

Variation in pronoun frequencies in early English letters: Gender-based or relationship-based?

Turo Vartiainen, Tanja Säily and Mikko Hakala
University of Helsinki

This is a pre-print version of the following paper: Vartiainen, Turo, Tanja Säily & Mikko Hakala. 2013. "Variation in pronoun frequencies in early English letters: Gender-based or relationship-based?" *Ex Philologia Lux: Essays in Honour of Leena Kahlas-Tarkka* (Mémoires de la Société Néophilologique de Helsinki XC), ed. by Jukka Tyrkkö, Olga Timofeeva & Maria Salenius. Helsinki: Société Néophilologique, 233–255.

Abstract

Using the *Corpora of Early English Correspondence*, 1600–1800, we analyse the extent to which the frequencies of first- and second-person pronouns are affected by the relationship between the sender and the recipient of the letter, comparing the effect to that of gender. While we find significant gender differences in our data, our results also pose some challenges to earlier studies, which interpreted differences in pronoun usage in terms of gendered styles of writing (informational vs. involved). Our data show that the differences in pronoun frequencies are not stable across time, the results are inconclusive in some categories, and there are statistically significant intra-gender differences depending on the relationship between the writer and the recipient. We also find that the balance of the corpus may affect the results to some degree and emphasise that close attention should be paid to the nature of the relationship between the writer and the recipient in future work.

1. Introduction

Several studies have found significant differences in the way men and women use nouns and personal pronouns (e.g. Rayson et al. 1997, Argamon et al. 2003). In a recent paper Säily et al. (2011) studied the letters in the *Parsed Corpus of Early English Correspondence* (PCEEC) and found that men used more nouns than women in all the periods studied (c.1415–1681), while women consistently used more pronouns than men. This was explained by the notion of gendered styles: it was suggested that women's writing style was more involved, while men's style was more informational.

However, Säily et al. largely ignored the effect of register (Biber 1988), audience design (Bell 1984) and corpus balance: most of the female-authored letters in the corpus were written to family members, whereas the majority of male-authored letters were written to acquaintances. Therefore, it is possible that the gender difference is conflated with differences in register and audience. Furthermore, Säily et al. treated all personal pronouns as a single category, whereas Biber (1988) has argued that only first- and second-person pronouns are indicative of an involved style, while third-person pronouns are a feature of narrative discourse.

The present study analyses the extent to which pronoun frequencies are affected by the relationship between the sender and the recipient of the letter, comparing the effect to that of gender. We focus on letters written to close family members and to other acquaintances, two categories presumed to be maximally apart in terms of the

level of intimacy. To assess the explanation proposed for the gender difference by Säily et al. (2011), i.e. involved vs. informational style, we restrict our analysis to first- and second-person pronouns. Our material consists of 17th- and 18th-century letters in the *Corpora of Early English Correspondence* (CEEC), covering the years 1600–1800. We also discuss the possible effect of corpus balance on the results.

The paper is organized as follows. Section 2 surveys previous research and justifies our decision to concentrate on the 17th and 18th centuries. Section 3 introduces the data used in the study, while section 4 describes our methods. The results of the study are presented in section 5 and discussed further in section 6. Section 7 concludes the paper with a summary of our findings.

2. Background

A tendency for men to overuse nouns and for women to overuse personal pronouns has been observed for Late Middle and Early Modern English letters in the PCEEC by Säily et al. (2011), and for spoken and written genres of present-day English in the *British National Corpus* (BNC) by Rayson et al. (1997) and Argamon et al. (2003), respectively. Argamon et al. (2003) and Säily et al. (2011) interpret this tendency in terms of Biber's (1988) multidimensional analysis of register variation. Biber defines the different dimensions of register variation based on the co-occurrence patterns of a number of linguistic features in texts. In addition to register variation, the dimensions have been used as indicators of communicative styles (e.g. Biber & Burges 2000). Previous research has paid particular attention to Biber's Dimension 1, which is labelled Informational vs. Involved Production. One of the key features indicative of informational style is a high frequency of nouns, whereas the involved (interactive and affective) style is characterised by, e.g., first- and second-person pronouns. Thus, the observation that men overuse nouns and women overuse personal pronouns can be explained in terms of gendered styles: men's style is more informational, whereas women's style is more involved.¹

These large-scale quantitative studies of nouns and pronouns are complemented by studies focusing on individual pronouns such as *I* (Palander-Collin 2009) or *thou* and *you* (Walker 2007), pronouns in specific functions, such as address terms (Nevala 2004), and specific constructions, such as *I THINK* (Palander-Collin 1999). Palander-Collin (1999) explores male and female styles in 17th-century correspondence of the upper ranks and finds that the overuse of *I THINK* by women persists even when the relationship between the sender and the recipient is taken into account. She, too, interprets the gender difference in terms of informational vs. involved styles. The 17th century would then seem to be a good starting point for our study of gender- and relationship-based variation in the use of first- and second-person pronouns in general.²

¹ We are, of course, aware that Biber's dimensions are based on the co-occurrence of several linguistic features in addition to first- and second-person pronouns. By focusing on these pronouns, we are only exploring one aspect of the Involved vs. Informational dimension.

² While the scope of our brief paper is more limited than that of Säily et al. (2011) in that we do not cover the fifteenth and sixteenth centuries, we follow their lead in

Our decision to extend our study into the 18th century is based on intriguing results from previous research. According to McIntosh (2008: 231), 18th-century British culture was “feminized”: there were more female authors than ever before, and politeness and sensibility were required of anyone aspiring to be counted among the upper classes (see also Tieken-Boon van Ostade 2010, Nevalainen & Tissari 2010). Moreover, as the number of social aspirers, such as wealthy merchants and manufacturers, increased during this period, the definition of a gentleman was loosened so that landownership was no longer essential (Hay & Rogers 1997: 18–24, Fitzmaurice 2012). This blurring of lines – gentlewomen somewhat more on a par with gentlemen in terms of education and literary influence, and gentry a more inclusive category than before – may have had an effect on language use as well.

Indeed, Biber & Finegan (1997) find that 18th-century letters in the ARCHER corpus are more involved than 17th-century letters, and Palander-Collin (2009) discovers a rise in the frequency of *I* in gentlemen’s letters between the 16th and 18th centuries, attributing this to an increasingly involved style. Furthermore, Säily (forthcoming) finds that the gender difference observed in the use of the nominal suffix *-ity* in the 17th-century section of the CEEC disappears in the 18th-century section. Could it be that there is no gender difference in the use of personal pronouns, either? Contrary to this hypothesis, Nurmi & Palander-Collin’s (2008) study of gender- and relationship-based variation in the use of first- and second-person pronouns in an 18th-century sample of the CEEC finds significant overuse in letters written by women as well as in letters written to family members, especially between family correspondents with equal social status. Let us now explore the situation in the full CEEC.

3. Data

The material for our study comes from the *Corpus of Early English Correspondence*, the *Supplement* to the original corpus and the 18th-century *Extension* of the corpus, together forming a 5.1-million-word collection of personal letters and covering four centuries from 1400 to 1800.³ The corpora have been compiled as a resource for sociolinguistic research, and they are accompanied by an extensive database of background information. This includes metadata about the authors and recipients (e.g. gender, place of birth, social rank) as well as the letters themselves (e.g. time and place of writing).

