BRILL

RESEARCH DATA JOURNAL FOR THE HUMANITIES AND SOCIAL SCIENCES 9 (2024) 1–12

BRILL OPEN

brill.com/rdj

The Corpus of Early English Correspondence Extension Sampler (CEECES)

Samuli Kaislaniemi | ORCID: 0000-0002-3596-1341
School of Humanities, University of Eastern Finland, Joensuu, Finland
Corresponding author
samuli@kaislaniemi.fi

Lassi Saario | ORCID: 0000-0002-5936-7996
Department of Philosophy, History and Art Studies, University of Helsinki, Helsinki, Finland
lassi.saario@helsinki.fi

Tanja Säily | ORCID: 0000-0003-4407-8929
Department of Languages, University of Helsinki, Helsinki, Finland
tanja.saily@helsinki.fi

Received 31 March 2023 | Revised 21 September 2023 |
Accepted 15 January 2024 | Published online 15 February 2024

Abstract

This data paper describes the *Corpus of Early English Correspondence Extension Sampler* (CEECES), a linguistic corpus of personal letters covering the long eighteenth century. The letters have been sampled and transcribed from various printed editions and are now openly distributed through Zenodo. The CEECES contains 2,624 letters by 200 writers, some 1.14 million words. It comes in several versions – plain text, XML, standardised-spelling, and part-of-speech tagged – with ample metadata on the correspondents and the letters, enabling the sociolinguistic study of historical English using a range of social variables including gender, age, social rank, and geographical region.

Published with license by Koninklijke Brill BV | DOI:10.1163/24523666-bja10034
 | ISSN: 2452-3666 (online)

Keywords

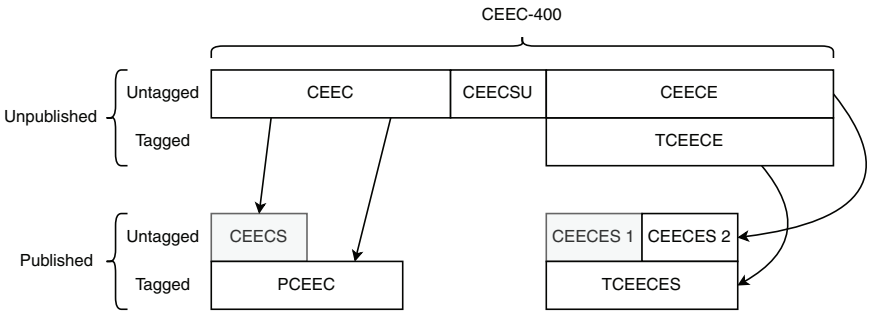
corpus linguistics – letters – correspondence – Late Modern English – English language – historical sociolinguistics – England – eighteenth century

– Related data sets "CEECES1" with DOI www.doi.org/10.5281/zenodo.4644243; "CEECES2" with DOI www.doi.org/10.5281/zenodo.5887100; and "TCEECES" with DOI www.doi.org/10.5281/zenodo.5887230 in repository "Zenodo"

1. Introduction

The *Corpus of Early English Correspondence Extension Sampler* (CEECES) is the third release from the *Corpora of Early English Correspondence* (CEEC-400), a family of linguistic resources built for the sociolinguistic study of historical English. The CEEC-400 contains over 5 million words from nearly 12,000 letters spanning 1402–1800. To date, some 2.2 million words from 1410–1681 have been released (CEECS in 1998, PCEEC in 2006 and PCEEC2 in 2022). The CEECES extends this coverage, adding over 1.1 million words dating from 1653 to 1800.

The CEECES is a selection (a 'sampler') from the CEEC *Extension* (CEECE), which contains more than 2.2 million words from 1653 to 1800 (see Kaislaniemi, 2018). The CEECE was completed in 2012, but its publication was hindered by difficulties in obtaining permissions from copyright holders. To remedy the situation, it was decided to release those parts of the CEECE which were 1) out of copyright, and 2) for which we have already received full permission from the copyright holders. These datasets were published as the CEECE *Sampler* parts 1 and 2 (CEECES 1 and CEECES 2), respectively. Further, it was



Note: Those corpora that are out of copyright have been highlighted in grey.

FIGURE 1 The released datasets from the CEEC-400 family of corpora

decided to complement these with 3) the same text collections taken from the *Tagged* CEECE (TCEECE); these were published as the *Tagged* CEECE *Sampler* (TCEECES). The mutual relationships of these various corpora are illustrated in Figure 1.

