical verbs than wives.

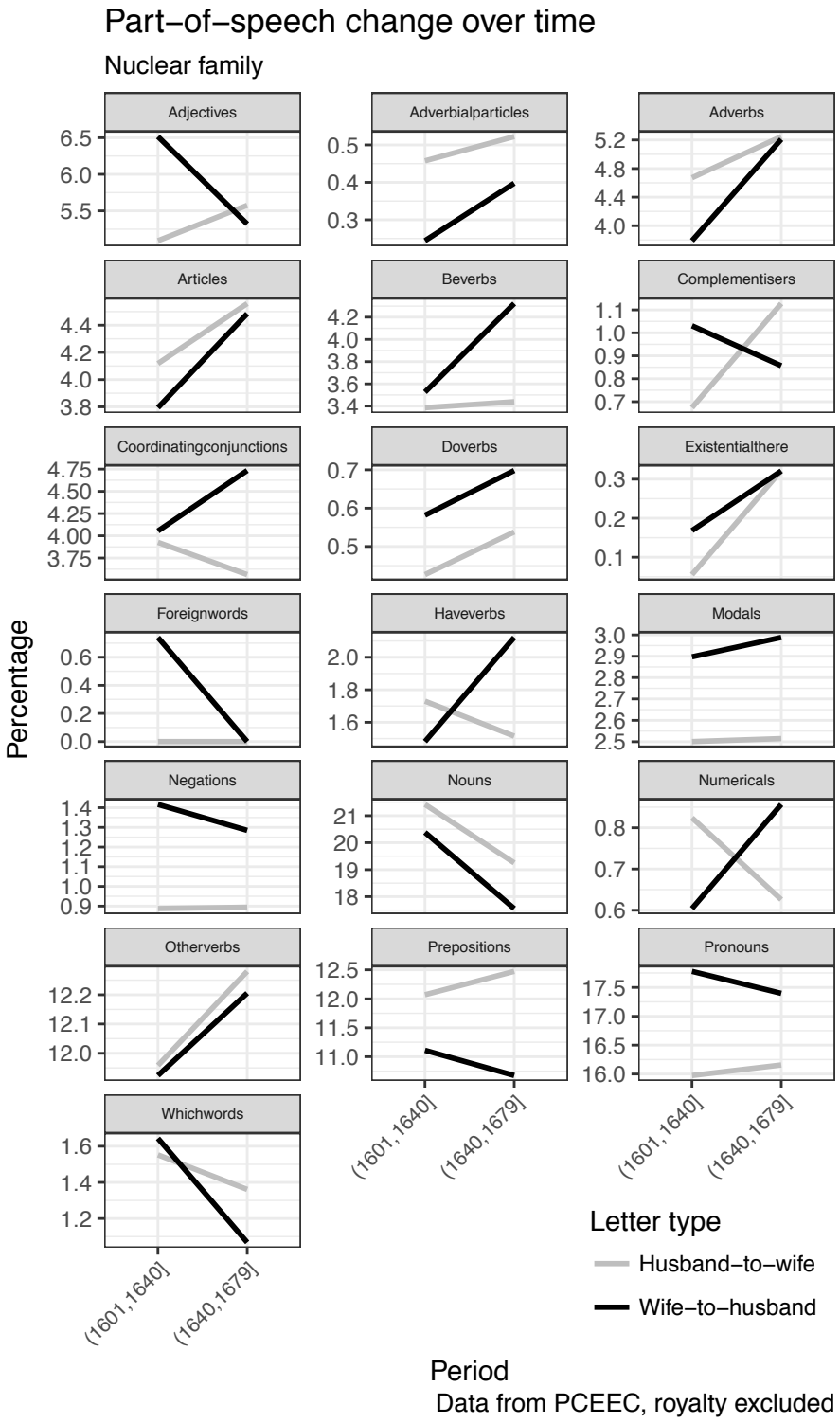

Figure 8. Variation and change in POS ratios in 17th-century spousal letters.

How can we explain these results? It is conceivable that some of our findings could be due to chance as the amount of data is relatively low. Nevertheless, the main indicators of style, nouns and pronouns, remain gendered over time. However, lexical verbs are a frequent enough category that there should be enough data to discern a pattern, and here we get the result that husbands tend to use them equally or even slightly more than wives. It may thus be that husbands writing to their wives use a somewhat more involved or oral style than men in general in terms of the proportion of verbs. As an example, let us take a look at an excerpt from a letter written in 1621 by gentleman John Hoskyns (1566–1638) to his wife Benedicta (lexical verbs in boldface).