The letters in the corpus can be divided into five categories on the basis of the relationship between the author and the recipient: letters written by or addressed to the members of the nuclear family (FN), other family members (FO), family servants (FS), close friends (TC), and “others” (T), a category comprising all other acquaintances. Table 1 shows the amount of data available from female and male writers within each relationship category (as already noted in the introduction, we are mainly interested in pronoun use in letters belonging to the FN and T categories). The periodization used in table 1 roughly reflects the cut-off point of 1681 in Säily et al.

analysing our data set in its entirety rather than limiting it to a specific social rank, and leave rank-based variation for future research.

³ For the sake of clarity, we will refer to the entire family of corpora as CEEC. See CoRD (<http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/>) for more details.

(2011) as well as the transition from the original CEEC to the long 18th century covered by the CEECE.⁴

Table 1. Number of words in different relationship categories in the CEEC.

Category	Women				Entire period
	1600–1679		1680–1800		
FN (Nuclear family)	180,188	53%	265,490	44%	445,678
FO (Other family)	71,640	21%	92,663	15%	164,303
FS (Family servants)	3,486	1%	0	0%	3,486
TC (Close friends)	34,509	10%	144,869	24%	179,378
T (Others)	46,987	14%	103,354	17%	150,341
Total	336,810		606,376		943,186

Category	Men				Entire period
	1600–1679		1680–1800		
FN (Nuclear family)	275,754	21%	463,970	29%	739,724
FO (Other family)	215,638	17%	117,180	7%	332,818
FS (Family servants)	22,400	2%	2,669	0%	25,069
TC (Close friends)	201,622	15%	507,958	32%	709,580
T (Others)	585,579	45%	517,976	32%	1,103,555
Total	1,300,993		1,609,753		2,910,746

In addition to comparing letters written to family members and to more distant acquaintances, we will take a closer look at the letters within the FN category. These can be further categorized according to the roles of the author and the recipient within the family unit (e.g. husband > wife, father > son). Table 2 provides a breakdown of the data along these parameters. This data set does not include the three outliers excluded from the analysis or members of the royal family (see section 4).

Unfortunately the scope of this paper does not permit a detailed discussion of each category, and so we will focus on pronoun use in the correspondence of fathers and sons, mothers and sons, and husbands and wives (highlighted in table 2) in sections 5.2 and 5.3 below.

Table 2. Number of words in letters written to close family members in our data set.

Relationship	Women				Entire period
	1600–1679		1680–1800		
mother > son	41,734	47 %	18,902	8 %	60,636 18 %
mother > daughter	2,115	2 %	37,489	16 %	39,604 12 %
daughter > father	1,386	2 %	18,101	8 %	19,487 6 %
daughter > mother	5,352	6 %	0	0 %	5,352 2 %
sister > brother	1,993	2 %	21,602	9 %	23,595 7 %

⁴ We will introduce our results in sections 4 and 5 by comparing these two periods or, to allow for more detailed scrutiny of the diachronic developments, by using 40-year periods.

sister > sister	3,142	4 %	74,278	31 %	77,420	24 %
wife > husband	31,502	35 %	51,910	22 %	83,412	25 %
lover > lover			16,165	7 %	16,165	5 %
other	1,995	2 %	931	0 %	2,926	1 %
Total	89,219		239,378		328,597	

Relationship	Men					
	1600–1679		1680–1800		Entire period	
father > son	26,756	10 %	39,560	10 %	66,316	10 %
father > daughter	962	0 %	20,878	5 %	21,840	3 %
son > father	29,090	11 %	94,931	24 %	124,021	19 %
son > mother	25,039	9 %	15,381	4 %	40,420	6 %
brother > brother	63,631	24 %	107,056	27 %	170,687	26 %
brother > sister	13,725	5 %	54,960	14 %	68,685	10 %
husband > wife	101,015	38 %	15,716	4 %	116,731	18 %
lover > lover	2,468	1 %	11,572	3 %	14,040	2 %
other	3,498	1 %	34,825	9 %	38,323	6 %
Total	266,184		394,879		661,063	

All first- and second-person pronouns, apart from the receding and very infrequent *th*-forms, were retrieved from the data sets. In what follows, however, we will mostly concentrate on *I* and *you* as the most frequent forms.

4. Method

Rather than simply computing aggregate pronoun frequencies for each category of interest (e.g. men and women) and comparing them, we divide the corpus into smaller samples, calculate normalised pronoun frequencies for each sample, and compare the categories based on the median frequency of the samples.⁵ For the purposes of this paper, each sample consists of a person's letters from a 20-year time period to recipients with whom the sender has a certain relationship. This kind of sample is specific enough to allow for the analysis of relationship-based variation while keeping the number of samples from each person low enough that a single outlier cannot easily skew the results. Our method is therefore an improvement on Säily et al. (2011), who used individual letters as samples.

As a precaution, we nevertheless exclude three outliers from the analysis. In some of our smaller categories – for instance, the category of men's letters to close friends from 1600–1639, consisting of only 13 samples – a single person may still skew the results to an extent. The excluded outliers are gentleman John Chamberlain (underuser of the most frequent first- and second-person pronouns, he wrote T and TC letters in 1600–1625), merchant Richard Cocks (underuser, T letters 1613–1622) and

⁵ We do this for each first- and second-person pronoun individually, for all of them together and for two groups of first-person pronouns: *I + me + we + us* and the singular *I + me*. Possessives are excluded from the last two groups because they seem to be connected to formulaic language use in letters, thus behaving differently.

gentlewoman Dorothy Osborne (overuser, FN letters 1652–1665), the last of whom was also excluded by Säily et al. (2011). Furthermore, in the closer analysis of family letters, we leave out letters by royalty as they are an atypical representative of the family unit.

To visually compare the categories, we use a modified version of the beanplot (Kampstra 2008; figure 1). Embedded in the middle of each “bean” is a one-dimensional scatterplot, where each small horizontal line represents the normalised frequency of pronouns in a sample (expressed as a percentage out of all words in the sample), while the thick horizontal line shows the median frequency across all samples in the data set. The shape of the bean represents the density of the samples. In comparisons between categories, the bean is divided so that one category (such as women) is shown on the left and the other (such as men) on the right; otherwise the scatterplot and density trace are mirrored to form the bean.⁶ The median frequency of pronouns across all data sets is shown as a dotted horizontal line across the entire graph. We use median rather than average values because they are less susceptible to outliers.

