CHAPTER 4

The *Corpus of Early English Correspondence Extension* (CEECE)

Samuli Kaislaniemi

4.1 The CEEC project and the CEEC family of corpora

The *Corpus of Early English Correspondence Extension*, or CEECE (pronounced 'cease'), expands the CEEC family of corpora to cover the eighteenth century from 1681 to 1800. The CEECE was compiled with two aims in mind: First, the idea was to provide diachronic continuity to the first part of the CEEC (finished in 1998, called CEEC-1998), in order to follow through the linguistic changes that were still ongoing in 1681, where the CEEC-1998 terminates (Nevalainen & Raumolin-Brunberg 2003); and second, to track linguistic change in the long eighteenth century. The CEEC-1998 and CEECE allow tracking linguistic features over nearly four centuries from Late Middle English into Late Modern English (the entire CEEC family is consequently referred to as CEEC-400; see Kaislaniemi 2006; Nevala & Nurmi 2013).

The compilation of the CEECE followed the completion of the CEEC-1998, which has its roots in the compilation of the Early Modern sections of the *Helsinki Corpus of English Texts* (or *Helsinki Corpus* for short; HC). After the completion and release of the HC, the compilers of the Early Modern section, Terttu Nevalainen and Helena Raumolin-Brunberg, decided to pursue investigating language change in Early Modern English in more detail. To this end, they would compile a separate corpus of Early Modern English; and due to their interest in testing the applicability of modern sociolinguistic methods on historical materials, they decided to compile the corpus from personal correspondence. The project was initiated in 1993, and with the help of student assistants, the corpus was finished by 1998. The CEEC-1998 covers the period from about 1410 to 1681 and contains some 2.6 million words in nearly 6,000 letters from almost 800 informants. Due to copyright reasons (the CEEC corpora are compiled from published editions) the CEEC-1998 could not be released without acquiring permissions from the publishers. Unfortunately, not all publishers granted permission to publish their texts as part of the corpus, and the version released publicly in 2006, the *Parsed Corpus of Early English Correspondence*

DOI 10.1075/ahs.8.04kai

(PCEEC), contains some 2.2 million words (5,000 letters from 650 writers; see the CEEC entry in CoRD for various CEEC corpora).[1]

From around the time of completion of CEEC-1998, the CEEC team had begun to think about extending the letter corpus to cover the eighteenth century. At the time, the only corpus projects containing eighteenth-century correspondence were the *Corpus of late 18c Prose* (CLEP) being compiled at the University of Manchester under David Denison (see the CLEP entry in CoRD), and the ARCHER corpus (version 1b: see the ARCHER entry in CoRD) – but the former was to cover only the years 1761–1790, and the latter had only just under 29,000 words of LModE letters. In the light of the results coming from studies using the CEEC-1998, it was clear that there was a call for a corpus of eighteenth-century letters (Kytö & Rissanen 1997: 14).

We had several expectations regarding letters from the eighteenth century. Literacy increased dramatically during the period, and we expected this to be reflected in an increase in the amount of material from the lowest social ranks – poorly represented in the CEEC-1998 – and hoped to get ample evidence of language change from below. For much the same reasons, we expected to find editions of 18th-century correspondence to contain more letters from women. And finally, we expected to see the effects of the proliferation of prescriptive grammars in the personal letters of the period.

4.2 Corpus compilation

The CEEC *Extension* is based on the same principles as CEEC-1998. Both corpora were designed for historical sociolinguistics: for the application of sociolinguistic methodology on historical texts. For this purpose, correspondence is a useful data source because, firstly, language change from below first appears in written form in private writings (ego-documents), such as personal letters, and secondly, personal letters are arguably the text type closest to speech (Biber 1995; see also Byrne 1964; Culpeper & Kytö 2010). Further, personal letters are eminently suited for sociolinguistics as by definition they (tend to) contain information about the language users and the usage context (names, dates, relationships, places, etc.). And more than anything, personal letters allowed us to access the language of individual writers in their daily lives.

---

1. The part-of-speech tagging of the PCEEC was carried out by Arja Nurmi (University of Helsinki) and the syntactic annotation by Ann Taylor (University of York). The corpus is distributed through the Oxford Text Archive (OTA).

CoRD is the acronym for the Corpus Resource Database: www.helsinki.fi/varieng/CoRD/.

