From *Plenipotentiary* to *Puddingless*: Users and Uses of New Words in Early English Letters*

Tanja Säily[0000−0003−4407−8929], Eetu Mäkelä[0000−0002−8366−8414], and Mika Hämäläinen[0000−0001−9315−1278]

University of Helsinki, Finland
{tanja.saily,eetu.makela,mika.hamalainen}@helsinki.fi

Abstract. We study neologism use in two samples of early English correspondence, from 1640–1660 and 1760–1780. Of especial interest are the early adopters of new vocabulary, the social groups they represent, and the types and functions of their neologisms. We describe our computer-assisted approach and note the difficulties associated with massive variation in the corpus. Our findings include that while male letter-writers tend to use neologisms more frequently than women, the eighteenth century seems to have provided more opportunities for women and the lower ranks to participate in neologism use as well. In both samples, neologisms most frequently occur in letters written between close friends, which could be due to this less stable relationship triggering more creative language use. In the seventeenth-century sample, we observe the influence of the English Civil War, while the eighteenth-century sample appears to reflect the changing functions of letter-writing, as correspondence is increasingly being used as a tool for building and maintaining social relationships in addition to exchanging information.

Keywords: neologisms · Early Modern English · Late Modern English · historical sociolinguistics · digital humanities.

1 Introduction

In this paper, we report on our work to study novel vocabulary, or neologisms, in a corpus of historical letters, namely the *Corpora of Early English Correspondence* (CEEC) [25]. The CEEC was specifically designed for the purposes of historical sociolinguistics, and as such it 1) aims at social representativeness in terms of e.g. gender and social rank and 2) records this social metadata for each letter included. Previous research into the history of English lexis has focused on published texts, which were mostly written by highly educated men. By analysing personal correspondence, a genre that was accessible to anyone who was literate, we are able to consider the language use of women and the lower social ranks as well. In this study, we are especially interested in the early adopters of new vocabulary, the social groups they represent, and the types and functions of their neologisms.

However, the varied nature of the corpus also poses problems for neologism identification. The CEEC compilers wanted to study authentic language use, and thus the

corpus is compiled from editions that retain original spelling. English spelling did not develop a prescribed standard until the eighteenth century, and there is a great deal of spelling variation in historical English texts – and in the CEEC. In total, the CEEC contains about 150,000 distinct word forms. By excluding those tagged as foreign, proper nouns or abbreviations, about 125,000 remain. Our initial idea was to take this vocabulary, and match it against the Oxford English Dictionary (OED) [27] and the Middle English Dictionary (MED) [23] to discover which words appeared in the CEEC earlier than mentioned in the dictionaries. However, even though the OED and MED contain some 375,000 and 260,000 spelling variants for their 280,000 and 60,000 words, only some 36,000 word forms from the CEEC could be directly mapped to the dictionaries, while 88,000, or 70%, could not. Further, the proportions of words that could be mapped differed widely between social ranks, with words used by the lower ranks having a significantly lower chance of finding a match in the dictionaries. This in turn made straightforward statistical comparisons between social ranks impossible.[1] To try to counteract these problems, we turned to automatic normalization approaches, which we hoped would allow us to map a significantly larger portion of the words against the two dictionaries.

2 Related Work

There is a lot of recent NLP research conducted on historical data, especially in meaning change over long periods of time. On the other hand, there are also recent studies on neologisms in present-day data. In this section, we describe some of the recent related work.

In linguistics, previous research on historical neologisms has typically gathered its data from dictionaries, which are biased towards well-known authors, while corpus-based studies have chiefly concentrated on individual derivational affixes [24,28]. Present-day studies have used more automated methods to discover a potentially wider range of neologisms in massive corpora compiled from e.g. newspapers or tweets [32,12], but they have tended to ignore social categories beyond regional variation.

A recent study shows a method for normalizing neologisms in social media texts [42]. Their method tries to first detect a neologism and then normalize it into standard language. Interestingly, they apply the idea of normalization in order to remove neologisms, where as we use normalization to find neologisms [15].

Another neologism identification study focuses on multi-word neologisms, namely adjective-noun pairs [22]. Their approach uses BERT [5] and ELMo [30] in detecting whether a given adjective-noun pair is a neologism. Our work focuses on individual words that are neologisms as opposed to multi-word expressions.

Spanish neologisms have been studied in a corpus of conversations of language learners [40]. They study neologisms in a neologism-annotated corpus examining the relationships between neologisms, loanwords, and errors. Loanwords and errors are typically produced by language learners and true neologisms by native speakers. Unlike our corpus, their corpus is annotated for neologisms.

[1] For more on the problematic nature of such non-standard data for analysis, see [21].

In Hämäläinen, M., Partanen, N., Alnajjar, K. (eds.) Multilingual Facilitation (2021), pages 153–169.

Lexical semantic change has been studied with word embeddings [8]. The authors suggest that word embeddings should not be aligned when used to study semantic change, as such an alignment introduces noise. Their approach introduces less noise which ultimately leads to more accurate results. Despite this, the use of machine learning approaches to study the phenomenon is not free of problems [17].

Normalization is a very common practice embraced in the NLP research when dealing with historical or otherwise non-standard language [7,3,29,9,43]. The benefit of normalization is that it makes non-standard orthography standard and thus enables the use of NLP tools and resources designed for modern normative data.

3 Neologism Pipeline

The details of our automated mapping procedure are described in [14]. In short, the best-performing approach we could come up with was based on neural machine translation, with a post-filtering step that accepted only lemmas appearing in the OED [15]. Access to a local version of the OED was kindly provided by Oxford University Press. Overall, this approach was able to recover accurate lemmas for 61% of previously unmatched words.

