The *Corpus of Early English Correspondence* Extension and Supplement

Samuli Kaislaniemi
University of Helsinki

Research Unit for Variation, Contacts and Change in English (VARIENG)

Table 2. List of collections in the CEEC Extension (as of May 2006).

COLLECTION	YEARS	WORDS
Addison	1699?–1718	14,201
Austen	1796?–1800?	27,955
Banks	1704–1756	39,162
Bentham	1745–1800	51,613
Blomefield	1730–1751	13,318
Bolton	1695–1700	8,300
Bowrey	1687–1708	19,229
Burney	1762–1784	39,451
Burney Fanny	1770–1800	55,729
Bute (George III)	1756–1764	5,372
Carter	1737–1800	c. 26,413
Champion	1774–1776	10,790
Clavering	1705?–1741	72,845
Clift	1792–1799	52,038
Cowper Spencer	1732–1764	10,187
Cowper William	1759–1799	56,254
Crisp	1779–1782	18,395
Culley	1784–1785	25,100
Darwin	1763–1797	18,362
Defoe	1703–1729?	32,688
Dodsley	1743–1764	48,691
Draper	1757?–1775?	22,511
Dukes	1732–1750	50,858
Evelyn	1665–1703	38,929
Evelyn 2	1665–1700	14,382
Fleming 2	1653–1701	76,297
Fleming Extra	1684–1698	14,020
Foundling	1758–1767	50,221
Garrick	1733–1777	42,832
Gay	1705–1731	7,227
George III	1765–1783	7,765
George III A	1779–1800	51,638
George IV	1778–1800	73,065
Gibbon	1750–1793	26,540
Giffard 2	1665?–1722?	11,954
Gower	1783–1800	16,989
Gray	1734?–1771	42,694
Haddock 2	1688–1719	4,647
Hatton 2	1682–1704	25,575
Henry	1660–1693	10,637
Hurd	1739–1797	37,415
Johnson Samuel	1732–1784	25,706
Jones	1768–1794	33,014
Lennox	1761–1800	66,361
Liddell	1709–1716	36,816
Melbourne	1776–1799?	3,824
Montagu	1710?–1761	76,408
Newdigate	1731–1797	30,054
North Country Life	1716–1788	14,553
Original 4	1682–1716	2,900
Pauper	1731–1795?	2,220
Pepys 2	1681–1692	9,435
Pepys 3	1665–1700	27,129
Perrot Jane	1799–1800	8,924
Petty 2	1682–1687	14,378
Pierce	1751–1775	20,354
Pinney	1679–1706	25,098
Piozzi	1784–1798	39,572
Pitt	1751–1757	9,071
Pitt 2	1754	15,618
Pope	1708–1744?	32,869
Porter	1789–1800?	13,815
Prideaux 2	1681–1722	15,934
Purefoy	1736–1754	28,549
Royal 4	1681?–1683?	4,408
Sancho	1768?–1780?	20,216
Secker	1738–1761	31,639
Stubs	1791–1800	4,270
Swift	1712–1734	57,445
Tixall 2	1684–1686	392
Twining	1762–1800	57,793
Wanley	1694–1726	32,725
Warton	1745?–1790	31,344
Wedgwood	1763–1793	35,249
Wentworth 2	1705–1739	62,223
Wollstonecraft	1773?–1797	31,908
Young	1707?–1765	25,842
Totals:	1653–1800	**c. 2,220,345**

Table 1. The *Corpus of Early English Correspondence* (CEEC-400): Published versions (CEECS & PCEEC) and constituents (CEEC 1998, CEECE, CEECSup) compared.

	CEEC (1998)	CEECS	PCEEC	CEECE	CEECSup	CEEC-400
Words	2,597,795	450,182	2,159,132	2,220,345	442,480	5,260,620
Collections	96	23	84	77	19	192
Letters	5,961	1,147	4,970	4,921	859	11,741
Informants	777	194	666	311	94	1,182
Time span	c1410–1681	1418–1680	c1410–1681	1653–1800	1402–1663?	1402–1800
Proportion of women:						
Words	17 %	23 %	N/A	28 %	13 %	21 %
Informants	22 %	22 %	N/A	32 %	30 %	25 %

Table 3. List of collections in the CEEC Supplement (as of May 2006).

COLLECTION	YEARS	WORDS
Arundel 2	1608–1638	5,199
Bacon Extra	1596–1602	8,954
Bacon Dorothy	1597?–1622	4,654
Betts	1522–1640	2,624
Cary	1625?–1659	10,259
Factory	1613–1622	198,497
Gardiner 2	1529–1546	4,408
Gawdy 2	1580–1614	35,586
Grene	1530s	3,029
Knyvett 2	1621–1644	23,945
LisleH	1531–1539	12,552
Oxinden Extra	1635–1663?	9,037
Paston Extra	1467–1503?	13,126
Plumpton 2	1461–1549?	36,432
Ralegh	1581–1618	18,514
Ralegh 2	1583–1617	21,095
Symcotts	1629–1660	13,789
Thynne	1570?–1611	19,574
Zouche	1402–1403	1,206
Totals:	1402–1663	**442,480**