2. Context

The CEEC-400 was compiled for the purposes of historical sociolinguistics. The original idea was to test the extent to which hypotheses derived from present-day sociolinguistics – such as "women tend to lead language change" – would be supported by empirical evidence in historical material covering hundreds of years. In the absence of spoken data, personal letters are ideal for sociolinguistic research of historical periods: they are speech-like, they have identifiable senders and recipients, and unlike published texts, they could be written by anyone who was literate. The corpus covers a wide social spectrum spanning from housemaids to kings. As such, it is of interest not only to scholars of language history but also to e.g. social historians. Compared to most other corpora of English historical correspondence, the CEEC-400 is larger (thanks to its compilation process, see below) and covers a wider section of the populace. The part-of-speech tagging enables studies at a higher level of abstraction, including stylistic trends in the evolution of the letter genre.

Examples of research conducted using the CEEC-400 include Nevalainen and Raumolin-Brunberg (2003, 2017), which is the seminal work in historical sociolinguistics. Nevalainen and Raumolin-Brunberg analyse the time course and social embedding of fourteen linguistic changes in English in 1410–1681, from the replacement of subject *ye* by *you* to the decline of multiple negation. The findings include that even in the past, women indeed tend to lead most changes in language, and that social aspirers often follow the lead of their social superiors. Using the CEECE, this research is extended into the eighteenth century by Nevalainen et al. (2018), who find that the female advantage holds there as well but that the pace of change seems to have been slower than in the preceding centuries, possibly retarded by the ideology of standardisation. Degaetano-Ortlieb et al. (2021) use the TCEECE to compare the language use of women and men at three linguistic levels: vocabulary, morphology (derivational suffixes) and grammar (part-of-speech trigrams). They find that middle- and upper-class women tend to innovate in the informal setting of family letters and that women lead changes at all three levels, contributing to the colloquialisation of the letter genre over time.

The publication of the CEECES provides the wider research community with the opportunity to conduct its own studies of eighteenth-century English using this rich dataset.

3. Corpus Compilation

The CEEC-400 was compiled primarily from previously published, printed editions (see Kaislaniemi, 2023). Only editions that preserved the original manuscript spellings were chosen. Text selection was guided by the aim of socio-regional coverage: to have as good a cross-section of literate English society as possible. Social categories used as selection criteria by the compilers include gender, social rank, and region (see Raumolin-Brunberg & Nevalainen, 2007). The corpus is organised into *collections*, which usually contain letters from a single source edition. However, the collections do not contain all of the letters in their source editions, as the selection criteria controlled for both quality (excluding for example later copies) and quantity (20 letters per writer was considered representative; more were taken when a writer's correspondence spanned decades, but often only a few letters were available). In the CEECES, there are on average 13 letters (c. 5,700 words) from each writer.

The CEEC team scanned the chosen texts, then digitised them with OCR software, and proofread the results three times against the source edition. The texts were stored as plain text, which required the conversion of formatting into simple text encoding, for example, superscripts like "Sr" *Sir* are marked as `S=r=` (these conventions follow Kytö, 1996). More recently, the texts in the CEEC-400 have been converted to XML, the previous example becoming `<hi rend="sup" range="1,2">Sr</hi>` (see Saario, 2020).

Because the CEEC-400 was designed for sociolinguistic research, the corpus texts are accompanied by rich metadata, which contains information on the correspondents and on the letters. This metadata was gathered from all available sources and recorded into a spreadsheet by the CEEC team. Some metadata is also included in the corpus texts, in the headers of each collection and each letter (for details, see Kaislaniemi, 2018, 2022; Nurmi, 1998; for more, see Nevalainen & Raumolin-Brunberg, 2017, pp. 26–52).

The decision to provide the CEECE with part-of-speech tagging was based on a desire to study connections between word classes of the letter texts and the social backgrounds of the letter writers. The CLAWS tagger of Lancaster University was chosen for the task (UCREL, [1997]). Given that CLAWS is designed for present-day English and only accepts text-level coding in XML

format, the spelling of the CEECE texts had to be standardised and their format converted into XML before tagging. The spelling was standardised in two stages: first semi-automatically by the Variant Detector software (VARD 2; Baron, 2011a, 2011b) and then manually by a team of people paying special attention to remaining variation recognised as problematic for the tagging, such as obsolete abbreviations and non-modern punctuation marks (Saario & Säily, 2020). The standardised-spelling CEECE was then converted into XML and part-of-speech tagged by CLAWS (Saario et al., 2021).