However, a broader evaluation of the whole pipeline cast a dark shadow over the overall usability of the method. Out of all words in a sample of 17th-century words, 75% were matched to a lemma. However, for a full 30% out of those, that lemma turned out to be a wrong one! Finally, even out of the words that were matched to the right lemma, 17% were matched to the wrong part of speech, which, when compared to the OED, could equate to a possibly wrong earliest attestation date. Actually, for this data, even if the lemmatization were perfect, about 20% of the words would still get mapped to the wrong entry without further disambiguation based on part of speech or other contextual information. On top of this, the normalization accuracy was not the same for each social category. While, interestingly, lemmatization accuracy was similar across social ranks, it differed according to gender, as well as according to the type of correspondence (between family members, between friends or between more distant acquaintances).

In light of this, while we originally endeavoured to come up with a mostly automated pipeline that could be used to study neologisms en masse and quantitatively, in the end we were forced to conclude that this would not be possible. Instead, we resigned ourselves to treat the automatically derived lemmas and OED matches as only suggestions that would be verified manually. This in turn meant a significant amount of manual work for analysing any sizable part of the corpus. We therefore adopted a time-limited sampling-based approach, where we first decided on a twenty-year timespan of the CEEC where we had a relatively even coverage of different social categories, and then sampled neologism candidates from that period using a stratified sampling approach that tried to include as equal an amount of running words from each social category as possible.

More specifically, we decided that we wanted our corpora to be as balanced as possible on three social axes: 1) the gender of the letter writer, 2) their social rank, and 3) register in terms of the relationship between the writer and the recipient. Based on an evaluation of the corpus balance over time along these axes, we settled on two time pe-

In Hämäläinen, M., Partanen, N., Alnajjar, K. (eds.) Multilingual Facilitation (2021), pages 153–169.

riods: 1640–1660 and 1760–1780. The former was also of interest because it coincided with the English Civil War. We then grouped the data from those periods according to the three criteria, and sampled letters as equally as possible from all buckets. From each sampled letter were then extracted the words that appeared for the first time in the CEEC in that letter. These then formed the candidates for neologisms without any regard yet to the dictionaries. The sizes of the samples were controlled so that each could be gone through manually in a reasonable time, leading in the end to candidate neologisms lists of 820 word types for the 17th century and 645 for the 18th century (both about 9% of the the total sampling pool of neologism candidates, which was 8,954 for the 17th century and 7,131 for the 18th century).[2]

Table 1. Number of running words in the two samples for each of the three social axes of interest.

Category	Value	17th c.	18th c.
Total		36,265	47,864
Gender	Male	23,459	29,225
	Female	12,806	18,639
Social Rank	Royalty	3,899	4,067
	Nobility	5,038	6,998
	Gentry	11,509	10,924
	Clergy	9,659	8,976
	Professionals	3,675	10,847
	Merchants	860	2,496
	Other non-gentry	1,625	3,556
Relationship	Nuclear family	15,045	12,754
	Other family	0	6,534
	Close friends	7,467	14,771
	Other acquaintances	13,753	13,805

Given the limitations inherent in the corpus, this did not result in a completely balanced sample, as shown in Table 1, but at least its biases are known and can be taken into account in the analysis.

We then used our FiCa tool [35] to manually verify each lemma and OED association suggestion from our automated pipeline, as well as to fill in these for the words where our approach had not yielded candidate mappings. After receiving this corrected and amended mapping, we finally used it to source earliest attestation dates from the OED for each of the lemmas. For this study, we decided to define a neologism as a word whose earliest attestation date in the OED was at most forty years before its appearance in our sample. After a final round of manual verification, this procedure yielded us 42 novel word types in the 17th-century sample, with a total of 53 token-level appearances.

[2] We would like to thank Jukke Kaaronen for going through the 18th-century sample, which was then analysed further by us.

For the 18th-century sample, we obtained 21 novel word types, each of which appeared only once.

For analysis, these neologism tokens were then combined with the social metadata associated with the letters they appeared in, but also etymological and semantic data as sourced from the OED. Here, we looked at each sense entry for each word entry and each semantic node in the Historical Thesaurus of the OED for each sense. The semantic nodes give us important information about the ontological meaning of each neologism candidate, as the Historical Thesaurus (HT) records hierarchical information about the meaning of each entry such as 'society » communication » writing » handwriting or style of'. This makes it possible for us to inspect the neologisms on different levels of the hierarchy.

As the number of neologisms in our samples is so low, the data does not lend itself to statistical hypothesis testing. Thus, all numerical results from the following analysis should be taken as provisional.[3]

4 Seventeenth-Century Neologisms

4.1 Overview

The 17th-century sample provided us with 53 neologism tokens representing 42 types, listed below. Antedatings to the OED (entries and senses, checked against OED Online in February 2021) are marked in boldface on the list.

> acrimonious, believingly, candid, candour, causally, compensate, compliance, condescension, coney ground, congregational, **covenanting** (adj.), **crawling** (n.), dishearten, dragooner, **efficaciously**, eminently, endeared, entanglement, **helpfulness**, **hint** (v.), idolum, incendiary, **incognito**, initiatory, **joke** (n.), **landgravine**, **malignancy**, manifesto, **oversweetness**, **packet-boat**, **plenipotentiary** (n.), remind, rickets, sequestrator, servient (n.), **statement**, Swede, **tea**, variously, **vibrate** (v.), visit (n.), voluminous

Unsurprisingly, nouns being the largest lexical category in general, most of the neologisms (24) are nouns as well, followed by adjectives (8), verbs (5) and adverbs (5). While half of the words have been formed within English through derivation (18), compounding (2) or conversion (1), there are also a large number of borrowings (19): thirteen from Latin and two each from French, Italian and German. The OED gives no certain etymology for two nouns, *joke* and *rickets*, although the former may come from Latin. Looking at the top level of the HT classification and the main semantic class of each word, 16 of the neologism types relate to 'society', 15 to 'the world' and 11 to 'the mind'. Within these classes, the words are distributed across as many as 22 second-level categories, the most frequent ones being 'society » authority' (5 neologism types) and 'the mind » mental capacity' (4 types). The former may be at least partly connected to the Civil War, as in example (1), where *malignancy* seems to be used in the sense 'political disaffection'. Sir Anthony was a Royalist who was charged with misapprobation of public monies and imprisoned during the war.

