© 2002 Mikko Laitinen

Extending the *Corpus Of Early English Correspondence* to the 18th Century

Mikko Laitinen

Abstract. The *Corpus of Early English Correspondence* (CEEC), stretching over almost 300 years of personal letters and including 2.7 million words, is now being expanded. This diachronic corpus that covers the period between 1417 and 1681 was designed by the *Sociolinguistics and Language History* project at the Department of English of the University of Helsinki. Currently, work is under way to compile a separate collection to the CEEC to cover the 18th century as well as the last two decades of the 17th. In addition, a third type of compilation work, which will include new valuable material not incorporated in the CEEC for various reasons, is being carried out to further improve the corpus.

This article concentrates on three issues. Firstly, I will report on the work the *Sociolinguistics and Language History* group has done so far in enlarging the CEEC, especially in expanding the corpus of personal letters chronologically. In addition, I will specify the terminology that we use of the different versions of it. Lastly, both the parameter and text level coding used in the corpus are discussed.

1. Introduction

The *Corpus of Early English Correspondence* is the product of the work of the Sociolinguistics and Language History research group at the Department of English at the University of Helsinki. The project was initiated in 1993, when the compilation process was begun, the work being funded by the Academy of Finland (1993-1995) and the University of Helsinki (1996-1998). The main aim of the group was to test the applicability of modern sociolinguistic methods to historical language data and, since there was no sufficient material available at the time, it was decided that a new purpose-built corpus was needed to facilitate this new kind of research (see Nevalainen & Raumolin-Brunberg 1994).

The work of the group has been going on for roughly eight years, under the guidance of Professor Terttu Nevalainen and Doctor Helena Raumolin-Brunberg. Two other original members of the project are Doctors Arja Nurmi and Minna Palander-Collin, and there have been a number of assistants helping with the work.[1]

The nature of the compilation process has meant that there are various versions of the corpus that have been used for different studies. I will concentrate here on the two main ones, since they show quite noticeably how the work has proceeded.

One of the earliest versions of the CEEC in extensive use has been the 1996 version. At this point the corpus included personal letters from 67 collections and roughly 2.4 million words covering the time period starting from circa 1417 and ending in 1681. A collection of pilot

[1] Ms Maarit Alanko, Ms Kirsi Heikkonen, Mr Jukka Keränen, Ms Minna Nevala helped in the compilation process.

studies can be found in Nevalainen & Raumolin-Brunberg (eds) (1996), whose Appendix also contains a complete list of the collections included in that version of the corpus.

The second major part of the compilation project was concluded with the completion of the 1998 version of the corpus. At the same time the *Corpus of Early English Correspondence Sampler* (CEECS), consisting of roughly half a million words no longer in copyright, was made available on the second ICAME CD-ROM (Bergen, The HIT Centre, 1999 (see Nurmi 1999b)). By 1998 the CEEC was somewhat larger, containing 2.7 million words in 96 different letter collections. The list of the collections and their word counts can be found in two doctoral dissertations based on the data drawn from the CEEC (see Nurmi 1999a and Palander-Collin 1999). The 1998 CEEC contains around 6,000 letters from almost 800 informants, of these roughly 20 per cent are by women. In addition, a separate sender database has been constructed for the corpus, which offers researchers easy access to various sociolinguistic variables, including the writer's provenance, social status, gender, education, age and the relationship with the addressee (see Raumolin-Brunberg 1997). It goes without saying that these are of paramount importance in facilitating historical sociolinguistic research, which is one of the objectives of the research group.

It has been decided that no new material will be added to the CEEC of 1998, since it forms a balanced corpus that can be neatly divided into two parts, both covering chronologically fairly equal periods: the first from c. 1417 to 1550 and the second from 1551 to 1680. All this means that whenever the *Corpus of Early English Correspondence* is mentioned in publications after 1998, it means the 1998 version.

2. Enlarging the CEEC

In 2000 the Sociolinguistics and Language History project group began a new project related to the CEEC aiming at improving it. The work is being done at the *Research Unit for Variation and Change in English (VARIENG)* at the University of Helsinki, the unit being one of the National Centres of Excellence funded by the Academy of Finland. Currently there are seven members of the project: Professor Terttu Nevalainen, Doctors Helena Raumolin-Brunberg, Arja Nurmi, Minna Palander-Collin, and Minna Nevala, Mikko Laitinen, Emma Murros and Anni Vuorinen.