In terms of coverage – whose letters to include, and how many? – like CEEC-1998, the CEECE consists of "judgement samples selected on the basis of extralinguistic criteria" (Nevalainen & Raumolin-Brunberg 1996: 41). The aim of all the CEEC corpora is at social representativeness – while acknowledging the fact of "bad data" (Labov 1994: 11) and the underrepresentation of the lower classes. But within possibilities, the CEEC corpora have been compiled aiming at a balanced corpus, in which there is equivalent coverage of all social parameters (Leech 1993: 13). The social parameters are explained below.

The CEECE contains 77 collections, compiled from 81 different sources (see the Appendix for a list of the corpus collections with their word counts and source details). The compilation process was begun in 2000, and effectively finished in 2006. The external databases have been further augmented since that time.

4.3 Coverage (representativeness and balance)[2]

The CEECE contains writers from all social parameters employed in CEEC corpora. In terms of balance it is thicker at the end of the period covered, and its regional coverage is not quite satisfactory. However, as it is compiled according to a model designed for quantitative sampling of Early Modern English society, it is worth remembering that 18th-century England was in some ways remote from what in essence was a post-medieval world. In other words, some of the inherited categories are not necessarily relevant for making sense of 18th-century England and Late Modern English. For instance, East Anglia was highly important in the Tudor period, but by the Georgian era its cultural and political relevance had waned. Similarly, the clergy were not nearly as important in 18th-century England as they had been in the 16th century, thanks to advances in literacy and the availability of the Bible in English. So, the fact that a high proportion of CEECE data comes from the professional classes and from Londoners is rather a reflection of changes in English society than of skewed data. Perhaps the greatest strength of the CEECE is its gender division. More than a quarter of the words in CEECE are by female informants, up from c. 17% in CEEC-1998.

2. This section is by Samuli Kaislaniemi and Mikko Hakala.

4.3.1 Diachronic and quantitative coverage

The CEEC Extension covers the long 18th century from 1681, where CEEC-1998 ends, to 1800 (inclusive). It contains 4,923 letters by 308 writers, amounting to just over 2.2 million words.[3] The distribution of data is not equal: as can be seen in Figure 4.2 below, there is a dip in the word count in the period 1720–1739; and a full third of the material falls in the last period (1780–1800). Although initially the aim was to choose 10 letters per writer (of 'average' length, i.e. excluding very short or long letters), this figure was soon expanded – in part due to abundant sources, in part in order to include good coverage of writers whose letters span a long period of time. Increasing sample size was also one way to attempt at variety in addressees of a single writer, to allow for individual variation according to recipient; and also attempt was made to include letters between two writers, in both directions.

In some cases, it was felt reasonable to include writers from whom there were less than the ideal 10 letters available. These concern primarily writers from the lower social ranks, and also women. Particularly good examples of collections with letters from the lower social ranks include Clift and Pauper (see the Appendix for details), both of which contain letters from poorly literate writers.

4.3.2 Gender balance

One of the aims of CEECE, as noted above, was to compile a corpus in which the proportion of female informants was higher than in the CEEC-1998. Figure 4.1 tracks the number of male and female informants for the years 1680–1800 in 20-year periods. Figure 4.2 shows the number of words in our data set divided into 20-year periods.

---

3. The studies in this volume focus on the period 1680–1800 and the CEECE provides the vast majority of the material studied. There is, however, some overlap between the original CEEC-1998 and the Extension. CEEC-1998 contains altogether 29,199 words written between the years 1680 and 1681, and CEECE 31,600 words from 1653–1679. For the sake of accuracy, only material dating after 1679 was included in our data set, which in all comprises 4,945 letters by 315 writers, coming to 2,216,119 words. To give an accurate characterization of the material used for the studies in this volume, the tables and graphs in this chapter are based on this latter data set (all letters in CEEC-400 spanning 1680–1800) rather than on the CEECE proper. However, as more than 98% of the material is from the CEECE, the figures may also be taken as a fairly accurate description of the structure of the CEECE. See the Appendix for more detailed information on the collections included in CEEC-1998 and CEECE.