[3] The analysis dataset is available on Zenodo [36].

(1) For what I made over to Mr. Kenrick's it was uppon reall considerations such as will appeare good if lawe have any being; but of that I will not dispute, considering my durance and being cloathed by some particular men with the garment of **Malignencye** and therfore in a suffring condition.

(Sir Anthony Percival to his neighbour, Henry Oxinden, 1643; OXINDE_186)[4]

Words used in the context of the Civil War can be found in a number of different semantic classes, however. An obvious one is *dragooner* 'society » armed hostility » warrior' used by Charles I himself, but there are also *packet-boat* 'society » travel » travel by water', *sequestrator* 'society » law » administration of justice' and *statement* 'the mind » language » statement', to name a few. Most of them do seem to be found under the top-level class of 'society', which makes sense as the Civil War was chiefly a societal phenomenon.

4.2 Sociolinguistic Variation

In what follows, we will compare the normalized frequencies of neologism tokens across different social groups; see Table 2. These are used as a starting point for a more qualitative analysis of the users and uses of neologisms.

Table 2. Normalized frequency of neologism tokens per 10,000 words in the 17th-century sample for the three main social axes of interest, sorted by frequency.

Category	Value	Frequency / 10,000 words
Gender	Male	17
	Female	11
Social Rank	Royalty	26
	Professionals	24
	Nobility	16
	Gentry	13
	Clergy	11
	Merchants	0
	Other non-gentry	0
Relationship	Close friends	25
	Other acquaintances	18
	Nuclear family	6
	Other family	–

[4] Examples are given in their original spelling. Boldface has been added, while italics (if any) are as printed in the letter edition and probably reflect underlining in the original manuscript. OXINDE_186 is the unique identifier of the letter in the corpus.

The two **social ranks** with the highest frequency of neologisms are royalty and professionals, both with c. 25 instances per 10,000 words. Although Charles I uses three neologism tokens (*dragooner* once and *dishearten* twice), most of the royalty's use of new words is due to Elizabeth Stuart, Queen of Bohemia, who lived abroad, basically in exile, at the time. She wrote about *visits* and being *incognito*, and discussed people ranging from *servients* to *landgravines*, *plenipotentiaries* and *Swedes*, the latter of whom were actively engaged in the Thirty Years' War. Amongst the nine professionals in the sample, new vocabulary is only used by two, Parliamentarian army officers John Dixwell (*sequestrator*) and Thomas Harrison (*believingly*, *condescension*, *endeared*, *hint* v., *variously*). Many estates owned by Royalists were sequestered during the Civil War, and war and religion were mixed in Harrison's discussion of Cromwell with Colonel John Jones (2).

(2) To agree (as is alreadie) to act in dearest love expressed to him named Protector, (or Mount Sirion as the Sidonians called Hermon, and David in the spirit followed that faithfully, **believingly**, undoubtingly, unanimously, that He would retreat in action of undertaking (and soe witnes repentance by **condisention**) and wee would as willingly repent of o^{r} sinfull dissentions) I shall therefore apply what I have now brought to offer, onely to that.

(Thomas Harrison to John Jones, 1656; JONES_040)

As regards **gender**, men tend to use more neologisms than women. Interestingly, the women in the sample mostly use borrowings, whereas the majority of the men's neologism tokens were formed within English. The female early adopters were in general well educated, which could explain their use of the borrowings. They belonged to the social ranks of royalty (Elizabeth Stuart), nobility (Anne Conway), gentry (Brilliana Harley), and clergy (Anne Cary and Winefrid Thimelby, nuns). Lady Harley used the potentially Civil War related term *incendiaries* in a political discussion with her son, who later became a Parliamentarian army officer; both her and Elizabeth's use of new vocabulary focuses on the semantic class 'society » authority'. Lady Conway, on the other hand, discussed philosophical concepts with her friend and fellow philosopher, Henry More, as in (3). The OED records More's use of *idolum* in a publication in 1647, so both of the correspondents would have been familiar with the term.

(3) Now the papyr certainly could not possible be capable of perceiving motion nor could it transmitt its motion from the obiect to the eye, for first the paper transmitts the colour of white w^{ch} is its own motion, and if it should transmitt the motion caused by any other obiect, then why does not everything we Looke upon yeeld the **eidolum** or representation of something else […]

(Lady Anne Conway to Henry More, 1651; CONWAY_093)

Considering **register**, the frequency of neologisms is at its highest in letters written to close friends (examples (2) and (3) above) and, to a lesser extent, other acquaintances (1), whereas letters to family members do not seem to be particularly fertile ground for neologisms. This is true for both the men and the women in the sample.