The work-in-progress falls under two separate projects. Firstly, we are expanding the CEEC to cover also 18th-century personal correspondence. This means that we will add personal letters written between 1682 and 1800 to the corpus. In the future this will make it possible to study variation and change over roughly four centuries of English which is as close to the oral registers of language as one can reach in the written medium. In addition to expanding the corpus, our work also means that some of the existing collections of the CEEC will be augmented, since we are now able to include material from some of the published collections of correspondence used for the CEEC that were beyond the original termination year of 1681. Since the CEEC will remain as it is, the material of the first project will be called the *Corpus of Early English Correspondence Extension.*

Another way we will expand the CEEC is by adding material that does not meet the strict sampling criteria that were set for the corpus. As the editorial principles of published collections of correspondence have varied over the centuries, there are a few editions that will be of use but for various reasons were rejected previously. A separate collection called the *Corpus of Early English Correspondence Supplement* will therefore be created in order to

accommodate this type of material. The Supplement will also include material from new editions published after 1998. In the following I will define both the Extension and the Supplement and provide an overview of the material in both.

3. CEEC Extension

The compilation process of the Extension will follow the same principles as for the CEEC. This means that we are primarily concerned with the authenticity of our material. By the nature of the project we must mainly rely on edited collections of correspondence. In some cases it is also possible to edit letters from manuscripts if we feel that the material provided in them could prove to be especially valuable, but because of both time constraints and our research interests, we try to keep the editorial duties to the minimum. An ideal case for us is to find editions that include letters that were actually sent from one person to another with precise identification of the people and their social backgrounds. While ideal collections are fairly rare, they certainly exist. Fortunately for us, some editors have showed outstanding expertise both in historical and in linguistic matters.

Most of the editions that are useful for us, however, contain letters of varying quality. Firstly, there may be autograph letters that were sent and have been found in the addressees' estates, and they naturally are a high priority for us. At this point of the compilation we can safely state that most letters in the Extension are autographs. What comes to draft letters and copies written by the sender, we will treat them as authentic letters (see Nevalainen & Raumolin-Brunberg 1994: 138). If we compare the compilation process of the Extension with that of the CEEC, there are obvious differences in that fewer people in the 18th century relied on secretarial help in their personal writing. It goes without saying that this is because of more widespread literacy in almost all social ranks in England.

Another part we carefully emphasise in the compilation of the CEEC Extension is the original spelling requirement. As the letters have attracted researchers from a variety of fields during the past two centuries, the editorial styles have been rather diverse. Unfortunately, some 19th century editors decided to alter the spelling or in some cases left the editorial principles unexplained. Some minor changes, however, have been allowed. The most frequent include expanded abbreviations and modernized capitalization and punctuation. We will thus follow similar practices to those used in the CEEC, and no modernised versions of the letters of the late 17th and 18th centuries will be included in the Extension.

Similarly, the same criteria for social representativeness as were used in the CEEC will also be applied to the Extension. We aim to provide a **balanced,** purpose-built corpus of personal letters written by both men and women from as broad a social spread as possible, designed specifically with socio-historical methodology in mind. The aim of the team is to include letters written by people from all walks of life not only from the higher ranks of English society. It is also in our interest to try to include as much material from female informants as possible.

At this stage of the compilation process it could be said that there seem to be many published collections of personal correspondence covering the 18th century from both men and women. The ideas of the Enlightment that spread across Europe and England caused that the material from the literary circles and educated men and women of the period is ample. It should be noted, however, that some writers of the period, especially literary men and women, often either addressed their letters to a large circle of people or intended them to be published later.

One of the challenges in the process is therefore to find material that meets the criteria of personal correspondence (see Nevalainen & Raumolin-Brunberg 1996: 40). There are, of course, some difficulties in finding material from the lower end of the social scale, but we expect to overcome them, for example, by searching through the correspondence of servants and the tradesmen employed by business houses and various trading companies.

Table 1. The collections in the CEEC Extension (as of January 2002).