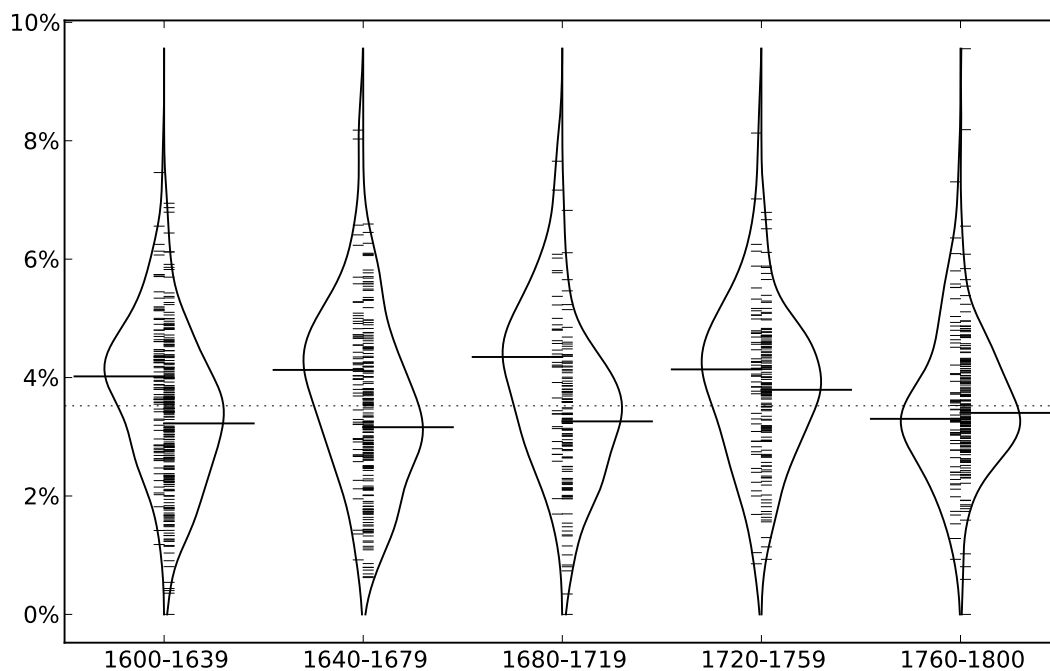


Figure 1. Frequency of *I* in letters written by women (left) and men (right) in the *Corpora of Early English Correspondence*, 1600–1800.

To discover statistically significant differences between categories, we use the Wilcoxon rank-sum test (Wilcoxon 1945, Mann & Whitney 1947). While the p value obtained from a significance test refers to the likelihood of observing a single measurement under the null hypothesis, here we end up with thousands of measurements: comparisons between men and women and different relationship categories for each pronoun and each group of pronouns in each time period. When we test multiple hypotheses, the probability that at least one of them is marked as

⁶ The figure captions in this paper will report the categories in this order, from left to right.

statistically significant is much greater than the probability for each of them individually (see Lijffijt et al. 2012: 4.3). To correct for the large number of hypotheses tested, we therefore apply a statistical method called false discovery rate control (Benjamini & Hochberg 1995). We use a false discovery rate of 0.1, i.e., we accept that 10% of the results may be due to chance, which in these data corresponds to a significance threshold of $p < 0.0055$.

5. Results

As stated in the introduction, we are mainly interested in studying the use of first- and second-person pronouns, that is, the pronouns which have been argued to be indicative of an involved style of writing. As a starting point, let us compare the frequencies of *I* (figure 1) and *you* (figure 2) in men's and women's letters from 1600 to 1800.⁷

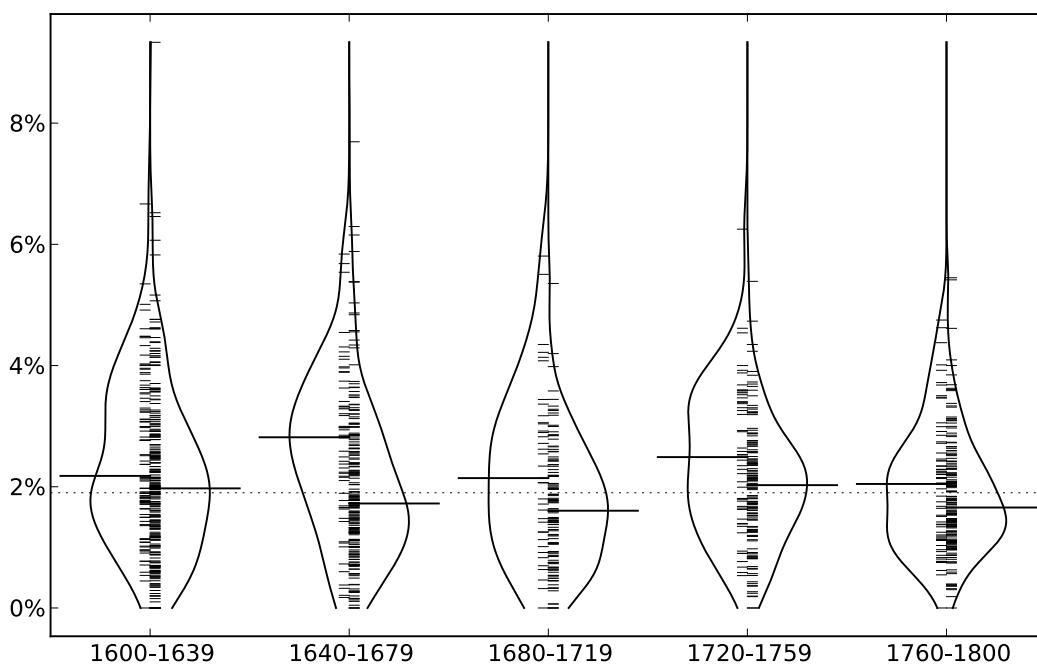


Figure 2. Frequency of *you* in letters written by women and men, 1600–1800.

The frequency of *I* is significantly higher in women's letters in the 17th century. For *you*, the difference is significant for 1640–1679. In the 18th century, however, the differences start to even out, and in the last period studied (1760–1800), there is no significant difference in the use of *I* and *you* across genders. Overall, the frequencies of *I* and *you* do not remain stable: there is a statistically significant increase in both pronouns from 1680–1719 to 1720–1759. The data indicate that the men's use of both *I* and *you* and the women's use of *you* increased during this period. However, the frequency of the pronouns decreases for both genders from 1720–1759 to 1760–1800.

⁷ The frequencies of *you* include both nominative and accusative forms in the singular and the plural, which may explain why we find more statistically significant results with *I*.

The developments in figures 1 and 2 may be due to actual changes in men’s and women’s writing styles. However, considering that the proportion of relationship categories in the corpus changes over time for both men and women (table 1), it may also be that the levelling of pronoun frequencies can be attributed to the way the corpus is structured.

To see if this is indeed the case, let us first take a closer look at the letters written to the members of the nuclear family, such as parents, spouses or siblings (FN), and compare them to the letters sent to people in the “other” category (T) (see section 3 for description).

5.1 Nuclear family vs. “others”

One could hypothesise that the close relationship between the writer and the recipient would result in a more involved style of writing for both genders. In the CEEC, letters sent to the members of the nuclear family (FN) might therefore be assumed to represent a highly involved style and to have a higher frequency of first- and second-person pronouns than letters sent to the people in the “other” category (T).

The data provide some support to this hypothesis. For example, we find that the use of *I* is significantly higher for men in the FN category compared to the men in the T category for 1600–1679. On the other hand, the median frequency of *I* is actually slightly higher in the T category for 1680–1800, although this difference is not statistically significant. In women’s letters, where we have very little data for some periods, the frequency of *I* fluctuates without a clear tendency, and only the data for 1760–1800 yield a statistically significant result (FN > T).