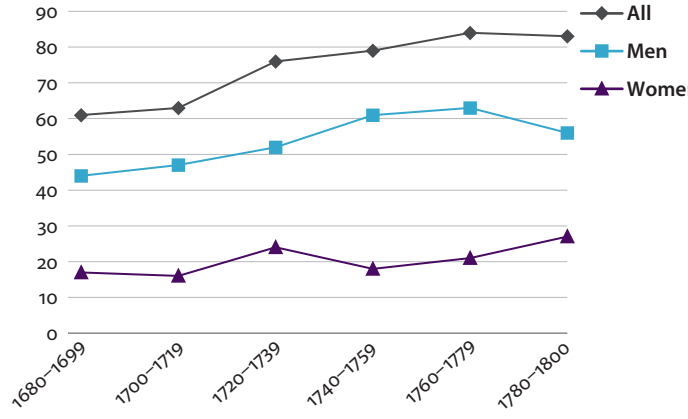

Figure 4.1 Number of informants per period: gender division (absolute frequencies)

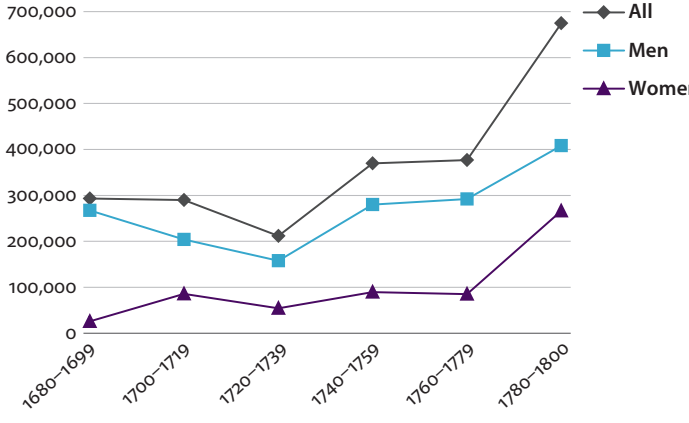

Figure 4.2 Number of words per period: gender division (absolute frequencies)

Apart from a minor slump in 1720–1739, there is relatively little fluctuation between 1680 and 1779. The last period reflects a major increase in the overall availability of data. One major component of the last period are the family letters of the kings George III and IV. The first Hanoverian monarchs were not native English speakers, but from George III their letters are included in the corpus, and happily the extensive family correspondence of the House of Hanover has been published and could be included in the corpus (George 3a, George 4; see the Appendix).

As seen in Figure 4.3, there are no dramatic changes in the proportion of female informants, which remains at around 30% over the entire 120-year period. Figure 4.3 reveals that the proportion of data from women is mostly between 20–30%, but the first and last period, 1680–1699 and 1780–1800, are exceptions. For the first period, for some reason there are simply less letters per writer in the source editions. For the last period, the rise in material is due in part to royal family letters, which include letters from four princesses as part of the family correspondence of the House of Hanover. The word count is also bolstered by the letters of writers such as Fanny Burney, Hester Piozzi (Mrs Thrale) and Sarah Duchess Lennox, and also by those of Jane Austen and Mary Wollstonecraft.

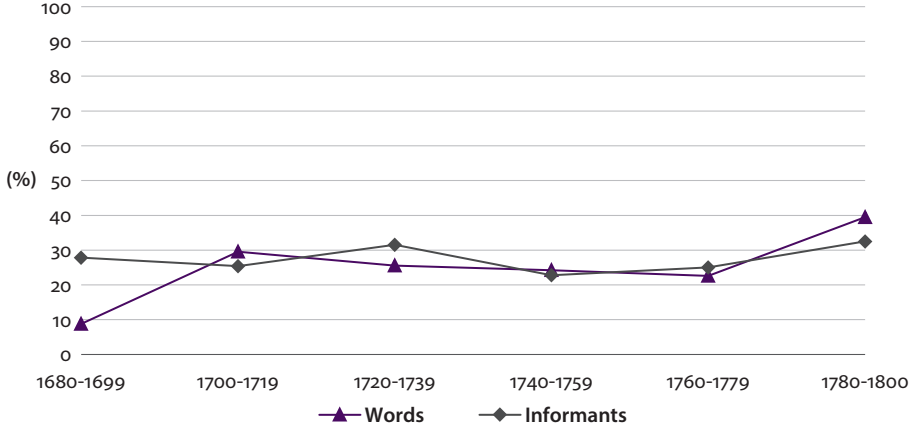

Figure 4.3 Proportion of running words and female informants per period

4.3.3 Social ranks

Table 4.1 provides a breakdown, based on running words, of the data according to the social status of the writer.