(4) Good sweet hart this is whitsonday. I had busynesse heere till friday last & must be heere agayn on friday next. this morning I **promised** to be at Rochester for busynesse there to-morrow. I am now **goinge** to the Tilt boat & my ma~ **goes** about wth my horses. Yesterday I **tooke** phisicke, and I am well havin a little remnant of the Rheume **falling** down on a side tooth without payne. I will **see** y^{r} 2 sisters & y^{r} broth^{r} & **come** up again presently. I **meane** to be so fine as that they shall not **laugh** at y^{u} for having a sloven to y^{r} husband.

John Hoskyns to his wife, 1621

(HOSKYNS_020; Hoskyns, 88)

The style of example (4) is quite informal, and in addition to verbs, Hoskyns uses a great deal of personal pronouns, although the proportion of nouns is also fairly high. As noted by Vartiainen et al. (2013: 247), the husbands in these data tend to describe what they have been doing, whereas wives are perhaps more concerned with thoughts and feelings. This difference could be explored further by classifying the verbs in a more fine-grained manner: Palander-Collin (e.g. 1999, 2000) has discovered that it is private verbs such as *think* and *feel* that tend to be overused by women in the CEEC, and private verbs also head Biber's (e.g. 1988) list of involvement features. Example (5) from a letter written in 1627 by Lady Brilliana Harley to her husband, Sir Robert, illustrates wives' use of private verbs (lexical verbs in boldface).

(5) Deare S^{r} – Your two leters, on from Hearifort and the other from Gloster, weare uery wellcome to me: and if you **knwe** howe gladly I **reseaue** your leters, I **beleeue** you would neeuer **let** any opertunity **pase**. I **hope** your cloche did you saruis betwne Gloster and my brother Brays, for with vs it was a very rainy day, but this day has bine very dry and warme, and so I **hope** it was with you; and to-morowe I **hope** you will be well at your journis end, wheare I **wisch** my self to **bide** you wellcome home. You **see** howe my thoughts **goo** with you: and as you haue many of mine, so **let** me haue some of

yours. **Beleeue** me, I **thinke** I neuer **miste** you more then nowe I doo, or ells I haue **forgoot** what is **past**.

Lady Brilliana Harley to her husband, 1627

(HARLEY_004; Harley, 3)

To conclude, by studying POS ratios we have discovered that the correspondence genre seems to have undergone a degree of gentry-led colloquialisation in c.1410–1681. As for gender, we have found different results at different levels of granularity. At all levels, we find stable gender variation in the proportions of nouns and personal pronouns. In spousal letters of the 17th century, some of the other stylistic differences observed in a more heterogeneous sample disappear or display a mixed pattern over time. This could be due to the coarseness of POS ratios as a measure: in the verbal domain, the key difference lies in the use of private verbs rather than lexical verbs as a whole. Thus, POS ratios can be used to study colloquialisation and gendered styles to some extent, but for a more reliable and detailed analysis we need more fine-grained categories.

5. Discussion and conclusion

In our exploration of POS ratios in the PCEEC, we have analysed colloquialisation and colloquial style with regard to complexity and

involved text production. Our study of complexity had mixed results: some features could be connected to decreasing complexity, while others indicated increasing complexity. Furthermore, changes in some categories (most notably "prepositions") could not be interpreted in terms of colloquialisation: as we have no way of knowing whether the decrease in the proportion of "prepositions" is due to changes in the frequency of prepositions or subordinating conjunctions, our results remain ambiguous in this respect. Indeed, we would argue that although some linguistic analyses of word classes may be logical and theoretically plausible, they may turn out problematic when taken as the basis for POS annotation. In our case, it would have been very easy to extract subordinators and prepositions from the data and later collapse the two categories (if there had been a reason to do so). However, separating subordinators from prepositions would have required a great deal of of manual labour, and as the purpose of this paper was to explore the usefulness of POS labels without resorting to manual analysis, this was not done.

