Brief manual to the *Corpus of Early English Correspondence Extension Sampler* part 1 (CEECES 1)

Samuli Kaislaniemi

1. Background

Introduction

The *Corpora of Early English Correspondence* (CEEC-400) have been compiled in the Faculty of Arts at the University of Helsinki, by a team led by Prof. Terttu Nevalainen. The CEEC team has been slowly working towards releasing the entire CEEC-400 to the use of other scholars, but this process has been hindered by copyright legislation. A part of the CEEC-400 has already been published (CEECS in 1998; PCEEC in 2006), but more recently it has been difficult to obtain permissions from copyright holders, and the *CEEC Extension* (CEECE), which spans 1681–1800, has not been released (see Kaislaniemi 2018).

To remedy the situation, it was decided to release those parts of the *CEEC Extension* which were i) out of copyright, and ii) for which we have already received permission to publish from the copyright holders. The resulting subcorpus was named the *CEEC Extension Sampler* (CEECES), in line with the *CEEC Sampler* (CEECS) subcorpus published in 1998, compiled similarly of out-of-copyright materials. **The CEECES will be released in three parts over 2021** (CEECES 1, CEECES 2 and CEECES 3).

This brief manual outlines the structure and contents of the *CEEC Extension Sampler* part 1 (CEECES 1), and directs the user to other publications for more information on the corpus.

Copyright, reference line & distribution

Please cite the CEECES 1 as:

> CEECES 1 = *Corpus of Early English Correspondence Extension Sampler* part 1. Compiled by Terttu Nevalainen, Helena Raumolin-Brunberg, Samuli Kaislaniemi, Mikko Laitinen, Minna Nevala, Arja Nurmi, Minna Palander-Collin, Tanja Säily and Anni Sairio at the Department of Languages, University of Helsinki.

The CEECES 1 has been deposited in Zenodo:

- www.doi.org/10.5281/zenodo.4644244

At the release of parts 2 and 3, all parts of the CEECES will be made available in Zenodo, and the corpus will also be deposited at the Oxford Text Archive.

Bibliography and further information on the CEEC-400, CEECE and CEEC-400 XML

For information on the CEEC-400 and its subcorpora (CEEC (aka CEEC 1998), CEECE, CEECS, CEECSU, and PCEEC), see the CEEC entry in CoRD.

CoRD = *Corpus Resource Database*. Compiled by The Research Unit for Variation, Contacts and Change in English (VARIENG), University of Helsinki. varieng.helsinki.fi/CoRD.

Kaislaniemi, Samuli. 2018. "The *Corpus of Early English Correspondence Extension* (CEECE)". In *Patterns of Change in Eighteenth-century English: A Sociolinguistic Approach*, ed. by Terttu Nevalainen, Minna Palander-Collin & Tanja Säily [Advances in Historical Sociolinguistics 8]. Amsterdam: John Benjamins, pp. 45–59. DOI: www.doi.org/10.1075/ahs.8. OPEN ACCESS

Kaislaniemi, Samuli. 2021. Editions and other sources used in the *Corpora of Early English Correspondence* (CEEC-400). Zenodo. DOI: www.doi.org/10.5281/zenodo.4134471. OPEN ACCESS

Kytö, Merja. 1996. *Manual to the Diachronic Part of the* Helsinki Corpus of English Texts*. Coding Conventions and Lists of Source Texts*. 3rd edition. Helsinki: Department of English, University of Helsinki. Available in *ICAME Corpus Manuals*, korpus.uib.no/icame/manuals.

Nurmi, Arja. 1998. *Manual for the* Corpus of Early English Correspondence Sampler *CEECS*. Helsinki: Department of English, University of Helsinki. Available in *ICAME Corpus Manuals*, korpus.uib.no/icame/manuals.

Raumolin-Brunberg, Helena & Terttu Nevalainen. 2007. "Historical sociolinguistics: The *Corpus of Early English Correspondence*". In Joan C. Beal, Karen P. Corrigan & Hermann L. Moisl (eds.), *Creating and Digitizing Language Corpora*, Vol. 2, *Diachronic Databases*, 148–171. Houndsmills: Palgrave-Macmillan. A pre-print version is available at varieng.helsinki.fi/CoRD/corpora/CEEC/generalintro.html.

Saario, Lassi. Forthcoming. "Conversion of CEEC-400 into XML". *Corpus Resource Database* (CoRD).