English society in the 18th century was greatly different from previous centuries; therefore the categories are not directly equivalent with those in CEEC-1998. In particular, the development of the wealth and (political) power of the middling ranks – merchants and professionals – makes them a highly influential part of society in the eighteenth century (see further 2.4). Their rise to prominence is particularly reflected in their share of the data – a full third of the words are by professional or merchant writers (even though the ratio of merchants is much smaller than in CEEC-1998, cf. Nevalainen & Raumolin-Brunberg 2003: 46).

Table 4.1 Social status of informants, percentage of running words per gender

	Men	Women	Total
Royalty	7%	6%	6%
Nobility	6%	34%	14%
Gentry (Upper)	9%	3%	7%
Gentry (Lower)	21%	23%	21%
Clergy (Upper)	6%	0%	5%
Clergy (Lower)	14%	10%	13%
Professional	30%	21%	27%
Merchant	3%	1%	3%
Other	4%	4%	4%
Total	100%	100%	100%

In general, there is more data from the upper than the lower ranks, but the social coverage of men's letters is somewhat more even than women's. This is an artefact of the sources, but also of English 18th-century society: on the one hand, most letter editions publish the correspondence of the elite; on the other, literacy strongly correlated with social rank, and the spread of literacy down the social ladder was slow. Figure 4.4 shows the distribution of the different social status groups over time in CEECE.

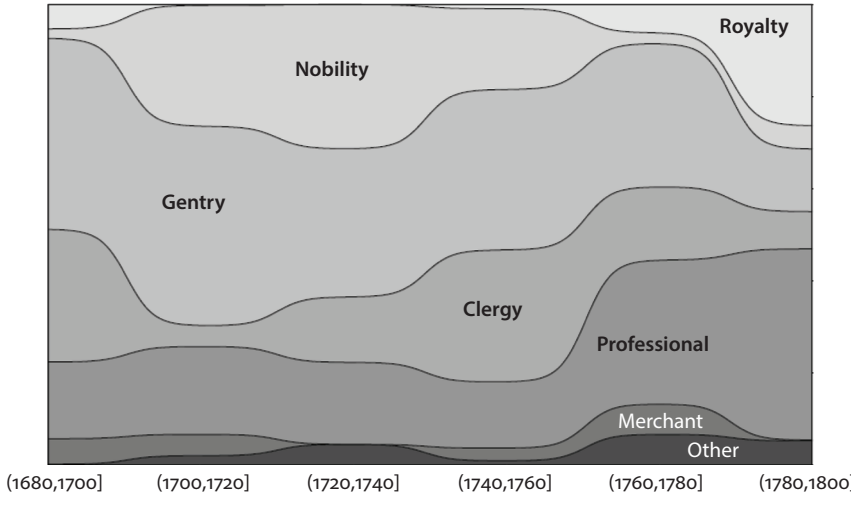

Figure 4.4 Density plot of the distribution of letters by social rank (courtesy of Harri Siirtola)

Figure 4.5 shows the proportion of data from the highest and lowest ranks in 20-year periods – royalty and the nobility on the one hand, and non-gentry on the other. There is a small increase in the proportion of data from the lowest ranks in the last two subperiods, but otherwise their proportion remains quite low throughout. There is some fluctuation in the amount of data from the highest ranks, the most notable feature in Figure 4.5 being the spike in 1720–1729. This however is an artefact caused in part by the slump in the overall amount of data (Figure 4.2), coinciding with a period where there are more letters from the nobility than from other social ranks.[4] The rise in the last period, on the other hand, is explained by the inclusion of royal family letters, as explained above.