In order to say more about complexity and colloquialisation, we should look inside the superordinate POS categories, but even then we would not have all the information we need, e.g. at the level of syntax. Therefore, we must conclude that POS ratios can only be regarded as a heuristic tool in the study of linguistic complexity and that they should be complemented with other measures. Using some of Lehto's (2015) measures, we have been able to show that indicators of topic shift have

changed in a similar manner in both the legal statutes studied by her and in our correspondence corpus: there is a decrease in the frequency of sentence-initial *and* along with an increase in the frequency of the colon and the semi-colon. Although we maintain that this change had little or no effect on the overall complexity of the genre, it is of course true that the replacement of sentence-initial *and* by punctuation is a development from a more "oral" to a more "written" variety, and in this sense it could be considered a change towards a less colloquial style. As our corpus is based on published editions of letters, we acknowledge that the ostensible changes in punctuation may also have been influenced by editorial practices (cf. Raumolin-Brunberg & Nevalainen 2007); however, as Lehto (2015: 81) obtained similar results using original printed material, it is unlikely that our results are entirely an artefact of editorial interference.

Our study of colloquialisation in terms of oral or involved style was perhaps more successful in that many of the POS categories – particularly nouns, verbs and personal pronouns – did seem to be directly related to style. Our results support previous research on gendered styles, providing a more complete picture of the letter genre than was previously available (e.g. Säily et al. 2011, who only analysed nouns and personal pronouns). Moreover, we have been able to extend Biber & Finegan's (1989) study into the past: our results indicate that correspondence seems to have undergone a gentry-led process of colloquialisation (in terms of the frequencies of nouns and verbs) before the reversal in the 18th century observed by them.

Nevertheless, it is clear that POS ratios do not tell the whole story: even though the POS annotation of verbs is relatively fine-grained in the PCEEC, it does not capture categories like private verbs or the progressive aspect, both of which have been linked to colloquial or involved style. Furthermore, the very general result of change led by the upper ranks should be complemented by a more detailed inquiry that accounts for e.g. audience design, as was done in our analysis of gender variation.

As argued in Hardie (2007), annotation principles crucially affect the kinds of research questions that can be explored by studying POS ratios, as well as the answers that can be obtained. For example, as was already discussed, because the PCEEC follows Huddleston & Pullum's (2002) analysis of prepositions in its classification, we were not able to study either subordinators or prepositions as separate categories. Furthermore, annotation schemes should in our opinion be relatively fine-grained: in the study of gendered styles, the category of personal pronouns should ideally be divided by person, number and gender, especially in a historical corpus where division by lexical form is not so straightforward owing to spelling variation. The CLAWS tagset,[6] for instance, has become steadily more detailed over time, but the more fine-grained divisions can easily be collapsed into superordinate ones if desired. From the perspective of complexity and colloquialisation, it would also be useful to have separate

[6] http://ucrel.lancs.ac.uk/claws/

tags for attributive and predicative adjectives, on the one hand, and for different kinds of complement clauses (e.g. *that*-clauses as complements of nouns and verbs), on the other, although this would introduce parsing information into the POS tags, which may not be desirable.

In any linguistic study, we believe that it is important to go back to the texts. When exploring something as general as POS ratios, it becomes especially important to interpret our results through close reading. In historical sociolinguistics, we also need to pay attention to the individuals and social groups behind the variation and change. This raises a methodological issue, as figures and tables are usually static and do not provide access to the texts and metadata on which they are based. In future research, the exploratory approach taken in this paper could be further enhanced by interactive visualisation: linking the texts, metadata, visualisations and statistical analyses to each other would greatly facilitate work in historical sociolinguistics. We are already working on this in a project led by Terttu Nevalainen: a second version of our Text Variation Explorer tool (Siirtola et al. 2014, 2016) will come out in 2017, and another tool for historical sociolinguistic research called Khepri (Mäkelä et al. 2016) will be available by the end of 2018.

In addition to interactive visualisation, variation and change in POS ratios could in the future be studied through more advanced statistical methods. Promising avenues to explore include multilevel Bayesian modelling (Carpenter et al. 2017) as well as machine learning techniques

such as subgroup discovery (Atzmueller 2015). While Labov's (1994: 11) famous bad-data problem will always be with us, we will continue to strive to make "the best use of bad data" using state-of-the-art tools and methods in collaboration with experts from other fields. We believe that this is the way forward for historical sociolinguistics.