Looking at **age**, it seems that older people (40–70) use neologisms more frequently than do younger people (17–39); the normalized frequencies are 21 and 10 per 10,000 words, respectively. This observation holds even if the age groups are split at 50 years, shifting two of the outliers, Elizabeth Stuart and Thomas Harrison, to the younger side; in that case, the normalized frequency is 18 for the older group and 15 for the younger. The youngest user of neologisms in our sample is the above-mentioned philosopher, Lady Conway, at twenty. The oldest person in the sample is the seventy-year-old Sir Hamon L'Estrange, a Royalist politician, who uses five new words in a single letter to his physician, Dr Thomas Browne. Some of these are related to his ailment and potential cures found elsewhere (*acrimonious*, *oversweetness*, *manifesto*), while others describe L'Estrange's own efforts in science as *crawling* (n.) and expect *candour* from Browne. His skilful use of recent vocabulary is not unexpected, as Kyle [19] characterizes L'Estrange as a "cultured and articulate man" who had many interests and who accumulated a large library.

In terms of **regional variation**, the amount of data per category permits us to say very little, but going beyond the existing metadata categories, we make the impressionistic observation that people residing abroad, either permanently or more temporarily, frequently use new words, which are naturally often borrowed. This goes for Elizabeth and the two nuns mentioned above, but also e.g. two Royalist noblemen, Thomas Howard, Earl of Arundel and Surrey, and his son, William Howard, Viscount Stafford, who travelled on the continent during the war. In fact, they produced three of the most interesting neologisms in our sample: *packet-boat* (4), *statement* (5) and *tea* (6). Not only do these antedate the first attestations provided by the OED, but we have also been unable to find earlier instances in massive databases of contemporary published texts like Early English Books Online [11].

(4) S. Jhon Pennington, just nowe Count Fabroni, & President Cognewe are come unto me from Q: Mother, to entreate very earnestly, that the gentleman cominge alonge wth this called Don Martino Dugaldi may instantly passe to Dunkerke for her M^{ties} especiall service w^{ch} depends soe much upon it as upon his retorne or any others sent before by y^{e} **Packette Boate** […]

(Thomas Howard in Dover to Sir John Pennington, 1641; ARUNDEL_068; OED3 first attestation 1642)

(5) I have receaved onely one letter in which there is a **statement** that the ssouldiers went to Mr John Penneducks house at King berry and ransaked it totally […]

(William Howard in Antwerp to Thomas Howard, 1642; ARUNDEL_072; OED3 first attestation 1750)

(6) I have scarce bought any thinge for my selfe but an Indian Brewhouse for **tee**, which hath beene very good Black Lack worke, but it is all spoyled and rased and yett I payed exceeding deare for it.

(William Howard in Amsterdam to Aletheia Howard, 1643; ARUNDEL_074; OED2 first attestation 1655)

Examples (4) and (5) seem to be related to the beginnings of the Civil War. As noted by Säily et al. [37] regarding Viscount Stafford and example (6),

He was writing from Amsterdam to his mother, who was also staying in the Low Countries. We're not really sure what an "Indian Brewhouse" is, but it's clearly not an actual house, being lacquerware (which was frequently imported from the East Indies); rather, it seems to be some sort of a container for tea. It makes sense that being in the Low Countries, William and his mother would have known the Dutch word for tea. Furthermore, the word isn't marked in any way in the text (the boldface was added by us), which indicates that it was quite a normal word for them – often, the new words we encounter in letters have been underlined or explained with another word as a sign of their novelty. It seems, then, that the upper social ranks of the time, at least those who travelled on the continent, could have been quite familiar with tea-drinking, so the early history of the word in English is not just about technical discussions of the plant but about everyday usage as well.

5 Eighteenth-Century Neologisms

5.1 Overview

Even though the 18th-century sample was larger than the 17th-century one, it only yielded 21 neologism types, listed below. Each of the types only occurred once in the sample. Antedatings to the OED are marked in boldface on the list, and words whose exact sense or part of speech could not be found in the OED are underlined. For the latter, we came up with suitable metadata based on neighbouring entries or senses in the OED.

anecdote-monger, blacky, cream-can, dénouement, floreat (n.), foundling-house, funny, **hookah**, **inspectress**, **interference**, jumpable, lovee, lumber-room, **merry-Andrew-like**, miliary fever, moonery, **mooning** (n.), namby-pamby, **pudding-less**, sentimental, tittup

Most of the words (14) are again nouns, followed by adjectives (5), verbs (1) and adverbs (1). Unlike the seventeenth century, the vast majority of the words have been formed within English through derivation (12), compounding (4) or conversion (2), and there are only two borrowings, *dénouement* from French and *hookah* from Arabic, and one verb of unclear etymology, *tittup*. Looking at the top level of the HT classification and the main semantic class of each word, 9 of the neologism types relate to 'the world' and 6 each to 'society' and 'the mind'. Within the classes, the words are again widely dispersed, but the most frequent categories are 'the mind » emotion' and 'society » leisure' at 3 types each. This paints a picture of letters written more for the purposes of keeping in touch and building friendships than for exchanging information, as in example (7). This may be partly explained by the composition of the sample: in terms of the relationship between the sender and recipient, the largest category is letters between close friends (31%), the proportion of which in the 17th-century sample is c. 21%.

(7) Your invocation has mounted me, **Merry Andrew-like**, upon stilts. – I ape you, as monkeys ape men by walking upon two.

(Ignatius Sancho to his friend, William Stevenson, 1777; SANCHO_016)

Table 3. Normalized frequency of neologism tokens per 10,000 words in the 18th-century sample for the three main social axes of interest, sorted by frequency.

Category	Value	Frequency / 10,000 words
Gender	Male	5
	Female	4
Social Rank	Other non-gentry	14
	Clergy	7
	Nobility	4
	Professionals	4
	Gentry	3
	Royalty	0
	Merchants	0
Relationship	Close friends	7
	Nuclear family	5
	Other family	3
	Other acquaintances	1

5.2 Sociolinguistic Variation

In contrast to the 17th century, the **social rank** with the highest relative frequency of neologisms in the sample (14 tokens per 10,000 words) is the lowest rank of all, other non-gentry; see Table 3. Two out of the five individuals representing this rank use new vocabulary: Henry Barnes (*foundling-house*) and Ignatius Sancho (*blacky*, *lovee*, *merry-Andrew-like*, *namby-pamby*). A husband to one of the nurses at the Foundling Hospital in London, Barnes addressed his letter to the matron of the hospital as in (8). The hospital was founded in 1739 and seems to have been the first of its kind in England [4], which explains the novelty of words referring to it.