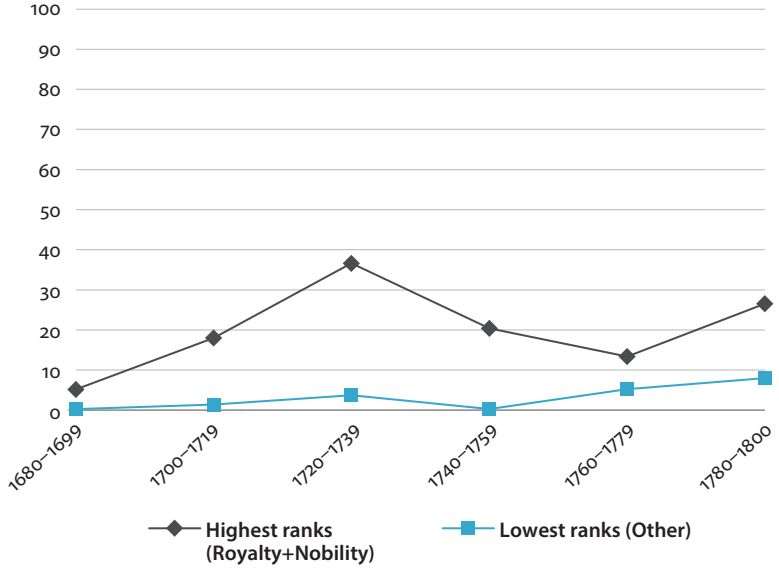

Figure 4.5 Proportion of running words, highest and lowest ranks

4.3.4 Regional coverage

The CEECE follows the same division into regions as is used in CEEC-1998. The encoded regions are: the North, East Anglia, London, the Court, the Home Counties, Other Areas in England, and Abroad (see Figure 4.6). Key characteristics of the regional coverage of CEECE are the rise of the proportions of informants living in London or so-called Other Areas. Between 1650 and 1800, London more than

4. To illustrate the skew another way, the word counts of letters from nobility for the periods 1720–1739 and 1740–1759 are the same (just over 81,000).

doubled in size, and the proportion of the population living in London grew from around 8% to 12% (Boulton 2000: 316; Schwartz 2000: 648; Wrigley & Schofield 1981: 207; see 2.3, above). This increased importance of London is reflected in the contents of the CEECE. At the same time, increased regional mobility makes it hard to pinpoint where some informants come from. This uncertainty is on the other hand counterbalanced in CEECE by registering domicile at a county level.

Figure 4.6 Map of regions covered by CEECE (1 North, 2 East Anglia, 3 London, 4 Court, 5 Home Counties, 6 Other areas)

Table 4.2 gives the regional distribution of the data according to the informants' domiciles. The most striking features here are the dominance of London and the dearth of data from East Anglia. The former is closely linked with increased domestic migration and the rise of the professional classes. The lack of data from East Anglia, on the other hand, is a reflection of what has been edited – it turned out that there are very few editions of letters from eighteenth-century East Anglian writers.

Table 4.2 Regional distribution, percentage of running words per gender

	Men	Women	All
The Court	5%	3%	5%
London	38%	58%	43%
East Anglia	1%	0%	1%
Home Counties	17%	12%	16%
North	10%	1%	8%
Other areas	25%	20%	23%
Abroad	4%	5%	4%
Total	100%	100%	100%

4.4 Coding

The coding of CEECE has followed that of CEEC-1998 (Nurmi 1998), which was derived from the coding used in the *Helsinki Corpus* (Kytö 1996).

There are two kinds of coding: text-level coding, and parameter coding. The former includes information contained in the source edition, such as special characters and textual commentary, but also information encoded by the CEECE compilers, such as omitted passages and foreign language strings. Parameter coding on the other hand contains extratextual information, including the file name and bibliographical information, page numbers, and importantly, information on the letters and their writers.

Each letter is preceded by a letter header consisting of four encoded lines. The first L-line contains the unique **letter ID**. The second is the Q-line: it contains a code letter indicating the **authenticity** of the letter, the **year** of writing, a code letter indicating the **relationship** between the writer of the letter and the addressee, and then the **correspondent code** – a unique identifier of the writer of the letter. The third line is the X-line, which contains only the **name of the writer** in human-readable form, as opposed to the opaque correspondent code. The fourth and final line is the P-line, containing the number of the **page** on which the text begins in the source edition (further P-tags are also inserted in the text at every page break).

For more detailed information on the coding used in CEEC-400, see Nurmi (1998) and Raumolin-Brunberg & Nevalainen (2007).

4.4.1 Letter quality

Since the CEECE was designed to gain access to the language of individuals, letter quality was a primary concern. Authorship is the most critically relevant feature here: only letters that were demonstrably written by an identifiable individual are included in the corpus. Scribal letters and copy-letters are marked as such. Linked to this, the extent to which a writer's social background was recoverable was important. And thirdly, only editions that retained original spelling were used – although this in practice meant including works which have normalized u/v-variation ("vp" > "up", "ouer" > "over"), expanded abbreviations ("w^{ch}" > "which", "y^{e}" > "the"), and modernized capitalisation and punctuation. Because different editions have followed differing practices, and furthermore as nearly all editions have to some extent normalized spelling, as a rule, the CEEC corpora cannot be used to study orthographical features.