Acknowledgements

This work was supported in part by the Academy of Finland grant 276349 to the project 'Reassessing language change: the challenge of real time', and by the Academy of Finland Digital Humanities Programme, project 'Interfacing structured and unstructured data in sociolinguistic research on language change (STRATAS)', grant 293441.

References

Argamon, Shlomo, Moshe Koppel, Jonathan Fine & Anat Rachel Shimoni. 2003. Gender, genre, and writing style in formal written texts. *Text* 23(3). 321–346.

Atzmueller, Martin. 2015. Subgroup discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 5(1). 35–49.

Bamman, David, Jacob Eisenstein & Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics* 18(2). 135–160.

Bell, Allan. 1984. Language style as audience design. *Language in Society* 13(2). 145–204.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas. 1992. On the complexity of discourse complexity: A multidimensional analysis. *Discourse Processes* 15(2). 133–163.

Biber, Douglas. 1995. *Dimensions of register variation*. Cambridge: Cambridge University Press.

Biber, Douglas & Jena Burges. 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, Douglas & Susan Conrad. 2009. *Register, genre, and style* (Cambridge Textbooks in Linguistics). Cambridge: Cambridge University Press.

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

Biber, Douglas & Edward Finegan. 1997. Diachronic relations among speech-based and written registers in English. In Terttu Nevalainen & Leena Kahlas-Tarkka (eds.), *To explain the present: Studies in the changing English language in honour of Matti Rissanen* (Mémoires

de la Société Néophilologique de Helsinki 52), 253–275. Helsinki: Société Néophilologique.

Biber, Douglas & Bethany Gray. 2010. Being specific about historical change: The influence of sub-register. *Journal of English Linguistics* 41(2). 104–134.

Biber, Douglas & Bethany Gray. 2011. The historical shift of scientific academic prose in English towards less explicit styles of expression: Writing without verbs. In Vijay Bhatia, Purificación Sánchez Hernández & Pascual Pérez-Paredes (eds.), *Researching specialized languages* (Studies in Corpus Linguistics 47), 11–24. Amsterdam: John Benjamins.

Biber, Douglas, Bethany Gray & Shelley Staples. 2016. Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics* 37(5). 639–668.

Carpenter, Bob, Andrew Gelman, Matt Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li & Allen Riddell. 2017. Stan: A probabilistic programming language. *Journal of Statistical Software* 76(1).

Chafe, Wallace. 1982. Integration and involvement in speaking, writing, and oral literature. In Deborah Tannen (ed.), *Spoken and written language*, 35–53. Norwood, NJ: Ablex.

Halliday, M.A.K. & Ruqaiya Hasan. 1976. *Cohesion in English*. London & New York: Longman.

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

Heylighen, Francis & Jean-Marc Dewaele. 2002. Variation in the contextuality of language: An empirical measure. *Foundations of Science* 7(3). 293–340.

Hinneburg, Alexander, Heikki Mannila, Samuli Kaislaniemi, Terttu Nevalainen & Helena Raumolin-Brunberg. 2007. How to handle small samples: Bootstrap and Bayesian methods in the analysis of linguistic change. *Literary and Linguistic Computing* 22(2). 137–150.

Huddleston, Rodney & Geoffrey K. Pullum (eds.). 2002. *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.

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

Karlsson, Fred. 2008. Complexity in linguistic theorizing. *The Mental Lexicon* 9(2). 144–169.

Kohnen, Thomas. 2007. 'Connective profiles' in the history of English texts. Aspects of orality and literacy. In Ursula Lenker & Anneli Meurman-Solin (eds.), *Connectives in the history of English*, 289–308. Amsterdam: John Benjamins.

Labov, William. 1982. Building on empirical foundations. In Winfred P. Lehmann & Yakov Malkiel (eds.), *Perspectives on historical*

linguistics: Papers from a conference held at the meeting of the Language Theory Division, Modern Language Assn, San Francisco, 27–30 December 1979 (Current Issues in Linguistic Theory 24), 17–92. Amsterdam: John Benjamins.

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

Labov, William. 1994. *Principles of linguistic change, volume 1: Internal factors*. Oxford: Blackwell.

Laslett, Peter. 1965. *The world we have lost*. New York: Charles Scribner's Sons.

Lehto, Anu. 2015. *The genre of Early Modern English statutes: Complexity in historical legal language* (Mémoires de la Société Néophilologique de Helsinki 97). Helsinki: Société Néophilologique.