COLLECTION	YEARS	RUNNING WORDS
Austen	1796-1800	27,811
Banks	1704-1760	c. 50,000
Burney	1751-1784	c. 42,500
Burney Fanny	1774?-1800	c. 58,000
Clift	1792-1799	52,038
Darwin	1763-1797	18,948
Defoe	1703-1729?	33,225
Dodsley	1743-1764	52,470
Evelyn	1665-1703	38,929
Fleming2	1653-1701	76,297
Garrick	1733-1777	43,974
Gibbon	1750-1793	27,669
Haddock2	1688-1719	4,647
Hatton2	1682-1704	25,575
Jones	1768-1794?	33,011
Liddell	1709-1716	36,822
Melbourne	1776-1799?	3,767
Montagu	1710-1781	80,395
Pepys2	1681-1692	9,435
Pierce	1751-1771	20,361
Pinney	1679-1706	25,098
Piozzi	1784-1798	40,296
Porter	1789-1800?	13,815
Prideaux2	1681-1722	15,934
Purefoy	1737-1751	30,548
Sancho	1770-1780	c. 30,000
Wenthworth	1705-1739	c. 67,200
Total number of words		**c. 958,765**

Furthermore, the letter selection process has not changed from the CEEC. The aim is still to find at least ten medium-length letters per writer, but in some cases we have had to settle for fewer than that. This has been the case especially with women writers whose material we have tried to include whenever possible. However, we have noticed that as we move on to the 18th century the collections of personal correspondence are rather more abundant than before. In some cases we have therefore included as many as 40-50 letters from one writer. Attention is also being put to the addressees, especially of those writers who are well represented, since we try to include letters intended to various recipients in order to provide material for network studies. Furthermore, the Extension follows the collection format

familiar in the CEEC. Table 1 above shows the collections with their word counts in the Extension at the moment.

As can be seen in Table 1, some of the collections in the Extension overlap with the time frame of the CEEC (1417-1681). For example, the letters of Evelyn, Fleming2 and Pinney contain material that originates prior to 1681, but they all continue beyond that, and therefore the Extension accommodates them.

Moreover, a sender database using a database program will be provided for the Extension, enabling easy access to the sociolinguistic variables that of course form the backbone of historical sociolinguistic research. As mentioned above, the conventions used by the editors have varied considerably. Some editors have included specific information about the senders and addressees; however, this information must often first be checked by comparing the letters themselves or comparing them with the information found in other sources. In other cases, unfortunately for us, the information for the database cannot be found in the editions, but must be sought in other sources.

4. CEEC Supplement

Since there are some gaps in the original time span of the CEEC (see Nevalainen 1997: 84), we always look to fill these in order to provide a chronologically balanced corpus. The material included in the Supplement originates from between 1402 and 1681 and is of two types. Firstly, there is material that meets the sampling criteria set for the CEEC (see Nevalainen & Raumolin-Brunberg 1996) and has become available since 1998, such as the letters of Sir Walter Ralegh. It should be noted that we are constantly in search of editions of high quality in order to improve the CEEC even further. Secondly, the Supplement contains material that was discarded because of the original-spelling requirement of the CEEC (see Nevalainen 1997: 82).

The collections with modernized spelling in the Supplement are otherwise excellent material for historical sociolinguistic research, and can provide researchers with valuable information on the language of women, like the letters of Lady Lisle or Joan and Maria Thynne, and close family members, such as the correspondence between the members of the Symcotts family. It should be noted that this part of the Supplement material is suitable for research on morphosyntactical structures, but not, for obvious reasons, spelling.

Some of the material in the Supplement is also our reserve that we expect to edit ourselves whenever the original letters are located. This was the case in the compilation of the CEEC when some of the members in the project edited (the letters of Lettice Gawdy)[2] or re-edited (those of the Marchall family)[3] various collections from manuscript. Table 2 below shows the collections in the CEEC Supplement at the moment.

[2] Edited from British Library Manuscripts by Minna Nevala.

[3] Re-edited by Jukka Keränen, Terttu Nevalainen and Arja Nurmi.

Table 2. The collections in the CEEC Supplement (as of January 2002).

COLLECTION	YEARS	RUNNING WORDS
Betts	1522-1640	2,624
LisleH	1531-1539	12,552
Plumpton2	1461-1549?	36,432
Ralegh	1581-1618	18,514
Symcotts	1629-1660	13,789
Thynne	1570?-1611	19,574
Zouche	1402-1403	1,206
Total number of words		**c. 104,691**

5. Coding and annotation

The coding used both in the Extension and the Supplement follows the patterns established in the CEEC and the CEECS (see Nurmi 1998 for a more comprehensive discussion on the CEECS). Both are therefore logical complements of the *Corpus of Early English Correspondence*, which facilitates their easy use in conjunction with it to provide further material on the time periods.

The coding has been introduced in some of our earlier articles (see Nurmi 1999b and Keränen 1998), which describe the two levels of coding used in the CEEC and the Sampler. The following is a brief overview of both the textual and parameter coding originally adapted in a slightly modified form from the Helsinki Corpus (see Kytö 1996).