4. Data Description

- **Corpus of Early English Correspondence Extension Sampler (CEECES) deposited at Zenodo**
  - CEECES 1 – doi:www.doi.org/10.5281/zenodo.4644243
    - License – CC BY-NC
  - CEECES 2 – doi:www.doi.org/10.5281/zenodo.5887100
    - License – CC BY-NC-ND
  - TCEECES – doi:www.doi.org/10.5281/zenodo.5887230
    - License – CC BY-NC-ND
- **Temporal coverage:** 1653–1800

The CEECES consists of 42 collections (see Appendix), which contain 2,624 letters written by 200 writers, coming to some 1.14 million words. The corpus texts are provided in plain text and XML formats, in both original and standardised-spelling versions, with the latter also provided with part-of-speech tagging (see Table 1).

Although the earliest letter in the CEECES is from 1653, there are only two letters from before 1680. Figure 2 shows the number of letters in the CEECES

TABLE 1 CEECES corpus versions and formats

	Spelling	Plain text – by letter			XML – by collection		
Annotation:		**Untagged**	**C5**	**C7**	**Untagged**	**C5**	**C7**
CEECES 1 & 2	original	YES			YES		
TCEECES	standardised		YES	YES	YES	YES	YES

over time, with the proportions of men's and women's letters per twenty-year period. (Gender representation in the CEECES is unequal because fewer women were literate in the first place, and thanks to gender bias their letters have been less likely to survive or to be edited and published: see Kaislaniemi, 2018, pp. 51–52). Figure 3 divides the data into social ranks by proportion of the word count per twenty-year period.

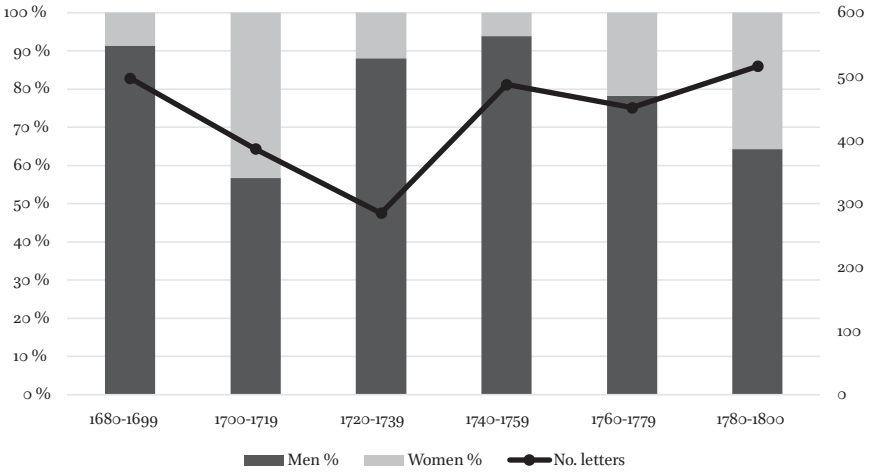


FIGURE 2 Number of letters in the CEECES over time, gender division (%)

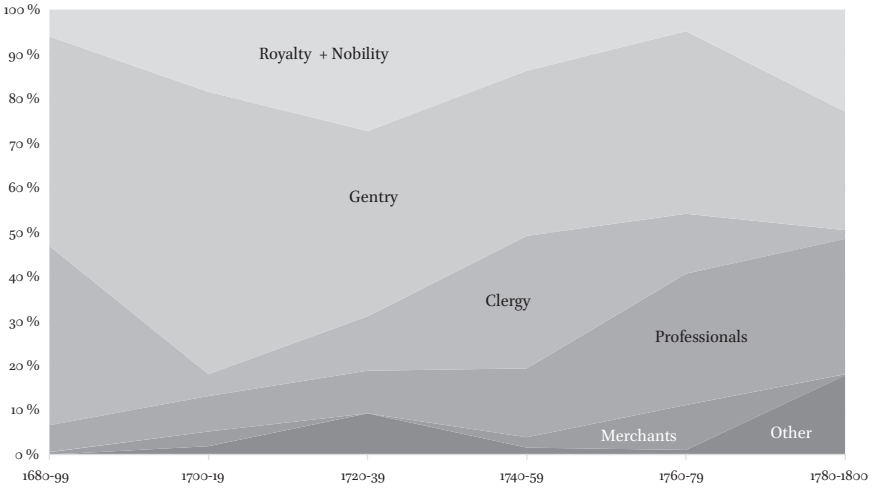


FIGURE 3 Words in the CEECES over time, social ranks (%)