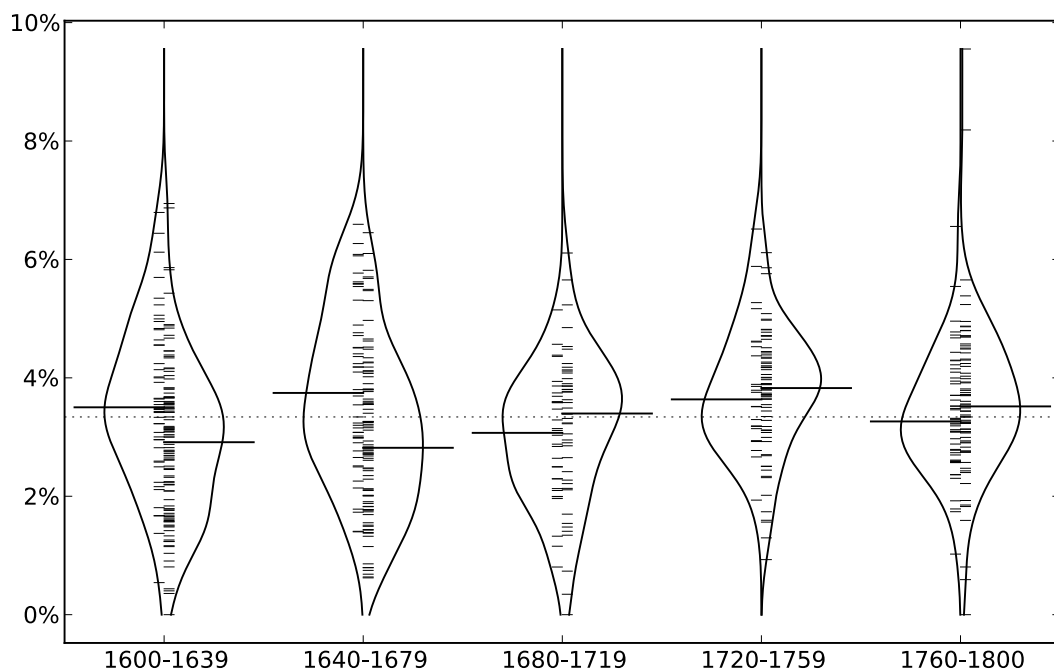


Figure 3. Frequency of *I* in the men’s letters in the FN and T categories, 1600–1800.

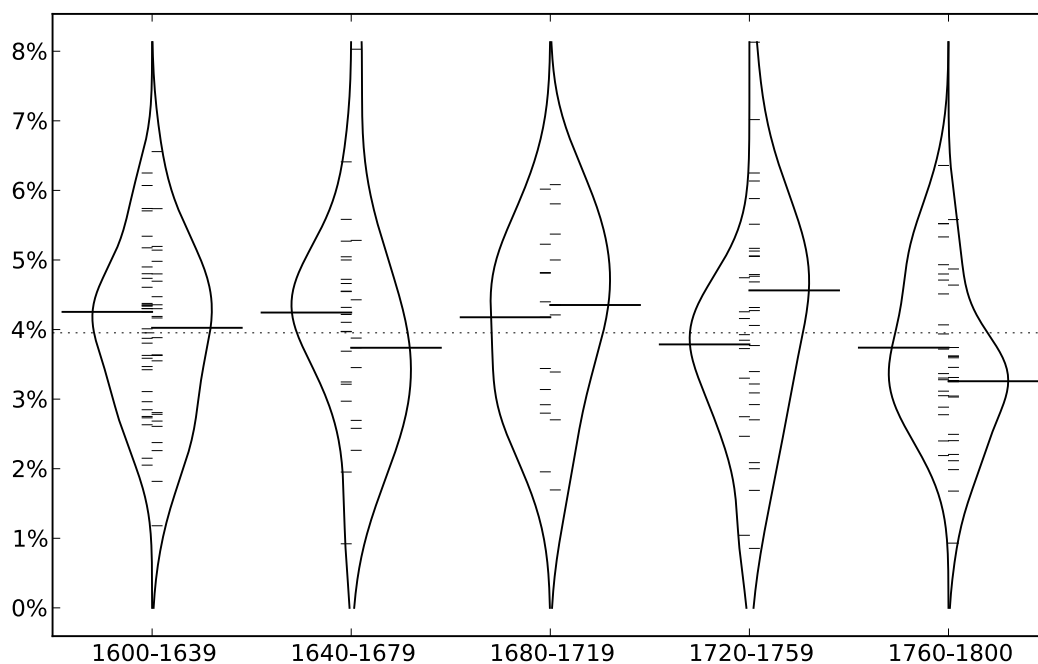


Figure 4. Frequency of *I* in the women's letters in the FN and T categories, 1600–1800.

Figures 3 and 4 suggest that the balance of the relationship categories may indeed affect the tendencies seen in figures 1 and 2. As the women's letters are skewed in favour of the FN category, more evenly balanced relationship categories would level out the gender differences in the 1600–1679 period somewhat. Moreover, men use *I* more often in the T category than in the FN category in the 1680–1800 period. At the same time, however, the proportion of the T letters decreases from 40% to 32%, which mitigates this effect.⁸

Although the relationship categories used in the CEEC, such as FN and T, do have some effect on pronoun frequencies, we argue that they may in fact mask variation on a more detailed level. In the following section we will take a closer look at the letters written between the members of the nuclear family (FN). We will show that the nuclear family is a very heterogeneous category and that the precise family role has a significant effect on pronoun frequencies to the point where it may constitute an even more important factor affecting pronoun usage than gender.

5.2 The nuclear family

Let us start the analysis of the nuclear family letters with cases where both the writer and the recipient are men, thus ruling out gender as an explanatory factor for variation in pronoun usage. In letters between fathers and sons, we see that sons use first-person forms more often than their fathers for the entire period studied (figure 5).⁹

⁸ Two outliers have been removed from these data, which explains the discrepancy between these figures and the ones in table 1.

⁹ Statistically significant differences include the use of *my* in 1600–1679 and *I, me, we* and *us* in 1680–1800. Somewhat surprisingly, Nurmi & Palander-Collin (2008) found

Furthermore, fathers use *you* more often than their sons in 1600–1679, but this result falls just shy of our threshold for statistical significance ($p \approx 0.0060$, figure 6). Similarly to differences between women and men (figures 1 and 2), the differences in pronoun usage in parental letters show a tendency of being evened out in the latter period (figures 5 and 6).



Figure 5. Frequency of *I* in father-son and son-father letters.

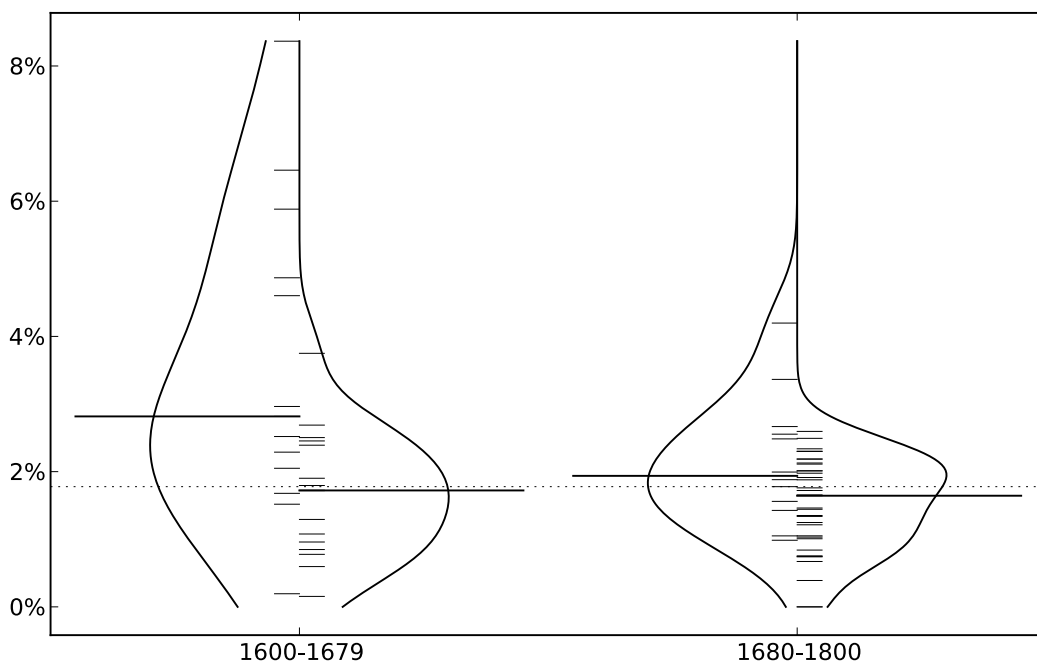


Figure 6. Frequency of *you* in father-son and son-father letters.

that children writing to their parents use less *I*. This may be due to the small size of their sample or the fact that they did not categorize family letters according to gender.