The parameter codes for each collection include the following information: firstly, the text files are identified on the **B-line** that is placed in the beginning of each collection. An example of a B-line is <B FFLEMIN2>, which identifies that particular collection as the letters of the Flemings in Oxford. The name followed by the number 2 in this case shows that the letters of this particular family are already included in the CEEC, and that we have added new material into it. The collection labels are followed by the source information. This has been decided to be done by using the text level coding [^…^], indicating 'our comment'. For example, the source information for the letters of the Flemings in the Extension is shown in (1) below.

(1) [^THE FLEMINGS IN OXFORD BEING DOCUMENTS SELECTED FROM THE RYDAL PAPERS IN ILLUSTRATION OF THE LIVES AND WAYS OF OXFORD MEN 1650-1700. VOLS. II-III. ED. BY MAGRATH, JOHN RICHARD. OXFORD HISTORICAL SOCIETY 62, 79. 1913, 1924.^]

All the letters included in the Extension and the Supplement are coded by using the **Q-line**, which contains the authenticity codes,[4] year of composition, relationship between the writer

[4] There are four classes of authenticity code: A for autograph letters by a person whose background has been identified, B for autographs whose writers' background information is incomplete, C for copies and letters written by secretaries, D for letters whose origin is unknown (see also Nevalainen & Raumolin-Brunberg 1996: 43).

and recipient, and a writer code. The **X-line** indicates the name of the writer in full. Since our corpus covers a long time period, and there were often several holders of a title, for example the Dukes of Devonshire, we have decided to identify the writers only by using their first and last names. The **P-line** includes the page numbers of the letters in the editions. Where there are several editions, the edition number is placed before the page number, followed by a comma (e.g. <P II,285>).

Square brackets [}…}] are used for the 'headings', which often contain an 'editor's comment' [\…\] embedded in them. This type of double coding indicates that two types of information are provided in the line. Example (2) below illustrates the parameter coding used for an autograph letter (A) written in 1764 by Charles Burney to his close family member (FN) Frances (his daughter), whom he calls Fanny. Furthermore, the P-line shows that it can be found on page 44 in the source, and the editor has included a heading stating the addressee of the letter and its date of composition. In this particular case Charles Burney continued the letter a few days after he began writing it, and the editor indicates this by placing the second date in brackets.

(2) <Q A 1764 FN CBURNEY>
<X CHARLES BURNEY>
<P 44>
[} [\TO FANNY BURNEY PARIS, 18-(20) JUNE 1764\] }]

The text level codes used in the Extension and the Supplement are the same as those in the *Helsinki Corpus* and the CEEC. It must be said, however, that as the orthography, especially in the 18th century, became closer to Modern Standard English, some of the coding, especially in the Extension, very often need not be applied. This is the case with the letters thorn and yogh, both of which had disappeared by then. Nevertheless, superscripts ('wth' coded as w=th=), accents, often in 'foreign language' phrases included in the letters (e.g. 'à charge' as (\a` charge\)), and emendations [{…{] and editor's comments are all still in use.

Example (3) below illustrates the text level coding used in the letter of Charles Burney, a music historian and a man of humble beginnings, to his daughter. It is an extract from the letter whose parameter coding is given in (2) above. In addition to the coding, the extract also clearly shows the nature of the material we constantly strive to find for inclusion in the corpus.[5] The letter was written to Burney's 14 year-old daughter Fanny from Paris where Charles was with his favourite daughter Susanna Elizabeth, or Sukey as the father calls her. Any other font than basic roman, such as italic, is indicated using (^…^) codes.

(3) Paris. Monday 18=th= June 1764

I am sure it will please my dear Fanny &c &c very much, to hear that Sukey is (^a great deal better^) , tho' I were to write nothing Else. & indeed I have but little Time to spare – She continued very ill here till Saturday, w=th= the most frightful fits of Cough & bleeding at the Nose I ever saw, w=ch= has made me hitherto pass my Time very ill. It has soured all my Enjoyments here, or prevented them: as I could but seldom leave her, & when I did, her Situation & Sufferings were always uppermost in my Thoughts. She now, however, wakes in the morn=g= & goes to sleep at Night without a Coughing Fit – & when she has one 'tis by no

[5] For more on the discussion concerning personal and private letters, see Nevalainen & Raumolin-Brunberg 1996.