These three aspects have been combined to give each letter a rating of **authenticity**. The authenticity codes are:

A = autograph letter in a good original-spelling edition; writer's social background recoverable
B = autograph letter in a good original-spelling edition; part of the writer's background information missing
C = non-autograph letter (secretarial work or copy) in a good original-spelling edition; writer's social background recoverable
D = doubtful or uncertain authorship; problems with the edition, the writer's background information, or both.

In practice, the proportion of non-A letters in CEECE is so small that authenticity does not play an important role in analysing results of corpus queries. More than 96% of the letters in CEECE are quality A.

4.4.2 Relationship between writer and recipient (register)

Social relations between correspondents are encoded into the corpus texts. Addressee relations allow tracking register variation, as distance and power relations manifest in the language of the letters (see Nevalainen & Raumolin-Brunberg 2003: 189–191 and 191ff.; also Nevala 2004a (reprinted in 2004b)). There are five basic relationships:

FN = Nuclear family members (close)
FO = Other family members (more distant)
FS = Family servants (distant)
TC = Close friends (close)
T = All other relationships (not close)

Establishing these categories is fairly straightforward, although on occasion it can be difficult to identify close friendships, and sometimes determining kinship can require some genealogical archaeology. The distribution of the letters in CEECE according to relationship of writer and recipient is shown in Table 4.3.

Table 4.3 Relationship of writer to recipient, percentage of letters per gender

	Men	Women	All
Nuclear family (FN)	27.6%	45.1%	32.0%
Other family member (FO)	8.5%	11.1%	9.2%
Family servant (FS)	0.2%		0.1%
Other (T)	36.2%	21.1%	32.4%
Close friend (TC)	27.4%	22.7%	26.2%
Total	100%	100%	100%

As can be seen, there are proportionally more family letters from women than from men in CEECE. This is because relatively more family letters survive from women than from men, which in turn is a reflection of the patriarchal structure of society as well as of literacy rates (for women letter-writers, see Daybell 2001, 2006).

4.5 Corpus formats and external databases

What makes the CEEC corpora stand apart from other historical corpora are their supporting databases: the **letter database** contains information regarding each individual letter, and the **correspondent database** contains information on each writer and recipient.

The letter-specific information includes information on the textual source (bibliographical details of source edition, page numbers, letter number in edition if any), details on the contents of the letter (date and place of writing, content type, word count), and brief information on the writer and recipient (unique correspondent IDs, names, social status and rank at the time of the letter, and the relationship of the writer and addressee). We have also recorded information of particular linguistic relevance (letter authenticity; presence of opening and closing formulas), and

metadata relating to the corpus itself (unique letter ID, collection name, copyright permissions status, state of completion of database entry, last update). Most of this information is stored in free text fields, but some (such as social rank of correspondents, their relationship, and letter authenticity) is recorded as code.

The correspondent database contains more in-depth information on each correspondent. Although it was begun as a sender database (e.g. Raumolin-Brunberg & Nevalainen 2007: 160–164), it has been updated to include recipients as well, so that it is now possible to find all letters in CEEC-400 written to recipients matching given parameters. Much of the information in the correspondent database is stored in both codes and free text fields to facilitate quick searches on the one hand, but also to retain more detailed information that is difficult to codify, such as specifics relating to the education, migration or career of any individual.

For reasons primarily to do with the state of computing at the time the CEEC project was planned (the early-to-mid 1990s), the sociohistorical information required in this kind of a sociolinguistic corpus was entered into external databases rather than encoded into the corpus texts. But in 2007, the CEEC-400 was uploaded into a relational database on an online server, accessed by a bespoke search tool called CEECer (pronounced 'seeker'), used with a web browser. This integration now allows us to search the corpus using as query parameters any information in the databases, greatly enhancing the process of performing corpus searches. In essence, with CEECer we can create subcorpora from CEEC-400, which can then be extracted and queried in a concordancer or other corpus tool.

The current base format of the corpus is no longer by collection or even by writer, but by individual letter. The base texts remain ASCII plain text files (a format inherited from HC), but these have been supplemented with the POS-tagged and syntactically parsed texts of PCEEC, and also with the standardised-spelling version of the corpus (see Palander-Collin & Hakala 2011). All these versions have been stored in CEECer.