Mair, Christian, Marianne Hundt, Geoffrey Leech & Nicholas Smith. 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.

Mäkelä, Eetu, Tanja Säily & Terttu Nevalainen. 2016. Khepri – a modular view-based tool for exploring (historical sociolinguistic) data. Presentation, *Digital Humanities 2016*, Kraków, 11–16 July 2016.

Markus, Manfred. 2001. The development of prose in Early Modern English in view of the gender question: Using grammatical idiosyncracies of 15th and 17th century letters. *European Journal of English Studies* 5(2). 181–196.

Meurman-Solin, Anneli. 2011. Utterance-initial connective elements in early Scottish epistolary prose. In Anneli Meurman-Solin & Ursula Lenker (eds.), *Connectives in synchrony and diachrony in European languages* (Studies in Variation, Contacts and Change in English 8). Helsinki: VARIENG. http://www.helsinki.fi/varieng/series/volumes/08/meurman-solin/ (17 December, 2016.)

Nevala, Minna. 2004. *Address in early English correspondence: Its forms and socio-pragmatic functions* (Mémoires de la Société Néophilologique de Helsinki 64). Helsinki: Société Néophilologique.

Nevalainen, Terttu. 2002. Language and woman's place in earlier English. *Journal of English Linguistics* 30(2). 181–199.

Nevalainen, Terttu & Helena Raumolin-Brunberg. 2003. *Historical sociolinguistics: Language change in Tudor and Stuart England* (Longman Linguistics Library). London: Pearson Education.

Newman, Matthew L, Carla J. Groom, Lori D. Handelman & James W. Pennebaker. 2008. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes* 45(3). 211–236.

Palander-Collin, Minna. 1999. *Grammaticalization and social embedding: I THINK and METHINKS in Middle and Early Modern English* (Mémoires de la Société Néophilologique de Helsinki 55). Helsinki: Société Néophilologique.

Palander-Collin, Minna. 2000. The language of husbands and wives in seventeenth-century correspondence. In Christian Mair & Marianne Hundt (eds.), *Corpus linguistics and linguistics theory. Papers from the twentieth International Conference on English Language Research on Computerized Corpora (ICAME 20), Freiburg im Breisgau 1999* (Language and Computers: Studies in Practical Linguistics 33), 289–300. Amsterdam: Rodopi.

PCEEC = *Parsed Corpus of Early English Correspondence*, tagged version. 2006. Annotated by Arja Nurmi, Ann Taylor, Anthony Warner, Susan Pintzuk & Terttu Nevalainen. Compiled by the CEEC Project Team. York: University of York & Helsinki: University of Helsinki. Distributed through the Oxford Text Archive. http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/ (17 December, 2016.)

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.

R Core Team. 2016. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. http://www.r-project.org (17 December, 2016.)

Raumolin-Brunberg, Helena & Terttu Nevalainen. 2007. Historical sociolinguistics: The Corpus of Early English Correspondence. In Joan C. Beal, Karen P. Corrigan & Hermann L. Moisl (eds.), *Creating and digitizing language corpora, volume 2: Diachronic databases*, 148–171. Houndsmills: Palgrave Macmillan.

Rayson, Paul, Geoffrey Leech & Mary Hodges. 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.

Rescher, Nicholas. 1998. *Complexity: A philosophical overview*. New Brunswick, NJ: Transaction Publishers.

Säily, Tanja, Terttu Nevalainen & Harri Siirtola. 2011. Variation in noun and pronoun frequencies in a sociohistorical corpus of English. *Literary and Linguistic Computing* 26(2). 167–188.

Santorini, Beatrice. 2016. *Annotation manual for the Penn Historical Corpora and the York-Helsinki Corpus of Early English Correspondence*. http://ling.upenn.edu/hist-corpora/annotation/ (17 December, 2016.)

Schiffrin, Deborah. 1987. *Discourse markers*. Cambridge: Cambridge University Press.

Siirtola, Harri, Poika Isokoski, Tanja Säily & Terttu Nevalainen. 2016. Interactive text visualization with Text Variation Explorer. In Ebad Banissi, Mark W. McK. Bannatyne, Fatma Bouali, Remo Burkhard, John Counsell, Urska Cvek, Martin J. Eppler, Georges Grinstein, Wei Dong Huang, Sebastian Kernbach, Chun-Cheng Lin, Feng Lin, Francis T. Marchese, Chi Man Pun, Muhammad Sarfraz, Marjan Trutschl, Anna Ursyn, Gilles Venturini, Theodor G. Wyeld & Jian J. Zhang (eds.), *Proceedings of the 20th international conference on Information Visualisation (IV 2016)*, 330–335. Los Alamitos, California: IEEE Computer Society.

