A CORPUS OF LATE MODERN BRITISH AND AMERICAN ENGLISH PROSE (COLMOBAENG)\(^1\)

TERESA FANEGO

University of Santiago de Compostela

1. Corpus make-up

1.1. Period, dialect and genre division

COLMOBAENG (1700-1879) is a 1,170,000 word database comprising texts drawn from both printed and electronic sources; the former are listed in the references at the end of this file, the latter include the Century of Prose Corpus, Project Gutenberg, the Online Books Page at the University of Pennsylvania, the Electronic Text Center at the University of Virginia, and Chadwyck-Healey’s collections of Early American Fiction, Eighteenth-Century Fiction and Nineteenth-Century Fiction. At present, the structure of the corpus is as shown in Tables 1-3.

<table>
<thead>
<tr>
<th>British English</th>
<th>American English</th>
</tr>
</thead>
<tbody>
<tr>
<td>BrE1 1700-1726</td>
<td>AmE1 1700-1724</td>
</tr>
<tr>
<td>BrE2 1732-1757</td>
<td>AmE2 1732-1759</td>
</tr>
<tr>
<td>BrE3 1761-1797</td>
<td>AmE3 1774-1804</td>
</tr>
<tr>
<td>BrE4 1850-1879</td>
<td>AmE4 1851-1879</td>
</tr>
</tbody>
</table>

*Table 1. Range of periods and dialects represented*

<table>
<thead>
<tr>
<th></th>
<th>BrE1</th>
<th>BrE2</th>
<th>BrE3</th>
<th>BrE4</th>
</tr>
</thead>
<tbody>
<tr>
<td>nonfiction</td>
<td>76,000 w</td>
<td>76,000 w</td>
<td>76,000 w</td>
<td>76,000 w</td>
</tr>
<tr>
<td>fiction</td>
<td>124,000 w</td>
<td>124,000 w</td>
<td>124,000 w</td>
<td>124,000 w</td>
</tr>
<tr>
<td>Total words</td>
<td>200,000</td>
<td>200,000</td>
<td>200,000</td>
<td>200,000</td>
</tr>
</tbody>
</table>

*Table 2. Range of texts from British English: fiction vs nonfiction (word [w] totals)*

<table>
<thead>
<tr>
<th></th>
<th>AmE2</th>
<th>AmE3</th>
<th>AmE4</th>
</tr>
</thead>
<tbody>
<tr>
<td>nonfiction</td>
<td>50,000 w</td>
<td>40,000 w</td>
<td>76,000 w</td>
</tr>
<tr>
<td>fiction</td>
<td>80,000 w</td>
<td>80,000 w</td>
<td>124,000 w</td>
</tr>
<tr>
<td>Total words</td>
<td>120,000</td>
<td>120,000</td>
<td>200,000</td>
</tr>
</tbody>
</table>

*Table 3. Range of texts from American English: fiction vs nonfiction (word [w] totals)*

\(1\) For fuller details see Fanego (2012).
The nonfiction samples in the first three subperiods of British English (BrE1-3) have been taken from Milic’s COPC, and this has influenced the choice of nonfiction texts in the rest of the database. The COPC is intended to constitute “an inventory of the daily language of the literate members of English society” of the eighteenth century (Milic 1995: 329). Therefore, in the interest of comparability, no attempt was made to include in BrE4 (1850-1879) or in the several subperiods of American English records of town, court and church proceedings, or the unstudied writing of uneducated individuals, since none of these would have an equivalent in the COPC.

A second criterion guiding the selection of the American English texts was the exclusion of authors born outside the United States, though exceptions were made in the case of Charles Thomson (1729-1824), whose family migrated from Ireland in 1739, Hugh Henry Brackenridge (1748-1816), born near Cambettown, Scotland, but brought up in Pennsylvania from the age of five, Alexander Hamilton (1757-1804), who at the age of fifteen moved to New York from the Caribbean island of Nevis, and John William Draper (1811-1882). Draper was born in Liverpool, England, and came to the United States in 1832; he graduated from the medical department of the University of Pennsylvania in 1836 and in 1839 became Professor of Chemistry at the University of the City of New York. It seems reasonable to assume that by 1874, when the History of the Conflict between Religion and Science – the only work by Draper included in COLMOBAENG – was published, his writing style would have been representative of standard American usage.