Säily, Tanja, Lassi Saario, Terttu Nevalainen & Samuli Kaislaniemi. Forthcoming. "The burden of legacy: Producing the *Tagged Corpus of Early English Correspondence Extension* (TCEECE)". *Research in Corpus Linguistics*, special issue: *Challenges in combining structured and unstructured data in corpus development.* doi.org/10.32714/ricl. OPEN ACCESS

Contact

Prof. Tanja Säily, University of Helsinki. orcid.org/0000-0003-4407-8929

Dr Samuli Kaislaniemi, University of Eastern Finland. orcid.org/0000-0002-3596-1341

2. Structure and contents of the CEECES

The corpus text files in the CEECES 1 come in two formats: plain text and XML. The plain text files each consist of a single letter, with the letter ID as part of the file name. Each XML file contains one collection (see below), and the file name includes a short version of the collection name.

The CEEC-400 was designed for historical sociolinguistics, and consists of not only the corpus texts, but also of a database containing further information on the letters, as well as data on the social backgrounds of the correspondents. Some of this information can be found encoded in the letter texts, but most of it is held in a separate database file.

The supporting database for the CEECES 1 contains information on the gender, age and social status of the writers, as well as known details about their regional origins and their formal education. The database contains similar 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) and of course the year of writing and word count. See Raumolin-Brunberg & Nevalainen (2007) and Kaislaniemi (2018) for more.

The database file comes as a .csv file which is named CEECES1-metadata.csv.

Additionally, the corpus texts contain some of this information in simple text encoding. For historical reasons, the CEEC-400 uses COCOA codes to include contextual information and for text-level encoding. See Raumolin-Brunberg & Nevalainen (2007) and Kaislaniemi (2018) for more; a fuller description of the codes is given in Nurmi (1998) and Kytö (1991).

Some of the code in the CEECES is slightly different from that used in the CEECS. Each letter is preceded by the **text identifier**, consisting of two strings of code:

```
<L WENTWO2_001>
<Q A 1705 FN IWENTWORTH>
```

The L-line gives the Letter ID, consisting of (a short version of) the **name** of the collection, and the **number** of the letter within the collection: in this case, the collection is WENTWORTH 2 and this is the first letter in the collection. This is followed by the Q-line, which gives 1) the **authenticity** of the letter (A = autograph), 2) the **year** of writing (1705), 3) the **relationship** between the writer and the recipient (close or distant; FN = nuclear family), and finally 4) the **writer code**, i.e. the identifier of the writer (IWENTWORTH). (The writer's name is given in expanded form in the code on the following line in the letter header).

For the XML version, the COCOA codes have been converted into TEI-compliant XML tags. A description of the XML code as well as the conversion process is given in Saario (forthcoming) (see also Säily et al., forthcoming).

Comparison of the CEECES and the CEECE

The *CEECE Sampler* part 1 consists of just under a quarter of the CEECE. Both the CEECE and its sampler span the years 1653–1800, with only a very few letters from before 1680. As seen in Table 1, the ratio of women in the CEECES 1 is slightly lower than in the CEECE, but women still account for over a quarter of the words in the subcorpus.

Table 1: CEECES 1 vs CEECE

	CEECES 1	23 collections			**CEECE**	77 collections		
	TOTAL	**M**	**F**	**F%**	**TOTAL**	**M**	**F**	**F%**
Words	492,711	366,600	126,111	26%	2.2m	1.62m	0.6m	27%
Letters	1,172	915	257	22%	4,923	3,681	1,242	25%
Writers	83	61	22	27%	308	214	94	31%

As seen in Figure 1, the number of letters in the CEECES 1 plotted over time in part mirrors that of the CEECE: there are plenty of letters from the end of the 17th century, and there is a dip in the 1720s. Aside from another dip in the 1740s, the number of letters in the CEECES 1 stays roughly the same for each decade through the eighteenth century.