(8) For the meatrern of the **fondlen house**

(Henry Barnes to Elizabeth Leicester, matron of the hospital, 1762; FOUNDLI_126)

Sancho, on the other hand, is quite a playful language user; see (7) above. He was an orphaned slave who ended up in England and found a position in the Duke of Montagu's household, later establishing a grocery shop [2]. He published several newspaper essays and rubbed shoulders with the literati of the time, and it seems that his social aspirations also show up in his use of neologisms.

In terms of **gender**, men tend to use more neologisms than women, as was also the case in the 17th-century sample. However, the proportion of female writers who use at least one new word is somewhat higher (5 out of 20) than the proportion of male writers (5 out of 33). The women come from the social ranks of gentry (Sarah Lennox and Hester Lynch Thrale), professionals (Frances Burney, author, and Mary Rawson Hart Boddam, wife of an East India Company employee, later governor, in Bombay), and clergy (Mary Colton, a vicar's wife and *inspectress* of the Foundling Hospital). Both of

the loanwords in the dataset again come from women: *dénouement* was used by Hester Lynch Thrale to her literary friend, Frances Burney, discussing a plot, while Mary Rawson Hart Boddam described her new husband's smoking habits in Bombay as in (9). For her part, Frances Burney wrote to her dramatist friend, Samuel Crisp, about herself as his *anecdote-monger* and about her father's *interference* in matters of matrimony, whereas Lady Sarah Bunbury née Lennox described a mutual relative as *funny* to her friend, Lady Susan O'Brien.

(9) He has been long at a subbordinate Factory and is a meer Moorman as to the language taste and customs, and will suck a Hubble Bubble, draw a Ailloon, smoak a **hooka** or **cream-cann** with you if you please; he has promised to wait on you with an account of us and to show you his different smoaking Machines being curious that way.

(Mary Rawson Hart Boddam to her uncle, Thomas Pickering, 1760; DRAPER_002)

As for **register**, letters to close friends (as in (7) above) again show the most frequent use of neologisms, but this time they are followed by letters written to family members, and letters to other acquaintances like (8) are the least innovative register in this respect. In addition to Boddam above, clergyman Thomas Twining uses novel vocabulary when writing to his half-brother, Daniel, and Sir Roger Newdigate to his wife, Sophia. While Newdigate mostly writes with news from London like a friend having *miliary fever*, Twining's letters are quite funny and aimed at entertaining the recipient, as in (10); the 16-year-old Daniel had tuberculosis and was staying at Bristol Hotwells for his health.

(10) Therefore, stick close to your Molière! for I shall set you a-translating again whenever I get you here; & the vanquished shall be **pudding-less** for *two* days, & not have three puddings for it on the third!

(Thomas Twining to his half-brother, Daniel Twining, 1764; TWINING_005)

Considering **age**, this time it seems that younger people under 40 use neologisms more frequently than those aged 40 and over (5 vs. 4 tokens per 10,000 words, respectively). If, however, we exclude the five new words by Thomas Twining, who has been shown to be an outlier in his prolific neologizing in the corpus as a whole [35], the numbers shift in favour of older people, as the younger group is left with only 3 tokens per 10,000 words. In our sample, the oldest person using new words is the above-mentioned Sir Roger Newdigate at 54, whereas the youngest is Mary Rawson Hart Boddam at fifteen or sixteen. Another gentleman of middling years is Horace Walpole, politician and later Earl of Orford, who wrote at 51 to his friend, poet Thomas Gray, about a literary novelty (11).

(11) I think you will like Sterne's **sentimental** travels, which tho often tiresome, are exceedingly goodnatured & picturesque.

(Horace Walpole to his friend, Thomas Gray, 1768; GRAY_065)

As the 18th-century section of the corpus was not designed to be regionally representative, we shall omit **regional variation**, noting only that the young Mary Boddam serves as a good example of vocabulary acquired abroad; see (9) above.

6 Discussion

What do the samples tell us about our three social axes of interest? In terms of **gender**, male writers seem to lead the way, although the difference is less clear-cut in the eighteenth century. While there is very little previous sociolinguistic research on this, we may note a study by Keune [18], who found that in Present-Day Dutch, the greatest lexical productivity was exhibited by highly educated older men, which also matches our results on age (unfortunately, we do not have enough data on the education of our informants to be able to use it as a category in the analysis). Furthermore, Säily and her collaborators have found that where there is variation in the productivity of certain nominal suffixes from Early Modern to Present-Day English, it is typically men who use them more productively than women [33,38]. Säily links this to men's more nominal style, which has been observed in both historical and Present-Day English.

As for **social rank**, the results vary wildly by time period. This could be due to the small size of the samples, but it could also reflect the social history of the two periods: in the seventeenth century, access to education, specialized registers, and new things and ideas in general was typically only available to the upper and middling ranks, whereas the eighteenth century saw some improvements in this respect. Be that as it may, what we can say is that clearly it is not only men and the higher ranks who use new vocabulary, which implies that it is important to conduct more studies like ours which do not focus on published texts alone but which aim to represent a wider section of the populace. In a previous study, too, we found neologisms used by the lower ranks that were underdocumented in the OED, including *Norfolker* 'a Norfolk sheep or cow', a term that could merit inclusion in the OED, and an antedating to *winterer* 'an animal kept over the winter' [35]. We have already submitted some of our antedatings to the OED, and entries are constantly being improved through ongoing work on the third edition of the dictionary.