The fact that gender does not (and cannot) affect pronoun usage in figures 5 and 6 poses a challenge to the idea that the frequency of first- and second-person pronouns is directly related to a more or less involved style of writing. First, there is no *a priori* reason to assume that fathers would write in a more involved style than sons or vice versa. Second, the pronouns *I* and *you* have different distributions: sons use *I* more often than their fathers, while the frequency of *you* is higher in the fathers' letters. As far as we know, neither first-person nor second-person pronouns are assigned clear priority in Biber's Dimension 1, which is related to the involved vs. informational production of text.¹⁰ It would therefore seem that the letters between fathers and sons are simply more concerned with the lives and activities of the sons, which shows in the sons' higher frequency of *I* and the fathers' higher frequency of *you*.

To illustrate, example (1) is taken from a letter written by Matthew Hutton to his father, where he recounts his recent activities and future plans. The short excerpt includes eight first-person forms (7 instances of *I* and 1 *my*) and only one token of *you*. By contrast, in (2) the Earl of Rochester writes to his son, giving him fatherly advice and telling him to "observe the instructions of [his] parents". This short letter includes eight first-person forms (6 instances of *I*, 1 *me* and 1 *my*). However, the focus is strongly on the recipient: there are altogether 15 tokens of *you* and 5 tokens of *your*.

(1) **I** was in hope to have gott out of towne this weeke, but **I** fear **I** shall not; for some of the gentlemen with whom **I** am to treat be out of towne, and **I** must stay there returne. **I** have placed **my** brother Thomas with M^f. Farmer: the indentures are not yet sealed, so as **I** cannot send **you** the conditions. (Matthew Hutton to his father, 1628)

(2) **I** hope Charles, when **you** receive this, and know that **I** have sent this gentleman to bee y^f tutour, **you** will bee very gladd to see **I** take such care of **you**, and bee very gratefull, w^{ch} is best showne in being obedient & dilligent, **you** are now grown bigg enough to bee a man, if **you** can bee wise enough; and the way to be truly wise is to serve God, learne y^f booke and observe the instructions of y^f Parents first and next y^f Tutour, to whom **I** have intirely resign'd **you** for this seven yeare, and according as **you** imploy that time, **you** are to bee happy or unhappy for ever; but **I** have soe good an opinion of **you** y^t **I** am glad to thinke **you** will never deceive **me**, deare Child. Learne y^f Booke, & bee obedient, & **you** shall see what a father **I** will bee to **you**. **You** shall want noe pleasure while **you** are good, & that **you** may be soe are **my** Constant Prayers. (The Earl of Rochester to his son, 1670s)

The uneven distribution of first- and second-person pronouns in (1) and (2) illustrates an important fact about the nature of the relationship between fathers and sons in the period studied: fathers could tell their sons what to do and how to behave, which shows in the high frequency of second-person pronouns (cf. Nurmi & Palander-Collin 2008). Furthermore, it may be that the sons were often obliged to report their activities to their fathers, as in (1), which further contributes to the sons' overuse of *I* (figure 5). Therefore, it is conceivable that the differences in pronoun usage are a

¹⁰ The relative importance of first- and second-person pronouns is not always the same. Biber (1988) finds second-person pronouns to be a more important indicator of involvement than first-person pronouns, whereas in Biber (2001) the situation is reversed.

consequence of the unequal relationship between fathers and sons without necessarily reflecting an involved style of writing.¹¹

Letters between husbands and wives, on the other hand, show that husbands use *I* even more often than either fathers or sons in the first period studied (figure 7). Moreover, the frequency of second-person forms is even lower in husbands' letters than in the letters written by sons to their fathers (figure 8).¹²

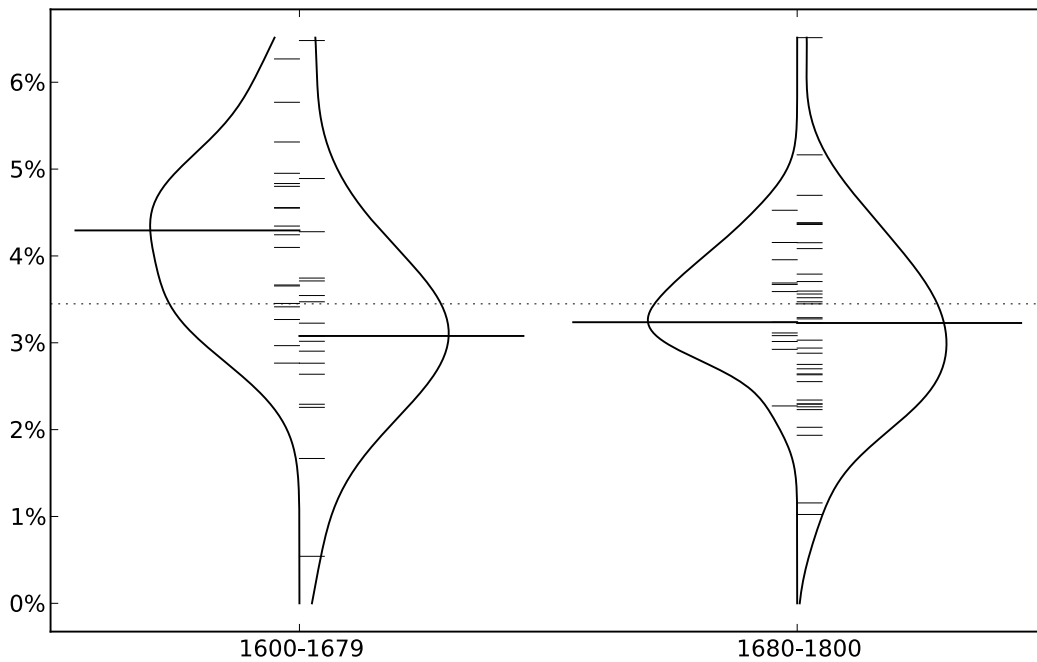


Figure 7. Frequency of *I* in husband-wife and son-father letters.

¹¹ Palander-Collin (2009: 263) shows that the sons' first-person usage may also reflect self-confidence or even arrogance. We will return to the possible reasons underlying pronoun usage in sections 6 and 7.

¹² The difference is statistically significant in the use of *your* in 1680–1800.