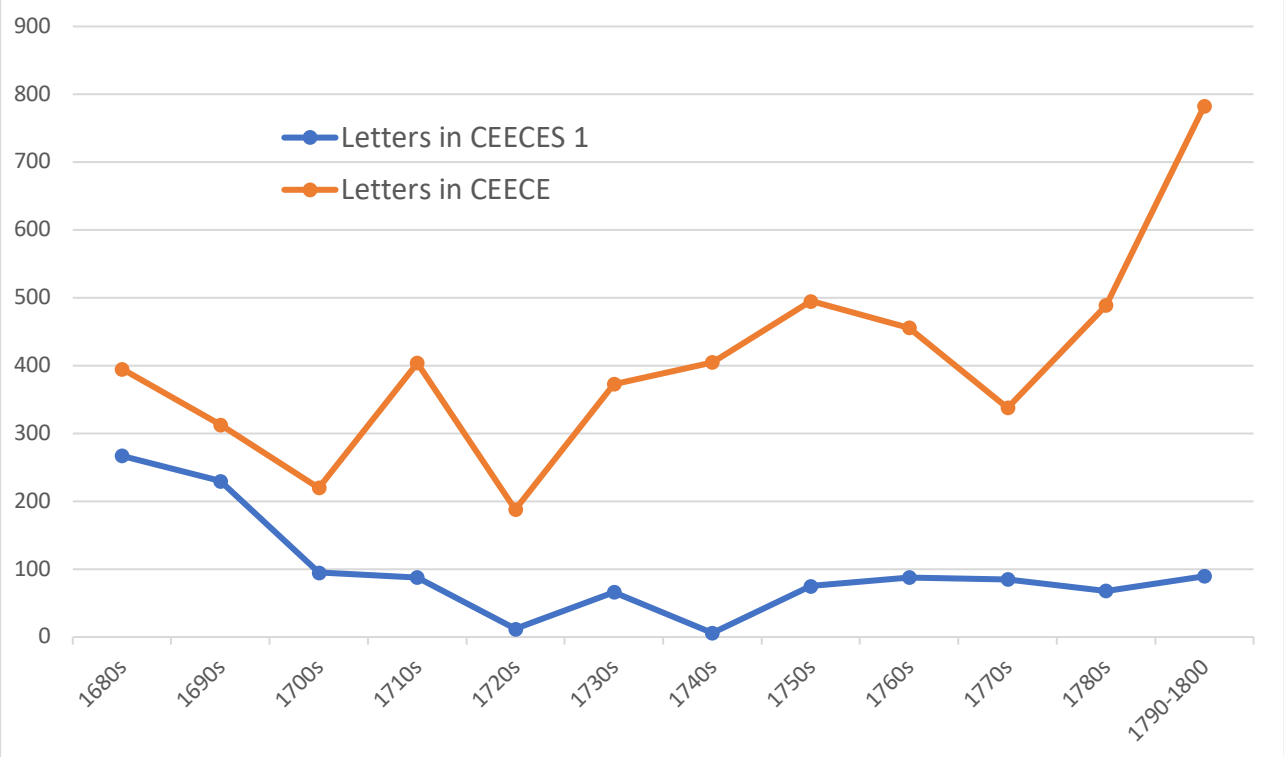

Figure 1: Number of letters in the CEECES 1 vs the CEECE

List of collections in the CEECES

The CEECES 1 contains 23 collections. These in turn are formed from 22.5 collections taken from the CEECE. Collections are sometimes compiled from more than one source: in the case of the collection HURD, one of the two sources could not be included in the CEECES 1 for copyright reasons.

Table 2 below lists the collections in the CEECES 1, giving the time span of the letters in each collection, as well as the number of writers, letters and words in each collection. Following that, Table 3 provides the same figures divided by gender for those collections in the CEECES 1 with letters written by women.

See the accompanying csv. file for more information on the social breakdown of the letters in the CEECES 1.

For a bibliography of the sources of the CEECES 1 (and all the CEEC-400), see Kaislaniemi (2021).

Table 2: List of collections in the CEECES 1

Collection	Years	Writers	Letters	Words
BOWREY	1687-1708	6	38	19,229
CHAMPION	1774-1776	1	14	10,790
CRISP	1779-1782	1	22	18,389
FLEMING 2	1653-1701	11	248	76,297
FLEMING EXTRA	1684-1698	1	52	14,020
GEORGE 3	1765-1783	1	36	7,765
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 (sample 2 only)	1764-1797	2	17	9,464
LENNOX	1761-1800	2	85	66,358
ORIGINAL 4	1682-1716	4	12	2,900
PETTY 2	1682-1687	2	36	14,378
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
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**	**83**	**1,172**	**492,711**

Table 3: The CEECES 1 collections with women writers

Collection	**Years**	**Writers**		**Letters**		**Words**	
		M	**F**	**M**	**F**	**M**	**F**
GIFFARD 2	1697-1722?	0	2	0	16	0	9,701
GOWER	1783-1800	3	4	20	19	8,477	8,512
HATTON 2	1682-1704	6	5	62	16	20,576	4,999
LENNOX	1761-1800	1	1	5	80	681	65,677
ORIGINAL 4	1682-1716	3	1	11	1	2,782	118
PITT 2	1754	1	1	13	11	8,163	7,455
ROYAL 4	1681?-1683?	0	2	0	19	0	4,408
TIXALL 2	1684-1686	0	1	0	2	0	392
WEDGWOOD	1763-1793	4	2	86	2	34,807	425
WENTWORTH 2	1705-1739	6	3	89	91	37,779	24,424
TOTAL			**22**		**257**		**126,111**