Siirtola, Harri, Terttu Nevalainen, Tanja Säily & Kari-Jouko Räihä. 2011. Visualisation of text corpora: A case study of the PCEEC. In Terttu Nevalainen & Susan M. Fitzmaurice (eds.), *How to deal with data: Problems and approaches to the investigation of the English language over time and space* (Studies in Variation, Contacts and Change in English 7). Helsinki: VARIENG. http://www.helsinki.fi/varieng/series/volumes/07/siirtola_et_al/ (17 December, 2016.)

Siirtola, Harri, Tanja Säily, Terttu Nevalainen & Kari-Jouko Räihä. 2014. Text Variation Explorer: Towards interactive visualization tools for corpus linguistics. *International Journal of Corpus Linguistics* 19(3). 417–429.

Smitterberg, Erik. 2008. The progressive and phrasal verbs: Evidence of colloquialization in nineteenth-century English? In Terttu Nevalainen, Irma Taavitsainen, Päivi Pahta & Minna Korhonen (eds.), *The dynamics of linguistic variation: Corpus evidence on English past and present* (Studies in Language Variation 2), 269–289. Amsterdam: John Benjamins.

Tannen, Deborah. 1991. *You just don't understand: Women and men in conversation*. New York: Morrow and Company.

Taylor, Ann. 2007. The York-Toronto-Helsinki Parsed Corpus of Old English Prose. In Joan C. Beal, Karen P. Corrigan & Hermann L. Moisl (eds.), *Creating and digitizing language corpora, volume 2: Diachronic databases*, 196–227. Houndsmills: Palgrave Macmillan.

Taylor, Ann & Beatrice Santorini. 2006. The Parsed Corpus of Early English Correspondence. University of York. http://www-users.york.ac.uk/~lang22/PCEEC-manual/ (17 December, 2016.)

Vartiainen, Turo, Tanja Säily & Mikko Hakala. 2013. Variation in pronoun frequencies in early English letters: Gender-based or relationship-based? In Jukka Tyrkkö, Olga Timofeeva & Maria Salenius (eds.), *Ex philologia lux: Essays in honour of Leena Kahlas-Tarkka* (Mémoires de la Société Néophilologique de Helsinki 90), 233–255. Helsinki: Société Néophilologique.

Appendix 1. Superordinate POS labels

The ReCEEC POS tags have been collapsed into the following superordinate word classes. For definitions of the POS tags, see Santorini (2016).

- Adjectives: ADJ, ADJR, ADJS
- Adverbial particles: RP
- Adverbs: ADV, ADVR, ADVS
- Articles: D
- BE verbs: BAG, BE, BED, BEI, BEN, BEP
- Complementisers: C
- Coordinating conjunctions: CONJ
- DO verbs: DAG, DAN, DO, DOD, DOI, DON, DOP
- Existential *there*: EX
- Foreign words: FW
- HAVE verbs: HAG, HAN, HV, HVD, HVI, HVN, HVP
- Modals: MD, MD0
- Negations: NEG
- Nouns: N, N$, NPR, NPR$, NPRS, NPRS$, NS, NS$
- Numbers: NUM, NUM$
- Other verbs: VAG, VAN, VB, VBD, VBI, VBN, VBP

- Prepositions: P
- Pronouns: PRO, PRO$
- *wh*-words: WADV, WD, WPRO, WPRO$, WQ

- Excluded: ADJP, ADJX, ADVP, ADVP-LOC, ADVP-TMP, ADVX, ALSO, CIPHER, CONJP, DET, ELSE, FOR, FOREIGN, FP, FRAG, INTJ, IP-PPL, LS, NNP-PRN, NP, NP-COM, NP-MSR, NP-SBJ, NP-VOC, NUMP, NX, ONE, ONE$, ONES, ONES$, OTHER, OTHER$, OTHERS, OTHERS$, PP, Q, Q$, QP, QR, QS, RRC, SUCH, TO, VP, WADVP, WARD, WNP, X, XX