Register, or the relationship between the writer and the recipient of the letter, has provided us with the most consistent results: it seems that letters between close friends are the most conducive to the use of new vocabulary. This is in accordance with Wolfson's "bulge theory" [41], which posits that a less stable relationship, such as that between status-equal friends, is subject to more negotiation and differs from both intimate family relationships and socially more distant relationships in terms of language use. Friends may wish to impress each other and put more effort into generating in-group solidarity, which could trigger more creative language use; cf. [34].

What can we say about differences in neologism use between the two **time periods** represented by our samples? In 1640–1660, we are able to observe the "Civil-War effect" discussed by Raumolin-Brunberg [31] and Lijffijt et al. [20]: the frequency of neologisms is much greater in this period than in the comparatively more tranquil 1760–1780, when wars were mostly fought further afield, and many of the 17th-century words are either directly or indirectly related to the war. We would argue that this phenomenon

is not limited to the English Civil War but can be seen as part of what Dixon [6] calls "punctuated equilibria", which is the notion that language history is characterized by periods of relative stability punctuated by external events that cause sudden changes in the linguistic situation and hence accelerate the rate of linguistic change. Nevalainen et al. [26] have shown that in Middle English, such punctuating events included the Norman Conquest and the Black Death; at the lexical level, the legacy of the conquest shows up in the many French loanwords in English.

Comparing the most frequent semantic classes of the time periods, while the samples are small and admittedly biased, the change from 'society » authority' to 'society » leisure' and 'the mind » emotion' may be indicative of a wider change in the styles and purposes of letter-writing. As noted by Somervell [39], in the eighteenth century "letter-writing conventions became less formal, with their subject-matter including private as well as public matters, and letters were becoming an artistic, moral and intellectual literary form". According to Säily [34, p. 214], "It is possible that letter-writing, at least for the upper classes in our corpus [i.e., the 18th-century section of the CEEC], was more about maintaining and building social relationships and identities than about conveying information." While the need to name things has always been one of the functions of neologism use (e.g. *tea*, *hookah*, *foundling-house*), eighteenth-century letter-writers in particular seem to delight in using inventive terms to describe people and their actions (e.g. *anecdote-monger*, *lovee*, *namby-pamby*, *moonery*) or to discuss mutual interests like literature (e.g. *dénouement*, *sentimental*).

7 Conclusions

This paper has studied the users of new vocabulary and the uses to which they put it in two samples of English letters from the seventeenth and eighteenth centuries. We have found that while male letter-writers tend to use neologisms more frequently than women, the eighteenth century seems to have provided more opportunities for women and the lower ranks to participate in neologism use as well. In both samples, neologisms most frequently occur in letters written between close friends, which could be due to this less stable relationship triggering more creative language use. In the seventeenth-century sample, we have observed the influence of the Civil War, while the eighteenth-century sample appears to reflect the changing functions of letter-writing, as correspondence is increasingly being used as a tool for building and maintaining social relationships in addition to exchanging information.

In this study, we have only analysed words that could be mapped to the OED, thus focusing on lexical items that were at some point well-established enough in the language to warrant their inclusion in the dictionary. There are, however, a number of potential neologisms that are not listed in the OED at all; while some of these may be nonce-words, others could be more established words that have simply been missed by the OED editors, especially when it comes to the second edition, which was compiled in the nineteenth and early twentieth centuries. It is a well-known fact that the eighteenth century has been less well covered by the OED than the seventeenth century [1], and so it is perhaps no surprise that we find more of these neologism candidates there. Out of the words not mapped to the OED, there are 13 viable candidates in the eighteenth-

century sample and only 4 in the seventeenth century. (It should be noted that even if these words were included in our analysis, the frequency of neologism use would still be higher in the seventeenth century, supporting our hypothesis of the Civil-War effect.) While many of them are Anglo-Indian words used by Mary Boddam and her sister, Eliza Draper, others are more general formations, such as *fellow-labourer*, *Pelhamized* and *soul-cheering*. Whether these are mere nonce-formations could be assessed by tracking them in large databases of contemporary published texts, like Eighteenth Century Collections Online [10].

As our quantitative approach has been based on comparing word forms (or strings separated by spaces) with those occurring earlier in the corpus as well as with the OED, we are missing many instances of homonyms, conversion, and multi-word units. According to the OED, however, the most frequent etymology types are borrowing and derivation [24, p. 351], and we would expect these to be fairly well covered by our approach, so we may in fact have captured the majority of the neologisms in the samples. In future work, we could nevertheless attempt to improve our coverage of the other means of word-formation: as an example, it would be relatively easy to automatically compare bigrams in addition to unigrams to capture more of the compounds.

Despite the improvements in the normalization step [15], normalization of the entire CEEC is still a problem that is far from solved. While using synthetic data improves low-resourced sequence-to-sequence models including character-level models [13,16], our experiments with back-translation on the training data available to us have not yielded better accuracies. This is due to the fact that our training data comes from a different distribution and presents variation that is very different from the variation in the CEEC. The training data comes mainly from authoritative sources, such as historical forms recorded in the OED or the MED. This means that if any synthetic data is used, it should be representative of the historical variation in the CEEC itself. This, however, is not an easy task and it is something to be solved in the future.

In the course of our work, we have learned to ask more focused questions rather than attempting to analyse the corpus as a whole, as the amount of manual labour required to identify and classify neologisms is considerable. While neural machine-learning approaches may help with the normalization task necessary for identifying neologisms, it seems that to fully realize their potential, we would need much more data to be able to train better models. It is to be hoped that more and more letters and other egodocuments will be digitized in the future, as their contents and social representativeness are of interest to a great number of scholars in the digital humanities.