Considerable efforts were also made to leave out all American works printed in Great Britain, since, as noted by Kytö & Rissanen (1983: 476), “there is always the possibility of British interference on the printer’s side.” However, in view of the scarcity of materials published in America during AmE2 (1732-1759) I finally opted for the inclusion of two texts printed in London, namely John Bartram’s Observations (1751) and Charles Thomson’s An Enquiry (1759).

1.2. Contents: fiction texts

Unless otherwise indicated, all extracts have around 20,000 words and have been taken from Chadwyck-Healey’s electronic text collections. Dates of publication are given after each title; birth-dates in brackets.

Subperiod BrE1 (1700-1726):
Anonymous. The Adventures of Lindamira, a Lady of Quality, 1702. [2,000 words]. COPC
Barker, Jane (1675-1743). The Amours of Bosvil and Galesia, 1719. [14,000 words]
Barker, Jane (1675-1743). Exilius: or, The Banish’d Roman, 1719. [6,300 words]
Davys, Mary (1674-1732). Familiar Letters Betwixt a Gentleman and a Lady, 1725. [13,110 words]
Davys, Mary (1674-1732). The Lady’s Tale, 1725. [7,240 words]
Defoe, Daniel (1659-1731). The Fortunes and Misfortunes of the Famous Moll Flanders, 1722.
Haywood, Eliza Fowler (1693?-1756). Fantomina: or, Love in a Maze, 1725. [11,900 words]
Haywood, Eliza Fowler (1693?-1756). The Fatal Secret: or, Constancy in Distress, 1725. [9,100 words]
Heare, Mary (dates unknown). Lover’s Week, 1718. [2,000 words]. COPC
Manley, Mary de la Rivière (1663-1724). The Secret History of Queen Zarah and the Zarazians, 1705.

Subperiod BrE2 (1732-1757):
Coventry, Francis (died 1754). Pompey the Little, 1752. [2,000 words]. COPC
Fielding, Henry (1707-1754). A Journey from this World to the Next, 1743.
Fielding, Sarah (1710-1768). The Governess, 1749. [2,000 words]. COPC
Haywood, Eliza Fowler (1693?-1756). The History of Miss Betsy Thoughtless, 1751.
Richardson, Samuel (1689-1761). The History of Sir Charles Grandison, 1754.

Subperiod BrE3 (1761-1797):
Burney, Frances (1752-1840). Evelina, or, a Young Lady’s Entrance into the World, 1778.
Godwin, William (1756-1836). Things as They Are; or, The Adventure of Caleb Williams, 1794.
Jenner, Charles (dates unknown). The Placid Man, 1770. [2,000 words]. COPC
Johnstone, Charles (c1719-c1800). Chrysal or, The Adventures of a Guinea, 1761. [2,000 words]. COPC

**Subperiod BrE4 (1850-1879):**
Braddon, Mary Elizabeth (1837-1915). *Lady Audley's Secret*, 1862. [2,000 words]
Gaskell, Elizabeth Cleghorn (1810-1865). *North and South*, 1855. [2,000 words]
Oliphant, Mrs. (Margaret) (1828-1897). *Salem Chapel: Chronicles of Carlingford*, 1863.

**Subperiod AmE3 (1774-1840):**
Foster, Hannah Webster (1759-1840). *The Coquette; or, The History of Eliza Wharton*, 1797.

**Subperiod AmE4 (1851-1879):**
Hawthorne, Nathaniel (1804-1864). *The House of the Seven Gables*, 1851.
Melville, Herman (1819-1891). *I and my Chimney*, 1856. [5,480 words] UVirg
Melville, Herman (1819-1891). *Bartleby, the Scrivener*, 1856. [14,540 words]
Spofford, Harriet Prescott (1835-1921). *In a Cellar*, 1859. [8,250 words]

1.3. Contents: nonfiction texts

**Subperiod BrE1 (1700-1726):**
8 extracts of 5,000 words each from Part A of the COPC.²
Defoe, Daniel (1659?-1731). *An Appeal to Honour and Justice* (1715).
Defoe, Daniel (1659?-1731). *A Tour through the Whole Island of Great Britain* (1724).
Steele, Richard (1672-1729). *Spectator* nos. 11, 49, 109, 113 (1710-1711).
18 extracts of 2,000 words each from Part B of the COPC:
Atterbury, Francis (1663-1732). *English Advice* (1714).
Baron, Samuel (dates unknown). *Kingdom of Tonqueen* (1700).
Bell, John (1691-1780). *St. Petersburg to Pekin* (1719).