1. Introduction

The *Corpus of Early English Correspondence* (CEEC) is a diachronic corpus of personal letters designed for historical sociolinguistics, compiled at the Department of English and the *Research Unit for Variation and Change in English* (VARIENG) at the University of Helsinki (for the compilers, see the bottom of this poster). This poster presents the scope and contents of the CEEC Extension and Supplement. For information on the background of the corpus and the Historical Sociolinguistics project at VARIENG, please refer to the works listed in the bibliography (see especially Nevalainen & Raumolin-Brunberg 1996: 39-54, 2003: 43–52; Raumolin-Brunberg 1997; Raumolin-Brunberg & Nevalainen 2007).

2. The CEEC family

The Corpus of Early English Correspondence (CEEC) consists of five corpora, as illustrated in Table 1 above: the 1998 version of the CEEC (CEEC 1998) with its two published extracts, the Corpus of Early English Correspondence Sampler (CEEECS) and the Parsed Corpus of Early English Correspondence (PCEEC); and the CEEC Extension (CEECE) and Supplement (CEECSup). The last two corpora were initiated in 2000, when it was decided to extend the CEEC (1998) to cover the eighteenth century on the one hand (CEECE; see Laitinen 2002), and to try and fill the gaps in the CEEC (1998) on the other (CEECSup). Work continues on the CEECE and CEECSup, but the entire CEEC (called the "CEEC-400") approaches completion, as it has almost reached a satisfactory level of socio-regional representativeness.

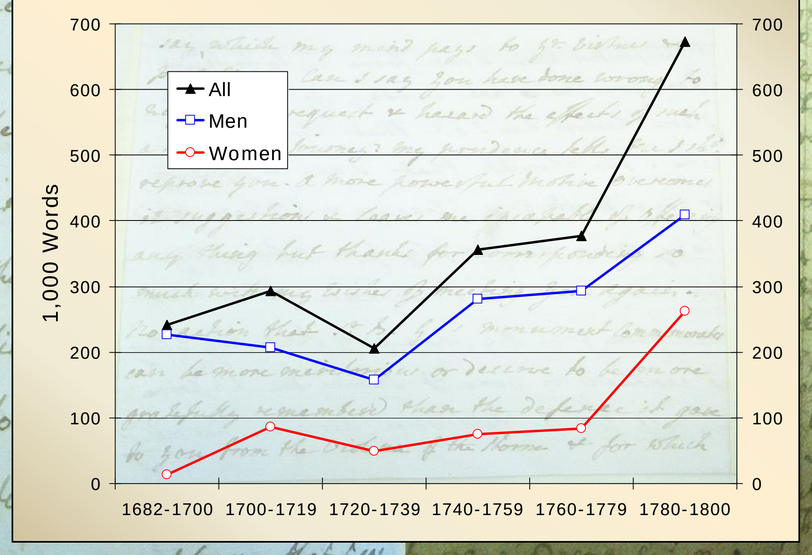
3. The CEECSup

The CEECSup differs from the CEEC (1998) and the CEECE in that it does not constitute a balanced corpus in itself. The time span overlaps that of the CEEC (1998), while the socioregional coverage is aimed to fill the gaps of the CEEC (1998). For a list of the CEECSup collections as of May 2006, see Table 3 above.

The Supplement contains four kinds of material:

1. Material which has only become available after 1998 (such as the letters of Sir Walter Raleigh)
2. Material which does not fulfil the criteria of the CEEC (1998) (e.g. by having modernized spelling) but which is scarce and valuable, and which is suitable for morphosyntactic research (such as Betts and Zouche)
3. Material which has only been discovered by the project after 1998, such as letter editions without transparent titles (e.g. the Factory collection, which are English East India Company merchant letters published under the title *The English Factory in Japan, 1613–1623* (Farrington 1991))
4. Material included to increase the word counts of certain individual writers already represented in the CEEC and CEECE, in order to facilitate research on rare linguistic features (this material includes collections such as Pepys 3 and all collections titled "Extra", some of which are included in the CEECE)