The standout feature of the CEECES is its metadata, which is considerably more detailed than that accompanying the previously published sections of the CEEC-400 (CEECS and PCEEC), making the CEECES particularly well suited to sociolinguistic research. The database of metadata for the CEECES contains information on the gender, age and social status of the writers, as well as known details about their regional origins and formal education. The database also contains information on the recipients of the letters, on the relationship between the writer and recipient, and then information on the letters themselves, such as authenticity (is the letter autograph or a copy), the year of writing and word count. (See the CEECES manual [Kaislaniemi, 2022] for more information on the social breakdown of the letter-writers, and for comparisons of the CEECES with the CEECE and the CEEC-400).

An additional layer of information is included in the CEEC-400 corpora in text-level encoding. Part of the editorial apparatus has been made computer-readable, as the corpus retains and encodes information from the edition such as scribal emendations, insertions, hand changes, and damage to the manuscript sources. In addition to information added to the texts by the editor, the corpus texts also contain information added by the compilers. This includes the flagging of foreign words, and in particular, the addition of linguistic part-of-speech annotation.

The part-of-speech tagging is provided using two different tagsets: C5 and C7 (see UCREL, [1997]). The accuracy of the tagging has been evaluated by taking a subsample of the text and manually checking the tags assigned to it by CLAWS. The accuracy in the full TCEECE using the C7 tagset is estimated to be 94.5% overall. The accuracy is 95.4% for letters from men, 92.8% for letters from women, 93.5% for letters from the 17th century, and 94.7% for letters from the 18th century. The corresponding numbers for the C5 tagset are slightly higher in each case. The tagging of the PAUPER collection, which probably had the lowest accuracy (87.9%), has been manually corrected in its entirety. Precisions and recalls by particular tags and the frequencies of most common incorrect–correct tag pairs are provided in the TCEECE manual (Saario & Säily, 2020; see also Saario et al., 2021, for an account of its creation).

To get a concrete grip on the data, see, for instance, the closing formula of the 100th letter in the FLEMING 2 collection as it appears in the CEECES 1 and in the C7 version of the TCEECES (letter ID FLEMIN2_100):

1) CEECES 1 – plain text, original spelling:
```
So with my duty to your Self, and love and Service
to all with you I re[{main{]
```

2) TCEECES – XML, normalised spelling, part-of-speech tags:

```
So_RR with_IW my_APPGE duty_NN1 to_II yourself_PPX1
,_, and_CC love_NN1 and_CC Service_NN1 to_II all_
DB with_IW you_PPY I_PPIS1 <supplied range="2,6"
orig="re[{main{]"> remain_VV0 </supplied>
```

3) TCEECES – XML, normalised spelling, tokenised, part-of-speech tags:

```
<w id="1397.1" pos="RR">So</w>
<w id="1397.2" pos="IW">with</w>
<w id="1397.3" pos="APPGE">my</w>
<w id="1397.4" pos="NN1">duty</w>
<w id="1397.5" pos="II">to</w>
<w id="1397.6" pos="PPX1">yourself</w>
<w id="1397.7" pos=",">,</w>
<w id="1397.8" pos="CC">and</w>
<w id="1397.9" pos="NN1">love</w>
<w id="1397.10" pos="CC">and</w>
<w id="1397.11" pos="NN1">Service</w>
<w id="1397.12" pos="II">to</w>
<w id="1397.13" pos="DB">all</w>
<w id="1397.14" pos="IW">with</w>
<w id="1397.15" pos="PPY">you</w>
<w id="1397.16" pos="PPIS1">I</w>
<supplied range="2,6" orig="re[{main{]">
     <w id="1397.17" pos="VV0">remain</w>
</supplied>
```

As a comparison of examples (1–3) makes clear, the CEECES 1 retains the original spelling with minimal annotation whereas the TCEECES represents the same text with normalised spelling and heavy annotation. The underlying normalisation is the result of a long and complicated process including changes to tokenisation (*your Self → yourself*). Rather than trying to pack all this information in one all-encompassing format, which would hardly have been readable by any existing corpus tool, the two layers of annotation have been separated into parallel versions of the same letter (for more discussion, see Saario et al., 2021, pp. 125–127).