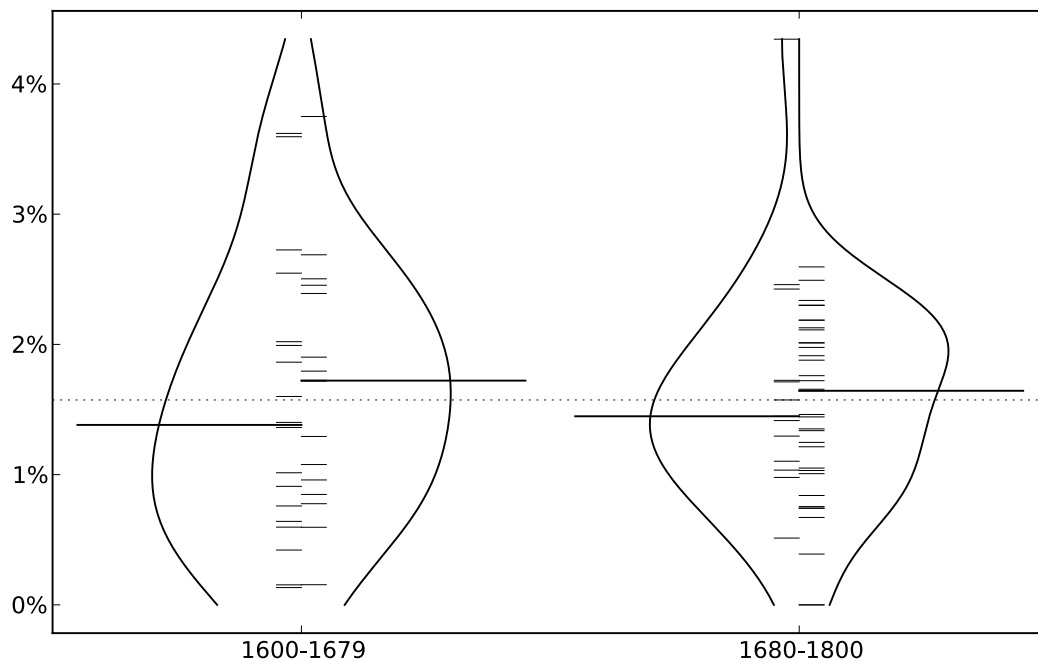


Figure 8. Frequency of *you* in husband-wife and son-father letters.

Again, we could argue that the husbands' letters focus more on what the husbands are doing, which explains the high frequency of first-person forms and the low frequency of second-person forms. Examples (3) and (4) illustrate this.

(3) Sweet hart, **I** wonder that Shipman is so busy, and yet **I** can not hear of any thing he doth, why doth he not wryte? nor any boddy else what he doth. If this 50^{li} will make me a saver, **I** shall reddily accept it, but **I** have no reason to loose by him, seeing his land with a little patience will help me to **my** mony, and theruppon have allreddy given order to James Lane to sew him to the outlary: to understand the bargain better, all my rents must be cast up, for as **I** take it, **I** make nothing of the grounds, arrable, meddow, nor pasture... (John Holles to his wife, 1627)

(4) **I** had a long discourse this morning in the Lo^s house, with Judg' Re:, who advis'd **me** not to be to hasty but see a little. **I** told him **I** had no great reason to exspect any Quick' disspatch', by the rule of other^s attendance. **I** am to wayte vpon m^r Herne, the councellor, to morrowe morning, who hath promisd **me** his best advise. He is well verst in the busines of sequestrations. **I** will throwe away a fee vpon him. He may perhapps gaine **me** an ordre to stopp sale of my goods 'till the case be heard. **I** mett this weeke with **my** olde freind m^r Legat, who seem'd to **me** the same man that ever he was. If his hart be otherwise, 'tis his sinn, not **mine**. After some discourse & many professions **I** ingag'd him both to Beadle & Cap^t Warner, who hath promist **me** to vse the vttermost of his power with them both, for **yo^r** Quiet & the saf'ty of **my** goods. (Thomas Knyvett to his wife, 1644)

Both (3) and (4) are information-oriented: they focus on the businesses of the husband. In (3), the only reference to the recipient of the letter is "Sweet hart" at the very beginning, while the excerpt in (4) includes 15 first-person forms (*I*, *me*, *my*, *mine*) and only one second-person possessive *yo^r* at the end of the passage.

Our data allow us to formulate an implicational scale of men's first- and second-person pronoun usage within the FN category in 1600–1679:

First person:	husbands	>	sons	>	fathers
Second person:	husbands	<	sons	<	fathers

Figure 9. An implicational scale of pronoun frequencies in men's FN letters, 1600–1679.

The scale in figure 9 suggests that men's first- and second-person pronoun usage cannot be neatly explained by a more/less involved style of writing. If it could, we would expect to see consistent overuse of both first- and second person forms in one of the categories, but instead we see a more elaborate tendency where the husbands overuse *I* and the fathers *you*, and the sons are ranked in the middle in both cases (cf. ego involvement vs. involvement with the hearer in Chafe 1985: 116–118).

5.3 Gender differences within the nuclear family

In the previous section we showed that the exact relationship between the writer and the recipient is relevant for pronoun usage. However, the above discussion is by no means intended as a rebuttal of earlier research, where clear gender differences in pronoun usage have been observed (Rayson et al. 1997, Argamon et al. 2003, Säily et al. 2011). For example, our data from 1600–1679 show that mothers writing to their sons use first-person forms significantly more often than fathers (figure 10). Likewise, in spousal letters from the same period wives use *you* more often than their husbands (figure 11).

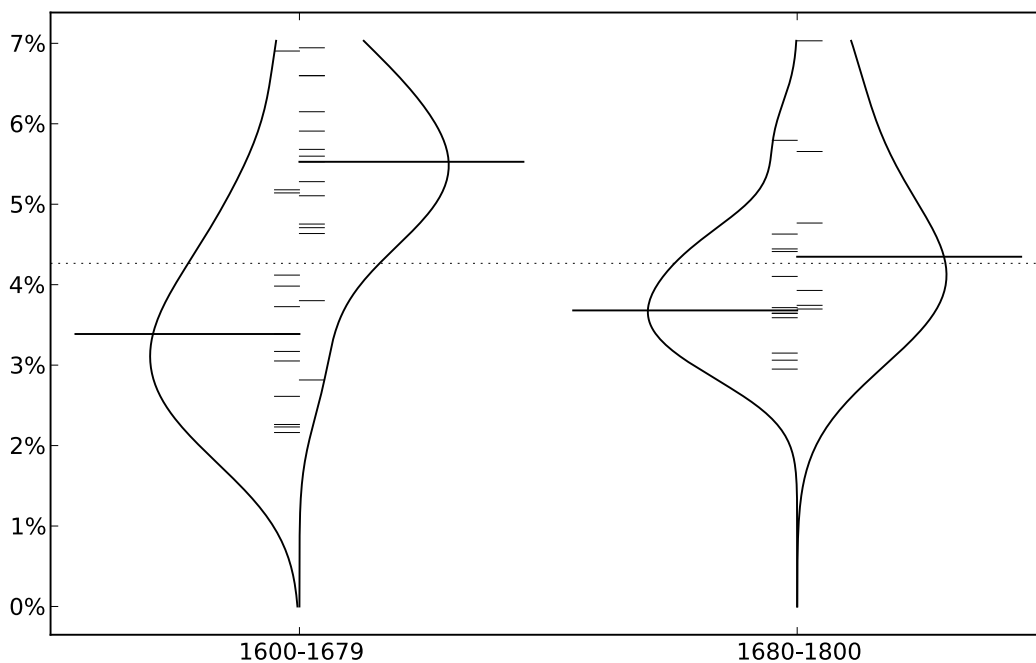


Figure 10. Frequency of *I*, *me*, *we* and *us* in father-son and mother-son letters.