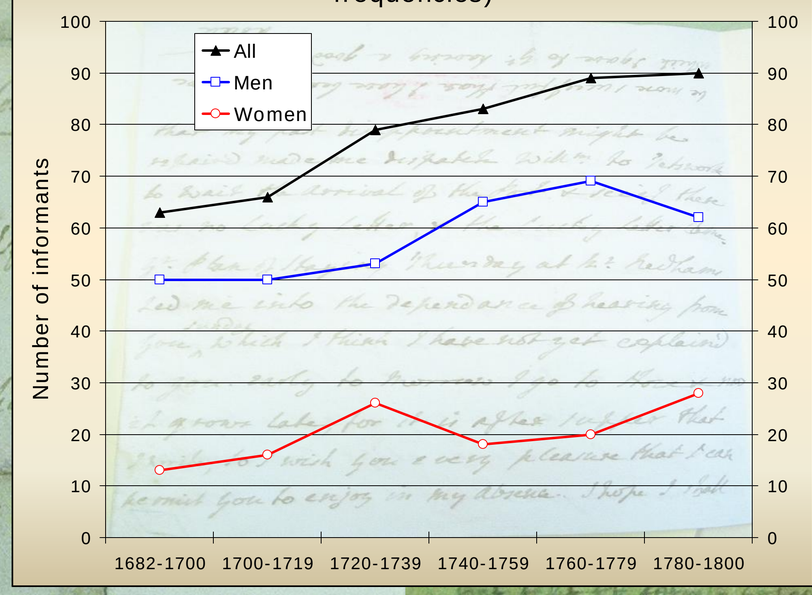
Figure 1: CEECE words by gender



4. The CEECE

The bulk of the CEECE consists of collections of eighteenth-century letters, but some collections are continuations of collections included in the CEEC (1998), e.g. Pepys 2 is a continuation of Pepys in the CEEC (1998). Table 2 on the far left lists all the collections included in the CEECE as of May 2006. Figures 1 and 2 show the number of words and the absolute frequency of informants in the CEECE, divided by gender and time period (true figures for the number of informants are given in Table 1 above).

Figure 2: CEECE Informants (absolute frequencies)



Background images are taken by Samuli Kaislaniemi:

Left side: Hester Grenville to William Pitt, 15 October 1754. TNA 30/8/7/1 ff. 10–14. Included in the CEECE collection Pitt 2.

Right side: Richard Cocks to Robert Cecil/Thomas Wilson, 10 December 1614. TNA CO 77/1 no. 42/43. Included in the CEECSup Factory collection.

The original version of this poster was presented at the 27th ICAME (International Computer Archive of Modern and Medieval English) conference, 24–28 May 2006. This online version has been slightly revised.

Samuli Kaislaniemi, November 2007

Compilers

CEEC, CEECS, PCEEC: Terttu Nevalainen (leader), Jukka Keränen, Minna Nevala (née Aunio), Arja Nurmi, Minna Palander-Collin and Helena Raumolin-Brunberg.

CEECE and CEECSup: Terttu Nevalainen (leader), Teo Juvonen, Samuli Kaislaniemi, Mikko Laitinen, Minna Nevala, Arja Nurmi, Minna Palander-Collin, Helena Raumolin-Brunberg, Anni Sairio (née Vuorinen), Tanja Säily and Tuuli Tahko.

In addition, the project has been helped by a number of research assistants over the years: Maarit Alanko and Kirsi Heikkonen assisted with the CEEC (1998), while Bethany Fox and Eero Timoskainen have assisted in compiling the CEECE and CEECSup.

References

Farrington, Anthony. 1991. *The English Factory in Japan 1613–1623*. London: The British Library.

Laitinen, Mikko. 2002. "Extending the Corpus of Early English Correspondence to the 18th century". *Helsinki English Studies* 2. <http://www.eng.helsinki.fi/hes/>.

Leech, Geoffrey. 1993. "100 million words of English". *English Today* 33/9.1:9–15.

Nevalainen, Terttu & Helena Raumolin-Brunberg. 1996. "The Corpus of Early English Correspondence". *Sociolinguistics and Language History: Studies Based on the Corpus of Early English Correspondence* ed. by Terttu Nevalainen & Helena Raumolin-Brunberg, 39–54. Amsterdam & Atlanta, GA: Rodopi.

Nevalainen, Terttu & Helena Raumolin-Brunberg. 2003. *Historical Sociolinguistics: Language Change in Tudor and Stuart England*. London: Longman.

Nurmi, Arja. 1998. *Manual for the Corpus of Early English Correspondence Sampler CEECS*. Department of English. University of Helsinki. <http:/khnt.hit.uib.no/icame/manuals/ceecs/>

Nurmi, Arja. 2002. "Does size matter? The *Corpus of Early English Correspondence* and its sampler". *Variation Past and Present: VARIENG Studies on English for Terttu Nevalainen* ed. by Helena Raumolin-Brunberg, Minna Nevala, Arja Nurmi & Matti Rissanen, 173–184. Helsinki: Société Néophilologique (*Mémoires de la Société Néophilologique de Helsinki* 61).

Raumolin-Brunberg, Helena. 1997. "Incorporating sociolinguistic information into a diachronic corpus of English". *Tracing the Trail of Time: Proceedings of the Diachronic Corpora Workshop, Toronto (Canada) May* 1995, ed. by Raymond Hickey, Merja Kytö, Ian Lancashire & M. Rissanen, 105–117. Amsterdam & Atlanta, GA: Rodopi.

Raumolin-Brunberg, Helena & Terttu Nevalainen. 2007. "Historical sociolinguistics: The *Corpus of Early English Correspondence*". *Creating and Digitizing Language Corpora: Diachronic Databases* vol. 2: *Diachronic Databases* ed. by J. C. Beal, K. Corrigan & H. Moisl, 148–171. Houndmills: Palgrave.