5. Concluding Remarks

The CEECES is a unique resource for sociohistorical research into the language of English personal letters in the long eighteenth century. Its structure makes it possible to study the language of one individual as easily as that of a certain period. Since it is openly available, it is also eminently suited for many digital humanities applications as well as for teaching.

References

Baron, A. (2011a). *VARD 2* [Computer software]. Lancaster University. Available from https://ucrel.lancs.ac.uk/vard.

Baron, A. (2011b). *Dealing with spelling variation in Early Modern English texts* (Publication No. 84887) [Doctoral dissertation, Lancaster University]. Lancaster University Library. https://eprints.lancs.ac.uk/id/eprint/84887.

CEEC-400 = *Corpora of Early English Correspondence*. Compiled by Terttu Nevalainen, Helena Raumolin-Brunberg, Samuli Kaislaniemi, Jukka Keränen, Mikko Laitinen, Minna Nevala, Arja Nurmi, Minna Palander-Collin, Tanja Säily and Anni Sairio at the Department of Modern Languages, University of Helsinki. https://varieng.helsinki.fi/CoRD/corpora/CEEC.

CEECS = *Corpus of Early English Correspondence Sampler* (1998). Compiled by Terttu Nevalainen, Helena Raumolin-Brunberg, Jukka Keränen, Minna Nevala, Arja Nurmi and Minna Palander-Collin at the Department of Modern Languages, University of Helsinki. Distributed through the Oxford Text Archive.

Degaetano-Ortlieb, S., Säily, T., & Bizzoni, Y. (2021). Registerial adaptation vs. innovation across situational contexts: 18th century women in transition. *Frontiers in Artificial Intelligence*, *4*, 609970. www.doi.org/10.3389/frai.2021.609970.

Kaislaniemi, S. (2018). The Corpus of Early English Correspondence Extension (CEECE). In T. Nevalainen, M. Palander-Collin, & T. Säily (Eds.), *Patterns of change in eighteenth-century English: A sociolinguistic approach* (pp. 45–59). John Benjamins. www.doi.org/10.1075/ahs.8.04kai.

Kaislaniemi, S. (2022). Brief manual to the *Tagged Corpus of Early English Correspondence Extension Sampler* (TCEECES). VARIENG. Available with CEECES.

Kaislaniemi, S. (2023). Editions and other sources used in the Corpora of Early English Correspondence (CEEC-400). Version 3. www.doi.org/10.5281/zenodo.4134471.

Kytö, M. (1996). *Manual to the diachronic part of the Helsinki Corpus of English Texts. Coding conventions and lists of source texts*. 3rd edition. Department of English, University of Helsinki. http://korpus.uib.no/icame/manuals/HC/INDEX.HTM.

Nevalainen, T. & Raumolin-Brunberg, H. (2003). *Historical sociolinguistics: Language change in Tudor and Stuart England*. Longman.

Nevalainen, T. & Raumolin-Brunberg, H. (2017). *Historical sociolinguistics: Language change in Tudor and Stuart England*. 2nd, revised edition. Routledge.

Nevalainen, T., Palander-Collin, M., & Säily, T. (Eds.) (2018). *Patterns of change in eighteenth-century English: A sociolinguistic approach*. John Benjamins. www.doi.org/10.1075/ahs.8.

Nurmi, A. (1998). Manual for the Corpus of Early English Correspondence Sampler CEECS. Department of English, University of Helsinki. http://korpus.uib.no/icame/manuals/CEECS/INDEX.HTM.

PCEEC = *Parsed Corpus of Early English Correspondence* (2006). Annotated by Ann Taylor, Arja Nurmi, Anthony Warner, Susan Pintzuk and Terttu Nevalainen. Compiled by Terttu Nevalainen, Helena Raumolin-Brunberg, Jukka Keränen, Minna Nevala, Arja Nurmi and Minna Palander-Collin. York: University of York and Helsinki: University of Helsinki. Distributed through the Oxford Text Archive.