Figure 11. Frequency of *you* in husband-wife and wife-husband letters.

Compared to example (2), where the Earl of Rochester gave very direct instructions to his son, example (5) illustrates the more indirect way in which Lady Brilliana Harley asks her son to buy some items for her. The letter includes assurances of monetary compensation with frequent pronoun use (*I... will send mony to pay for them; ... which I will pay you for; I doo willingly giue you the rige of goold*), but there are also epistemic markers (and conditional *if*-clauses) that in part soften the potential face-threatening act of ordering the son to do something for the mother (e.g. *I beleuee you remember; I beleuee it may be mended*, see also Palander-Collin 1999).

(5) All **my** frute disches are brocken; thearefore, good Ned, if theare be any shuch blwe and white disches as **I** vse to haue for frute, bye **me** some; they are not purslane, nor they are not of the ordinary mettell of blwe and white disches. **I** beleuee **you** remember what **I** vse to haue; if **you** chuse them against the horses come for **you**, **I** will take order with the men about the bringeing of them home, and will send mony to pay for them. **I** see **your** sister has a nwe hude; it semes shee lost hers and durst not tell, and so, as **I** gees, rwit to **you** for one, which **I** will pay **you** for. **I** haue sent **you my** wath, and **I** beleuee it may be mended. **I** doo willingly giue **you** the rige of goold that was aboute the agget. (Lady Brilliana Harley to her son, 1639)

Example (6), on the other hand, illustrates that the high frequency of *I* may also be explained by other factors. First, the mother uses *I* to report her reactions to the news regarding her son (*I was mighty glad to receive your letter; I was afraid of your taking physick at Venice*). Second, the mother uses *I pray* to issue a polite request, and third, the mother uses *I hope* to wish her son good health and to express her own desires (*I hope in God that you will now recover your health, I hope to bee shortly at Kensington*, see Palander-Collin 2009 for data on frequent *I*-clusters).

(6) **I** was mighty glad to receive **your** letter from Florence, because **I** was afraid of **your** taking physick at Venice, but now **you** are come to **your** uncle and Dr Baines

my mind is at rest, for **I** am sure **you** will want nothing that they can help **you** to, and **I** hope in God that **you** will now recover **your** health for **I** hear **you** are very melancholy and that makes **mee** beleive **you** are not well. **I** pray remember **mee** to **your** uncle and Dr Banes with great affection, and bee sure to watch **yourselfe** that **you** do not stoop, for it is neither handsome nor wholesome. **I** hope to bee shortly at Kensington, for the sickness abates very much; there dyed last week but 652 of the plague, and this week but 300; but the worst is that it is still dispersed in every parish a little, and God knows what it may come to agayn next summer. (Lady Elizabeth Finch to her son, 1665)

In the letters written by wives to their husbands, we find a high frequency of both *I* and *you* (cf. Palander-Collin 2000). In the 18th century data women use *I* significantly more than men ($p < 0.003$). Similarly to the husbands' letters, the letters written by wives often included news of what had taken place at home, which contributes to the high frequency of *I*. However, the wives also focused on the activities and opinions of the husbands, as in (7), where Catherine Clavering writes to her husband James, defending herself against her husband's accusations of "reprimanding" him. The letter includes 30 first-person forms and 19 second-person forms in total.

(7) **My** dearist,

If that **I** did not desire to aquit **my self** of what **you** continually charge **me** with, **I** would not give **you** the trouble of **my** epistle this post, as sister Alice writes. **Your** daily accuseing **me** is a concerne that **I** can not lay any thing of but **you** call it reprimanding **you**, which **I** know nothing of, nor **I** hope ever shall. It's fite **you** should be **your** own master and take **your** liberty for **me** and from henceforth **I** will never name **your** writeing in any kind, since it is such a crime. It is **my** parte to write as oblidgingly to **you** as **I** am capable, and **I** think have don it to the best of **my** poor ability. Indeed **my**n is not in so good a still as **I** could wish but **I** hope they may be accepted as comeing from an affectionat senceare harte as the more refined. So **I** beg it of **you** may have no more angry words and **I** promis shall have no reason **I** can prevent. For the latter parte of **your** last letter, **I** wounder why **you** should imagine it to be a pleaser to **me**, since **I** never gave **you** reason to accuse **me** of imodisty of all faults. If **you** had seen **me** read it **you** might have observed a greate maney blushis, which it provoked exstreamly, without any greate satisfaction. **I** shall now reliefe **you** from **my** nonsense, believing **me**...

P.S. Cosen Pell is come tonight for news. Refers **you** to sister Alice, who is better at giveing **you** it than **myself**. (Catherine Clavering to her husband, 1705)

6. Discussion

The results of this study complement earlier research by Säily et al. (2011), who found that women use more pronouns than men in early English letters. In this paper we have extended the scope of investigation to 1800, focusing on first- and second-person pronouns as well as the relationship between the writer and the recipient. Similarly to Säily et al. (2011), our data revealed clear gender differences in the 17th century letters. However, we also have some concerns that may present problems to the way the results were interpreted in Säily et al. (2011).

First, our data support the results of the earlier study only in the first period studied, i.e. 1600–1679, where we found statistically significant and relatively consistent gender differences in pronoun usage. However, the data from 1680 to 1800 indicate significant levelling: in 1760–1800, the differences in the frequencies of *I* and *you* are much smaller across genders. We pointed out that our results may in part be due to imbalances in the corpus: neither the relationship categories nor the precise relationships between the writer and the recipient within the relationship categories are balanced for gender, nor do they remain stable across time. For example, the proportion of son-father letters in men's FN category increases from 11% to 24% from 1600–1679 to 1680–1800, while the proportion of husband-wife letters decreases from 38% to 4%.

Furthermore, we have argued that the relationship categories in the CEEC may be too general to provide revealing results and that linguistic variation (at least in the case of pronoun usage) should also be studied on a more detailed level (cf. Biber & Gray 2013), so that the precise relationship between the writer and the recipient, such as son-father or wife-husband, is taken into account. While the CEEC is not balanced for these categories, it does include all the necessary information for studying the data at the appropriate level of granularity.

It should also be emphasised that despite our concerns regarding the structure of the corpus, we do not believe that the gender differences observed in the 17th century data or the developments in the 18th century can be explained by problems in the balance of the corpus alone. It is plausible that societal factors, such as increased literacy and better education, contributed to the loss of gendered styles in the 18th century. Be that as it may, we think that the extent to which the differences in the 17th century letters can be attributed to gendered styles should be studied more closely: our data suggest that the recipient of the letter could have an even more significant effect on pronoun usage than the gender of the writer. Moreover, in some cases, such as in the letters written by wives to their husbands, the high frequency of second-person forms seems to reflect the fact that the topics of the letters often revolved around the lives of the husbands (which is mirrored in the high frequency of *I* in the husbands' letters in the 17th century data).

As a final note, we would like to emphasise that personal pronoun use is a multifaceted phenomenon. In addition to gender, it is influenced by a number of factors, such as politeness, the relative (in)equality between the writer and the recipient, the topic of the letter and the way social identities are constructed and portrayed (see also Säily et al. 2011: 182, Nurmi & Palander-Collin 2008: 44). For example, while our data indicate that letters written by husbands were often more information-oriented than letters written by wives, it could be argued that this is something especially typical of the husband-wife relationship in the 17th century (i.e. the specific social roles of the men and the women), not just a reflection of a general difference between women's and men's language use.