Acknowledgements

We would like to thank the anonymous reviewers for helpful feedback. This work was supported in part by the Academy of Finland, Grants 293009 and 323390.

References

1. Brewer, C.: Treasure-house of the language: The living OED. Yale University Press, New Haven (2007)

2. Carretta, V.: Sancho, (Charles) Ignatius (1729?–1780), author. In: Oxford Dictionary of National Biography. Oxford University Press, Oxford, online edn. (2021). https://doi.org/10.1093/ref:odnb/24609
3. Chakravarthi, B.R., Rani, P., Arcan, M., McCrae, J.P.: A survey of orthographic information in machine translation. arXiv preprint arXiv:2008.01391 (2020)
4. Clark, G. (ed.): Correspondence of the Foundling Hospital inspectors in Berkshire 1757–68, Berkshire Record Society, vol. 1. Berkshire Record Society, Reading (1994)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1423
6. Dixon, R.M.W.: The rise and fall of languages. Cambridge University Press, Cambridge (1997)
7. Domingo, M., Casacuberta, F.: Modernizing historical documents: A user study. Pattern Recognition Letters **133**, 151–157 (2020)
8. Dubossarsky, H., Hengchen, S., Tahmasebi, N., Schlechtweg, D.: Time-out: Temporal referencing for robust modeling of lexical semantic change. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 457–470. Association for Computational Linguistics, Florence, Italy (2019). https://doi.org/10.18653/v1/P19-1044
9. Duong, Q., Hämäläinen, M., Hengchen, S.: An unsupervised method for OCR post-correction and spelling normalisation for Finnish. arXiv preprint arXiv:2011.03502 (2020)
10. ECCO: Eighteenth Century Collections Online. Gale (nd), https://www.gale.com/primary-sources/eighteenth-century-collections-online
11. EEBO: Early English Books Online. ProQuest (nd), https://proquest.libguides.com/eebo
12. Grieve, J., Nini, A., Guo, D.: Analyzing lexical emergence in Modern American English online. English Language and Linguistics **21**(1), 99–127 (2017). https://doi.org/10.1017/S1360674316000113
13. Hämäläinen, M., Rueter, J.: Finding Sami cognates with a character-based NMT approach. In: Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers). pp. 39–45. Association for Computational Linguistics, Honolulu (2019), https://www.aclweb.org/anthology/W19-6006
14. Hämäläinen, M., Säily, T., Rueter, J., Tiedemann, J., Mäkelä, E.: Normalizing early English letters to Present-day English spelling. In: Proceedings of the 2nd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature. pp. 87–96. Association for Computational Linguistics, Santa Fe, New Mexico (2018), http://aclweb.org/anthology/W18-4510
15. Hämäläinen, M., Säily, T., Rueter, J., Tiedemann, J., Mäkelä, E.: Revisiting NMT for normalization of early English letters. In: Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature. pp. 71–75. Association for Computational Linguistics, Minneapolis, USA (2019). https://doi.org/10.18653/v1/W19-2509
16. Hämäläinen, M., Wiechetek, L.: Morphological disambiguation of South Sámi with FSTs and neural networks. In: Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL). pp. 36–40. European Language Resources Association, Marseille, France (2020), https://www.aclweb.org/anthology/2020.sltu-1.5
17. Hengchen, S., Tahmasebi, N., Schlechtweg, D., Dubossarsky, H.: Challenges for computational lexical semantic change. arXiv preprint arXiv:2101.07668 (2021)

In Hämäläinen, M., Partanen, N., Alnajjar, K. (eds.) Multilingual Facilitation (2021), pages 153–169.

18. Keune, K.: Explaining register and sociolinguistic variation in the lexicon: Corpus studies on Dutch. Ph.D. thesis, Radboud University Nijmegen (2012), https://hdl.handle.net/2066/101694
19. Kyle, C.R.: L'Estrange, Sir Hamon (1583–1654), politician. In: Oxford Dictionary of National Biography. Oxford University Press, Oxford, online edn. (2005). https://doi.org/10.1093/ref:odnb/67340
20. Lijffijt, J., Säily, T., Nevalainen, T.: CEECing the baseline: Lexical stability and significant change in a historical corpus. In: Tyrkkö, J., Kilpiö, M., Nevalainen, T., Rissanen, M. (eds.) Outposts of historical corpus linguistics: From the Helsinki Corpus to a proliferation of resources. Studies in Variation, Contacts and Change in English, vol. 10. VARIENG, Helsinki (2012), https://varieng.helsinki.fi/series/volumes/10/lijffijt_saily_nevalainen/
21. Mäkelä, E., Lagus, K., Lahti, L., Säily, T., Tolonen, M., Hämäläinen, M., Kaislaniemi, S., Nevalainen, T.: Wrangling with non-standard data. In: Proceedings of the Digital Humanities in the Nordic Countries 5th Conference, Riga, Latvia, October 21–23, 2020. CEUR Workshop Proceedings, vol. 2612, pp. 81–96. CEUR-WS.org, Aachen (2020), http://ceur-ws.org/Vol-2612/paper6.pdf
22. McCrae, J.P.: Identification of adjective-noun neologisms using pretrained language models. In: Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019). pp. 135–141. Association for Computational Linguistics, Florence, Italy (2019). https://doi.org/10.18653/v1/W19-5116
23. MED: Middle English Dictionary. University of Michigan (nd), https://quod.lib.umich.edu/m/med/
24. Nevalainen, T.: Early Modern English lexis and semantics. In: Lass, R. (ed.) The Cambridge history of the English language, III: 1476–1776, pp. 332–458. Cambridge University Press, Cambridge (1999). https://doi.org/10.1017/CHOL9780521264761.006
25. Nevalainen, T., Raumolin-Brunberg, H., Keränen, J., Nevala, M., Nurmi, A., Palander-Collin, M., Kaislaniemi, S., Laitinen, M., Säily, T., Sairio, A.: CEEC, Corpora of Early English Correspondence. Department of Modern Languages, University of Helsinki (1998–2006), https://varieng.helsinki.fi/CoRD/corpora/CEEC/
26. Nevalainen, T., Säily, T., Vartiainen, T., Liimatta, A., Lijffijt, J.: History of English as punctuated equilibria? A meta-analysis of the rate of linguistic change in Middle English. Journal of Historical Sociolinguistics **6**(2), 20190008 (2020). https://doi.org/10.1515/jhsl-2019-0008
27. Oxford English Dictionary: OED Online. Oxford University Press (nd), http://www.oed.com/
28. Palmer, C.C.: Measuring productivity diachronically: Nominal suffixes in English letters, 1400–1600. English Language and Linguistics **19**(1), 107–129 (2015). https://doi.org/10.1017/S1360674314000264
29. Partanen, N., Hämäläinen, M., Alnajjar, K.: Dialect text normalization to normative standard Finnish. In: Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019). pp. 141–146. Association for Computational Linguistics, Hong Kong, China (2019). https://doi.org/10.18653/v1/D19-5519
30. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 2227–2237. Association for Computational Linguistics, New Orleans, Louisiana (2018). https://doi.org/10.18653/v1/N18-1202
31. Raumolin-Brunberg, H.: Social factors and pronominal change in the seventeenth century: The Civil-War effect? In: Fisiak, J., Krygier, M. (eds.) Advances in English historical linguistics (1996), Trends in Linguistics: Studies and Monographs, vol. 112, pp. 361–388. Mouton de Gruyter, Berlin (1998). https://doi.org/10.1515/9783110804072.361