4.6 Copyright and publication

The publication of a corpus built from previously printed sources requires acquiring copyright permissions from the editors and publishers of the source texts. Of course, this is an old concern, voiced for instance by Kytö & Rissanen (1997: 18–19). The process for acquiring permissions to publish the CEEC-1998 corpus was begun in the early 2000s. This turned out to be more difficult than envisaged, and for some texts the permission was not received, as noted above.

When the copyright-clearing process was started for CEECE in 2010, we expected similar difficulties. Yet over a few years, publishers' attitudes to corpus compilation and similar work had changed greatly. The launch of mass digitization projects like Google Books (in 2004) had raised publishers' awareness of the power of digitization, and resulted in what can only be called confusion and panicked protectionism. In sum, even publishers that had formerly kindly allowed using texts published in their books in linguistic corpora, now were reluctant to grant permission, or denied it outright. As of late 2017, we have not received permission for some 2/3 of the texts in CEECE, and publication of the corpus remains uncertain.

INFOBOX 1

Data retrieval

Mikko Hakala

This section provides a short description of the data retrieval process for each study in this volume. Although the search parameters were specific to individual studies, the retrieval process usually followed the same basic outline: a research assistant first performed the corpus queries and sorted the results and each researcher then checked and further pruned their own data. The main tools used in retrieving the data were WordCruncher 4.30 and WordSmith 5.0, with the former mostly preferred with single-word search terms and the latter with search terms containing multiple words or wildcards.

Associating the results with relevant sociolinguistic metadata is an important consideration in our research. Both programs provide significant help with this task; WordCruncher automatically retrieves letter-specific ID-tags from the corpus files and adds them to the extracted data, while WordSmith can be instructed to add the same ID-tags with a separate tag file. Biographical information about the authors and the recipients can then be extracted from a separate database and associated with examples on the basis of the ID-tags. As spelling variation is still common in 18th-century letters, dictionaries (mainly the OED and the MED) and wordlists generated from the corpus (using WordSmith and WordCruncher) were consulted to cover the different spellings of each search term as thoroughly as possible. All results were extracted from the concordance programs with a context of five lines of text before and after each search term. The results were always checked manually for false hits.

Data retrieval for the studies dealing with individual lexemes was relatively straightforward. In her chapter, Nevala focuses on the personal pronouns *my*, *mine*, *thou*, *thee*, *thy*, *thine*, *you*, *your*, and *yours*. All spelling variants of these pronouns were searched for using WordCruncher. For Nevalainen's study of the third-person indicative suffixes -(*e*)*th* and -(*e*)*s*, the relevant forms of the verbs *do*, *have* and *say* were searched for. And for Laitinen's study, the search terms were the indefinite pronouns -*one,* -*body* and -*man*. For Nurmi's study on *do*, the various forms (*do*, *did*, *don't*, *didn't*) and their spelling variants were retrieved from the data. The instances of periphrastic *do* in affirmative statements were then identified and manually sorted.

Palander-Collin's chapter deals with *its*, *of it*, *of the same*, and *thereof*, all instances of which were retrieved using WordSmith. For the search terms *of it*, *of the same* and *thereof*, the search parameters were further specified to include a definite article to the left of the target word/phrase and up to six words intervening (e.g. ***the*** *sulliage and infection* ***of it***), which had proved sufficient during trials.

For Sairio's study of the progressive aspect, all instances of BE + -*ing* were extracted from the corpus with, on the basis of trials, up to four words intervening. Forms of the present participle incorporating the clitic *a* (e.g. *a going*) were taken into account. Different spelling variants of both the verb and the past participle were searched for using wildcards in WordSmith.

Säily's study deals with the derivative suffixes -*ness* and -*ity*. The data for her study were retrieved with WordSmith by using a search list of potential spelling variants, including plural forms, for both suffixes. The search list was created on the basis of previous research, the OED and close readings of the corpus. The results were connected to the relevant sociolinguistic metadata with Python scripts.

Acknowledgments

For a more detailed overview of the history of the CEEC project, see Nevalainen & Raumolin-Brunberg (1994a; 1996 (eds.); 2003), Raumolin-Brunberg & Nevalainen (2007), and the CEEC entry in CoRD. The premises of the CEEC *Extension* were first outlined in Laitinen (2002).