² The COPC consists of two parts: Part A, which comprises 5,000-word extracts from “the foremost prose writers of the period” (Milic 1995: 329), and Part B, which comprises 2,000-word extracts from the works of secondary writers.
The Free-Thinker (1720).
Monthly Miscellany (1707)
Ray, John (1627-1705). Philosophical Letters (1718).^{3}
Rowe, Nicholas (1674-1718). Life of Shakespear (1709).

Subperiod BrE2 (1732-1757):
9 extracts of 5,000 words each from Part A of the COPC:
Fielding, Henry (1707-1754). Covent Garden Journal nos. 3-6, 10 (1752).

18 extracts of 2,000 words each from Part B of the COPC:
Cibber, Colley (1671-1757). Occasional Letter to Pope (1744).
Daily Advertiser (1741).
Egmont, Percival (1683-1748). Diary (1739).
Holmes, John (dates unknown). Art of Rhetoric (1739).
Lediard, Thomas (1685-1743). Life of Marlborough (1736).
London Magazine (1736).
Neal, Daniel (1678-1743). History of the Puritans (1732).
North, Roger (1651?-1734). Life of Francis North (1742).^{4}

Subperiod BrE3 (1761-1797):
8 extracts of 5,000 words each from Part A of the COPC:
Boswell, James (1740-1795). Hypochondriack nos. 1, 21, 27 (1777-1779).
Boswell, James (1740-1795). Selections from the Life of Johnson (1791).
Gibbon, Edward (1737-1794). History, Chapter XXII (1781).

^{3} Ray’s collected letters were published posthumously in 1718. Milic (1995: 330) points out that because “editors and printers… had during the eighteenth century substantial latitude in altering an author’s words, it is safe to assume that the language of the selection as represented in these posthumously published letters is that of the publication date and not that of writing, which we cannot accurately determine in any case.”

^{4} The Life of Francis North was published eight years after Roger North’s death, by his son Montagu.
Johnson, Samuel (1709-1784). *Preface to Shakespeare* (1765), *Some Thoughts on the Falkland’s Islands* (1771).


Walpole, Horace (1717-1797). *Life of Mr. Thomas Baker* (1778).

18 extracts of 2,000 words each from Part B of the COPC:

Akin, John (1747-1822). *Natural History and Poetry* (1777).


Granger, James (1723-1776). *Biographical History* (1769).


Jones, William (1726-1800). *Natural Philosophy* (1762).


Tytler, William (1711-1792). *Mary Queen of Scots* (1760).


*Westminster Magazine* (1776).

**Subperiod BrE4 (1850-1879):**


Private Letters by John Richard Green and Lord and Lady Amberley, 1861 and 1872. [5,000 words] Denison


**Subperiod AmE2 (1732-1759):**


Franklin, Benjamin (1706-1790). *Poor Richard’s Almanack* (1733-1758). [5,900 words] UPenn

Woolman, John (1720-1772). *Some Considerations on the Keeping of Negroes*, 1754. [4,731 words]

**Subperiod AmE3 (1774-1804):**

Jefferson, Thomas (1743-1826). *Notes on the State of Virginia*, 1781-1782. [5,000 words] UVirg
Woolman, John (1720-1772). *Journal*, 1774. [Chapter XI, written in 1772; 5,000 words] UVirg