> means so bad. She has better spirits & more appetite, both of w=ch= were quite tost till within these two Days. No progress is as yet made about placing your Sisters here. It turns out a far more difficult thing to find a proper house for them than I imagined.
> Wednesday Night [^CHARLES BURNEY CONTINUED THE LETTER ON 20 JUNE^]
> I cannot send this away without telling you that Sukey is still better than when the above was written, & likewise that I have now Hopes of placing Hetty & her much to my Satisfaction. Indeed it will cost a good deal more money than I expected, but I am now too far advanced
> <P 45>

6. Availability and copyright

The ultimate aim of the *Corpus of Early English Correspondence* is to provide almost four centuries of authentic material, and to make it available to those interested in language history. However, as the work with the corpus is still in progress and there are still major obstacles to be overcome, the biggest being the copyright question, we do not foresee the publication taking place in the near future. As mentioned above, the Sampler version of the CEEC containing material no longer under copyright has been published and is available for use on the second ICAME CD-ROM.

We are currently working on the copyright question for all the remaining material of the CEEC. It is our wish that numerous publishers whose material, sometimes even entire volumes, we have included in the corpus grant us the permission to use it for research purposes,[6] as this corpus will undoubtedly help researchers to tackle issues concerning language¬variation and change. Work is still in progress on the Extension and the Supplement, and we will have to complete them before their copyright problems are investigated.

The publication of the CEEC, its Extension and Supplement will probably be in the form of collections, i.e. the format used in the CEECS. This arrangement will provide easy access, allowing comparisons between various socially important groups. It will also offer the users the opportunity to investigate regional variation in language. Furthermore, we are currently working on annotation of the CEEC in co-operation with the University of York. To enable easier and more efficient use of the CEEC, its Extension and Supplement, this project will be completed before the publication.

More information about the publications and updates of our work can be found at http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/index.html. [Updated 10 March 2011]

E-mail: mikko.laitinen(at)helsinki.fi

[6] As we will probably not be granted the rights to all the material in the various parts of the corpus, it is therefore unlikely that the complete corpus will be published.

REFERENCES

Keränen, Jukka (1998) "The *Corpus of Early English Correspondence*: Progress report." In Renouf, Antoinette (ed.) *Explorations in Corpus Linguistics*. Language and Computers: Studies in Practical Linguistics 23. Amsterdam/Atlanta, GA: Rodopi. 29-37.

Kytö, Merja (comp.) (1996) *Manual to the Diachronic Part of the Helsinki Corpus of English Texts. Coding Conventions and Lists of Source Texts*. 3rd ed. Helsinki: Department of English, University of Helsinki.

Nevalainen, Terttu (1997) "Ongoing work on the *Corpus of Early English Correspondence*." In Hickey, Raymond, Merja Kytö, Ian Lancashire & Matti Rissanen (eds) *Tracing the Trail of Time: Proceedings from the Second Diachronic Corpora Workshop*. Language and Computers: Studies in Practical Linguistics 18. Amsterdam/Atlanta, GA: Rodopi. 81-90.

Nevalainen, Terttu & Helena Raumolin-Brunberg (1994) "Sociolinguistics and language history: The Helsinki Corpus of Early English Correspondence." *Hermes, Journal of Linguistics* 13. 135-143.

Nevalainen, Terttu & Helena Raumolin-Brunberg (eds) (1996) *Sociolinguistics and Language History: Studies Based on the Corpus of Early English Correspondence*. Language and Computers: Studies in Practical Linguistics 15. Amsterdam/Atlanta, GA: Rodopi.

Nurmi, Arja (ed.) (1998) *Manual for the Corpus of Early English Correspondence Sampler CEECS*. Helsinki: Department of English, University of Helsinki. Available at http://www.hit.uib.no/icame/ceecs/index.htm.

Nurmi, Arja (1999a) *A Social History of Periphrastic DO*. Mémoires de la Société Néophilologique de Helsinki 56. Helsinki: Société Néophilologique.

Nurmi, Arja (1999b) "The *Corpus of Early English Correspondence Sampler* (CEECS)." *ICAME Journal* 23. 53-64.

Palander-Collin, Minna (1999) *Grammaticalization and Social Embedding. I THINK and METHINKS in Middle and Early Modern English*. Mémoires de la Société Néophilologique de Helsinki 55. Helsinki: Société Néophilologique.

Raumolin-Brunberg, Helena (1997) "Incorporating sociolinguistic information into a diachronic corpus of English." In Hickey, Raymond, Merja Kytö, Ian Lancashire & Matti Rissanen (eds) *Tracing the Trail of Time: Proceedings from the Second Diachronic Corpora Workshop*. Language and Computers: Studies in Practical Linguistics 18. Amsterdam/Atlanta, GA: Rodopi. 105-117.