32. Renouf, A.: Tracing lexical productivity and creativity in the British media: 'The chavs and the chav-nots'. In: Munat, J. (ed.) Lexical creativity, texts and contexts, Studies in Functional and Structural Linguistics, vol. 58, pp. 61–89. John Benjamins, Amsterdam (2007). https://doi.org/10.1075/sfsl.58.12ren
33. Säily, T.: Sociolinguistic variation in English derivational productivity: Studies and methods in diachronic corpus linguistics, Mémoires de la Société Néophilologique de Helsinki, vol. XCIV. Société Néophilologique, Helsinki (2014)
34. Säily, T.: Change or variation? Productivity of the suffixes *-ness* and *-ity*. In: Nevalainen, T., Palander-Collin, M., Säily, T. (eds.) Patterns of change in 18th-century English: A sociolinguistic approach, Advances in Historical Sociolinguistics, vol. 8, pp. 197–218. John Benjamins, Amsterdam (2018). https://doi.org/10.1075/ahs.8.12sai
35. Säily, T., Mäkelä, E., Hämäläinen, M.: Explorations into the social contexts of neologism use in early English correspondence. Pragmatics & Cognition **25**(1), 30–49 (2018). https://doi.org/10.1075/pc.18001.sai
36. Säily, T., Mäkelä, E., Hämäläinen, M.: Neologism dataset for "From *plenipotentiary* to *puddingless*: Users and uses of new words in early English letters" (2021). https://doi.org/10.5281/zenodo.4578940
37. Säily, T., Mäkelä, E., Kaislaniemi, S.: Cha before tea: Finding earlier mentions in a corpus of early English letters (part 1). In: Oxford English Dictionary blog. Oxford University Press, Oxford (2019), https://public.oed.com/blog/cha-before-tea-finding-earlier-mentions-in-a-corpus-of-early-english-letters-part-1/
38. Säily, T., Suomela, J.: *types*2: Exploring word-frequency differences in corpora. In: Hiltunen, T., McVeigh, J., Säily, T. (eds.) Big and rich data in English corpus linguistics: Methods and explorations, Studies in Variation, Contacts and Change in English, vol. 19. VARIENG, Helsinki (2017), https://varieng.helsinki.fi/series/volumes/19/saily_suomela/
39. Somervell, T.: Public and private, real and fictional: The rise of women's letter-writing in the eighteenth century. Bluestocking: Online Journal for Women's History (2011), https://blue-stocking.org.uk/2011/03/01/public-and-private-real-and-fictional-the-rise-of-womens-letter-writing-in-the-eighteenth-century/
40. Wein, S.: Classification and analysis of neologisms produced by learners of Spanish: Effects of proficiency and task. In: Proceedings of the The Fourth Widening Natural Language Processing Workshop. pp. 88–91. Association for Computational Linguistics, Seattle, USA (2020). https://doi.org/10.18653/v1/2020.winlp-1.22
41. Wolfson, N.: The bulge: A theory of speech behavior and social distance. Penn Working Papers in Educational Linguistics **2**(1), 55–83 (1990), https://repository.upenn.edu/wpel/vol2/iss1/3/
42. Zalmout, N., Thadani, K., Pappu, A.: Unsupervised neologism normalization using embedding space mapping. In: Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019). pp. 425–430. Association for Computational Linguistics, Hong Kong, China (2019). https://doi.org/10.18653/v1/D19-5555
43. Zarnoufi, R., Jaafar, H., Bachri, W., Abik, M.: MANorm: A normalization dictionary for Moroccan Arabic dialect written in Latin script. In: Proceedings of the Fifth Arabic Natural Language Processing Workshop. pp. 155–166. Association for Computational Linguistics, Barcelona, Spain (2020), https://www.aclweb.org/anthology/2020.wanlp-1.14