7. Conclusion

In this paper we have studied the relevance of the recipient and corpus balance to pronoun usage in Early and Late Modern English personal letters. We have specifically addressed the question of gendered styles and examined the explanation

for gender variation in Säily et al. (2011), where it was suggested that women's more frequent use of personal pronouns is indicative of a more involved style of writing.

We found significant gender differences in our data, but our results also pose some challenges to this interpretation. Our data showed that i) the differences in pronoun frequencies are not stable across time, ii) the results are inconclusive in some categories, and iii) there are statistically significant intra-gender differences depending on the relationship between the writer and the recipient (e.g. the use of *I* for husbands vs. fathers).

We also found that the balance of the corpus may affect the results to some degree and emphasised that close attention should be paid to the nature of the relationship between the writer and the recipient in future work. Finally, we fully acknowledge that further studies are needed for a more comprehensive picture of all the relevant factors underlying pronoun usage.

Acknowledgements

The authors would like to thank Jukka Suomela for methodological assistance and for coding the Python scripts used for significance testing and producing the graphs in this paper. We also thank an anonymous reviewer for helpful feedback.

References

- Argamon, Shlomo, Moshe Koppel, Jonathan Fine and Anat Rachel Shimoni 2003. Gender, genre, and writing style in formal written texts. *Text* 23(3): 321–346.
- Bell, Allan 1984. Language style as audience design. *Language in Society* 13: 145–204.
- Benjamini, Yoav and Yosef Hochberg 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57(1): 289–300.
- Biber, Douglas 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.
- Biber, Douglas 2001. Dimensions of variation among 18th-century speech-based and written registers. Hans-Jürgen Diller and Manfred Görlach (eds.), *Towards a History of English as a History of Genres*, 89–109. Heidelberg: Winter.
- Biber, Douglas and 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 and Edward Finegan 1997. Diachronic relations among speech-based and written registers in English. Terttu Nevalainen and Leena Kahlas-Tarkka (eds.), *To Explain the Present: Studies in the Changing English Language in Honour of Matti Rissanen*, 253–275. Helsinki: Société Néophilologique.
- Biber, Douglas and Bethany Gray 2013. Being specific about historical change: The influence of sub-register. *Journal of English Linguistics* 41(2): 104–134.
- CEEC = *Corpora of Early English Correspondence*. Compiled by Terttu Nevalainen, Helena Raumolin-Brunberg et al. at the Department of Modern Languages, University of Helsinki. <http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/>
- Chafe, Wallace L 1985. Linguistic differences produced by differences between speaking and writing. David R. Olson, Nancy Torrance and Angela Hildyard

- (eds.), *Literacy, Language, and Learning: The Nature and Consequences of Reading and Writing*, 105–123. Cambridge: Cambridge University Press.
- Fitzmaurice, Susan 2012. Social factors and language change in eighteenth-century England: The case of multiple negation. *Neophilologische Mitteilungen* 113(3): 293–321.
- Hay, Douglas and Nicholas Rogers 1997. *Eighteenth-Century English Society: Shuttles and Swords*. Oxford: Oxford University Press.
- Kampstra, Peter 2008. Beanplot: A boxplot alternative for visual comparison of distributions. *Journal of Statistical Software* 28: Code Snippet 1.
- Lijffijt, Jeffrey, Tanja Säily and Terttu Nevalainen 2012. CEECing the baseline: Lexical stability and significant change in a historical corpus. Jukka Tyrkkö, Matti Kilpiö, Terttu Nevalainen and Matti Rissanen (eds.), *Outposts of Historical Corpus Linguistics: From the Helsinki Corpus to a Proliferation of Resources*. Studies in Variation, Contacts and Change in English 10. Helsinki: VARIENG.
http://www.helsinki.fi/varieng/journal/volumes/10/lijffijt_saily_nevalainen/
- Mann, H. B. and D. R. Whitney 1947. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* 18(1): 50–60.
- McIntosh, Carey 2008 British English in the long eighteenth century (1660–1830). Haruko Momma and Michael Matto (eds.), *A Companion to the History of the English Language*, 228–234. Chichester: Wiley-Blackwell.
- Nevala, Minna 2004. *Address in Early English Correspondence: Its Forms and Socio-Pragmatic Functions*. Mémoires de la Société Néophilologique de Helsinki LXIV. Helsinki: Société Néophilologique.
- Nevalainen, Terttu and Heli Tissari 2010. Contextualising eighteenth-century politeness: Social distinction and metaphorical levelling. Raymond Hickey (ed.), *Eighteenth-Century English: Ideology and Change*, 133–158. Studies in English Language. Cambridge: Cambridge University Press.
- Nurmi, Arja and Minna Palander-Collin 2008. Letters as a text type: Interaction in writing. Marina Dossena and Ingrid Tieken-Boon van Ostade (eds.), *Studies in Late Modern English Correspondence: Methodology and Data*, 21–49. Linguistic Insights 76. Bern: Peter Lang.
- Palander-Collin, Minna 1999. Male and female styles in seventeenth-century correspondence: I THINK. *Language Variation and Change* 11: 123–141.
- Palander-Collin, Minna 2000. The language of husbands and wives in seventeenth-century correspondence. Christian Mair and Marianne Hundt (eds.), *Corpus Linguistics and Linguistic Theory: Papers from the Twentieth International Conference on English Language Research on Computerized Corpora (ICAME 20), Freiburg im Breisgau 1999*, 289–300. Language and Computers: Studies in Practical Linguistics 33. Amsterdam: Rodopi.
- Palander-Collin, Minna 2009. Variation and change in patterns of self-reference in early English correspondence. *Journal of Historical Pragmatics* 10(2): 260–285. (Thematic issue: Historical sociopragmatics, ed. Jonathan Culpeper).
- PCEEC = *Parsed Corpus of Early English Correspondence* (tagged version) 2006. Annotated by Arja Nurmi, Ann Taylor, Anthony Warner, Susan Pintzuk and Terttu Nevalainen. Compiled by the CEEC Project Team. York: University of York and Helsinki: University of Helsinki. Distributed through the Oxford Text Archive.

- Rayson, Paul, Geoffrey Leech and 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.
- Säily, Tanja Forthcoming. Sociolinguistic variation in morphological productivity in 18th-century English. To appear in *Corpus Linguistics and Linguistic Theory*.
- Säily, Tanja, Terttu Nevalainen and Harri Siirtola 2011. Variation in noun and pronoun frequencies in a sociohistorical corpus of English. *Literary and Linguistic Computing* 26(2): 167–188.
- Tieken-Boon van Ostade, Ingrid 2010. Eighteenth-century women and their norms of correctness. Raymond Hickey (ed.), *Eighteenth-Century English: Ideology and Change*, 59–72. Studies in English Language. Cambridge: Cambridge University Press.
- Walker, Terry 2007. *Thou and You in Early Modern English Dialogues: Trials, Depositions, and Drama Comedy*. Pragmatics & Beyond New Series 158. Amsterdam: Benjamins.
- Wilcoxon, Frank 1945. Individual comparisons by ranking methods. *Biometrics Bulletin* 1(6): 80–83.