**Subperiod AmE4 (1815-1879):**

Emerson, Ralph Waldo (1803-1882). *The Conduct of Life*, 1860. [6,010 words] UVirg
Emerson, Ralph Waldo (1803-1882). *Thoreau*, 1862. [4,544 words]
Fiske, John (1842-1901). *Spain and the Netherlands*, 1868. [5,040 words] UVirg
Holmes, Oliver Wendell (1809-1894). *The Autocrat of the Breakfast-Table*, 1858. [5,050 words] GUT
Motley, John Lothrop (1814-1877). *The Life and Death of John of Barneveld, Advocate of Holland with a View of the Primary Causes and Movements of the Thirty Years' War*, 1874. [5,040 words] GUT
Several authors. A group of private letters, 1861-1867. [about 3,800 words] UVirg
Thoreau, Henry David (1817-1862). *The Correspondence of Henry David Thoreau*, 1851-1862. [Three letters from the years 1851 and 1862; 1,260 words.]
Thoreau, Henry David (1817-1862). *Walden, or, Life in the Woods*, 1854. [5,040 words] UVirg
Twain, Mark (1835-1910). *Innocents Abroad*, 1869. [5,000 words] UVirg
Whitman, Walt (1819-1892). *Memoranda during the War*, 1875-1876. [5,100 words] UVirg

2. **Research possibilities and limitations**

The corpus is too small for lexicographic purposes, and is also unsuitable for the study of phenomena related to spelling or punctuation. It should also be borne in mind that texts retrieved from internet sources often lack exact bibliographical information, so that it is not clear who the editors were, nor the degree of editorial intervention. Furthermore, Milic (1995: 330-331) acknowledges that he opted for “normalizing” the spelling, punctuation and capitalization practices of the texts in the COPC.

A second limitation is that, as noted in section 1, the corpus is biased towards texts written by literate members of English and American society in the eighteenth and nineteenth centuries, and includes neither speech-related genres such as town and court depositions, where speech is taken down as either direct or indirect speech, nor drama, in which dialogues are written to be spoken. Many of the fiction extracts included in the corpus contain dialogue written to represent speech, but in general COLMOBAENG is not adequate to address questions having to do with the natural spoken language of the period. By contrast, in terms of Biber’s multi-dimensional approach to register variation, discourse dimensions such as narrative vs. non-narrative concerns (Dimension 2; see Biber 1988: 109, 136) and informational vs. involved production (Dimension 1; see Biber 1988: 107, 128) are well represented in the corpus and might be explored by selecting samples from the appropriate registers (e.g. fictional and historical prose, philosophical writings, academic prose, etc.).

COLMOBAENG was compiled with the specific aim of investigating syntactic development within the Late Modern period, and thus far has been used successfully to explore changes in the English system of sentential complementation, such as the competition between infinitives and gerunds as preverbal and extraposed subjects (e.g. “Inviting / To invite the twins was a bad mistake”; see Fanego 2007: 186-219, 2010), and the specialization of gerunds to encode object clauses with “non-controlled” implicit subjects...
(e.g. “A visit to London entails bringing gifts for friends”) as opposed to infinitival object clauses with controlled implicit subjects (e.g. “I want to see Mary”; see Fanego 2007: 178-186). Other areas of grammar that were also undergoing change in LME and which would thus constitute suitable research topics include the be Ving progressive, the passive and other impersonalizing strategies, and the variation between subjunctives, indicatives and modals in certain types of clauses (e.g. “… if the temperature be increased / was increased / should be increased”).

Finally, in terms of size the corpus is comparable to traditional small historical corpora such as the Helsinki Corpus (551,000 words of Early Modern English; see Kytö 1996), CONCE (1,000,000 words of nineteenth-century English) and ARCHER (A Representative Corpus of Historical English Registers, 1600-1999). It can be used most successfully to supplement the latter two, in particular ARCHER, since this also comprises both eighteenth- and nineteenth-century British and American English. Unlike ARCHER, however, which can only be consulted in situ at one of the fourteen departments belonging to the ARCHER Consortium, COLMOBAENG is freely available and can be obtained by simply contacting its compiler at <teresa.fanego@usc.es>.

References

Primary sources:

COLMOBAENG comprises materials from the following sources:

a) electronic databases:

COPC Century of Prose Corpus (Milic 1995).
Denison A Corpus of Late Modern English Prose (Denison 1994).
EAF Early American Fiction, Chadwyck-Healey.
ECF Eighteenth-Century Fiction, Chadwyck-Healey.
GUT Project Gutenberg.
NCF Nineteenth-Century Fiction, Chadwyck-Healey.
UPenn Online Books Page, University of Pennsylvania.
UVirg Electronic Text Center, University of Virginia.

b) printed texts:


Secondary sources:


5 Cf. <http://www.alc.manchester.ac.uk/subjects/lel/research/projects/archer/>