PCEEC2 = *Parsed Corpus of Early English Correspondence 2* (2022). Revised and corrected by Beatrice Santorini. Annotated by Ann Taylor, Arja Nurmi, Anthony Warner, Susan Pintzuk and Terttu Nevalainen. Compiled by Terttu Nevalainen, Helena Raumolin-Brunberg, Jukka Keränen, Minna Nevala, Arja Nurmi and Minna Palander-Collin. York: University of York and Helsinki: University of Helsinki. www.github.com/beatrice57/pceec2.

Raumolin-Brunberg, H. & Nevalainen, T. (2007). Historical sociolinguistics: The Corpus of Early English Correspondence. In J. C. Beal, K. P. Corrigan, & H. L. Moisl (Eds.), *Creating and digitizing language corpora*, Vol. 2, *Diachronic databases* (pp. 148–171). Palgrave-Macmillan. Pre-print available at https://varieng.helsinki.fi/CoRD/corpora/CEEC/generalintro.html.

Saario, L. (2020). Conversion of the CEEC-400 into XML. A manual to accompany the XML edition. VARIENG. https://varieng.helsinki.fi/CoRD/corpora/CEEC/xml_doc.html.

Saario, L. & Säily, T. (2020). POS tagging the CEECE. A manual to accompany the Tagged Corpus of Early English Correspondence (TCEECE). VARIENG. Also included in the TCEECES bundle. https://varieng.helsinki.fi/CoRD/corpora/CEEC/tceece_doc.html.

Saario, L., Säily, T., Kaislaniemi, S., & Nevalainen, T. (2021). The burden of legacy: Producing the Tagged Corpus of Early English Correspondence Extension (TCEECE). *Research in Corpus Linguistics*, *9*(1), 104–131. www.doi.org/10.32714/ricl.09.01.07.

UCREL. [1997]. *CLAWS4* (Version 24) [Computer software]. Lancaster University. Available from http://ucrel.lancs.ac.uk/claws.

Appendix

For a list of the sources of the CEECES collections, see Kaislaniemi (2023).

TABLE A1 List of collections in the full CEECES

Collection	Years	Writers	Letters	Words
Banks	1704–1756	22	98	39,162
Blomefield	1730–1751	5	40	13,318
Bowrey	1687–1708	6	38	19,229
Champion	1774–1776	1	14	10,790
Clavering	1705?–1741	11	187	72,846
Clift	1792–1799	6	63	52,038
Cowper spencer	1732–1764	1	20	10,187
Crisp	1779–1782	1	22	18,389
Culley	1784–1785	3	21	25,100
Darwin	1763–1797	1	41	18,362
Dodsley	1743–1764	10	139	48,691
Draper	1757?–1775?	2	19	22,511
Dukes	1732–1750	4	121	50,858
Fleming 2	1653–1701	11	248	76,297
Fleming extra	1684–1698	1	52	14,020
Foundling	1758–1767	25	198	50,221
Garrick	1733–1777	1	93	42,832
George 3	1765–1783	1	36	7,765
George iii a	1779–1800	9	142	51,638
Giffard 2	1697–1722?	2	16	9,701
Gower	1783–1800	7	39	16,989
Gray	1734?–1771	4	73	42,694
Haddock 2	1688–1719	4	11	4,647
Hatton 2	1682–1704	11	78	25,575
Henry	1660–1693	2	23	10,637
Hurd	1739–1797	2	64	37,415

TABLE A1 List of collections in the full CEECES (*cont.*)

Collection	Years	Writers	Letters	Words
LENNOX	1761–1800	2	85	66,358
LIDDELL	1709–1716	1	51	36,816
NEWDIGATE	1731–1797	1	62	30,054
ORIGINAL 4	1682–1716	4	12	2,900
PAUPER	1731–1795?	12	12	2,220
PERROT	1799–1800	1	8	8,924
PETTY 2	1682–1687	2	36	14,378
PIOZZI	1784–1798	3	69	39,572
PITT	1751–1757	1	23	9,071
PITT 2	1754	2	24	15,618
PRIDEAUX 2	1681–1722	1	36	15,934
ROYAL 4	1681?–1683?	2	19	4,408
STUBS	1791–1800	7	21	4,274
TIXALL 2	1684–1686	1	2	392
WEDGWOOD	1763–1793	6	88	35,232
WENTWORTH 2	1705–1739	9	180	62,223
TOTAL	**1653–1800**	**208***	**2,624**	**1,140,286**

* Many writers occur in more than one collection: the actual total number of writers in the CEECES is 200.