DE GRUYTER OPEN

ICAME Journal, Volume 40, 2016, DOI: 10.1515/icame-2016-0003

Corpus of Early American Literature[1]

Mikko Höglund, Stockholm University
Kaj Syrjänen, University of Tampere

1 Introduction

The use of electronic corpora has become widespread in many subfields of linguistics, and English, perhaps to some extent owing to its status as a lingua franca, has been the trailblazer in this methodology. The first widely used electronic corpus, the Brown corpus, compiled in the 1960s (see Kučera and Francis 1967), comprised American English texts, and as the field has advanced over the last decades, many of the pioneering corpora have consisted of English material. Nowadays, there is an abundance of English corpora available, including corpora of different national varieties of English, learner English, English for specific purposes, and historical English. However, even though the selection of corpora is impressive and broad, the basic subject matter and the object of linguistic research, the English language, is a veritable leviathan to investigate, and there are still notable gaps in the coverage of the available corpora.

In recent years and decades, many corpora have been compiled which document different periods in the history of English. To name a few, the *Helsinki Corpus of English Texts* is one of the first historical corpora, covering the period from the 8th century to the 18th century; the *Corpus of Early English Correspondence* (CEEC) and later versions of it comprise letters from 1403–1800; the *Corpus of Late Modern English Texts* (CLMET; De Smet 2005) and its successors CLMETEV and CLMET3.0 include texts from 1710–1920; and the *Old Bailey Corpus* (OBC) consists of transcripts of the proceedings of the Old Bailey criminal court from 1720 to 1913. In addition to the aforementioned corpora, which document English spoken and written in the Old World, there are also corpora that cover material from the other main variety, American English. The largest of these is the 400-million-word *Corpus of Historical American English* (COHA; Davies 2010–), which documents American English from 1810 to present day. Another corpus of historical American English is the *Corpus of Early American English* (Kytö 1994), which comprises texts from the period 1620–1720, and some historical corpora such as the *Corpus of Late Mod-*

ern British and American English Prose (COLMOBAENG; 1700–1879) also include American English to some extent. We should also mention ARCHER (*A Representative Corpus of Historical English Registers*), a multi-genre corpus which covers a chronologically impressive time span from 1600 to 1999 of both British and American English.

Even though the entire time span of American English is fairly well documented, there is a noteworthy gap in the chronological coverage of the available corpora. The critical period for the formation of the AmE variety, from the early 18th century before the declaration of independence and the times following the declaration, is not consistently documented. For instance, in COLMOBAENG there is some material from the 18th century, and in COHA there is material from 1810 onwards, but there is no single corpus that would cohesively document that formative period. While ARCHER does cover this particular period, its total number of words for American English for the entire time span from 1600 to 1999 is only around 1.34 million (as of version 3.2).

The present paper introduces a corpus which is an endeavor to fill the existing gap in the selection of American English corpora. The corpus is called the *Corpus of Early American Literature* (CEAL) and it includes texts from 1690–1920. The corpus is divided into three subperiods, CEAL1 (1690–1780), CEAL2 (1781–1850), and CEAL3 (1851–1920), and the subperiods cover approximately 1.5, 5.7 and 6.3 million words, respectively (see Table 1). The compilation process and its challenges have been discussed in an earlier paper (Höglund and Syrjänen 2010) and will not be repeated in detail here. The aforementioned paper also outlines the historical events and circumstances that affected the formation of the new nation and with it the American English variety, and considers them in relation to the chronological subdivision of the corpus.

Table 1: Contents of CEAL in numbers

Subcorpora	**Number of texts (files)**	**Authors**	**Words**
CEAL1 (1690–1780)	42	40	1,484,463
CEAL2 (1781–1850)	65	56	5,740,042
CEAL3 (1851–1920)	87	45	6,319,792
TOTAL	*194*	*141*	*13,544,297*

The present paper is structured as follows: First, in Section 2 the structure and contents of CEAL are introduced together with a comprehensive list of the texts,

publication years, authors and their birth and death years, and word counts. This is followed by a brief evaluation of the corpus in Section 3. Finally, Section 4 concludes the paper.

2 Structure and contents of CEAL

The compilation of CEAL was inspired by the CLMET, and the original idea was to create an American English counterpart of the British English CLMET for comparing linguistic phenomena in BrE and AmE. The compilation criteria outlined in De Smet (2005) served as a starting point for CEAL; during the process of compiling CEAL these criteria were gradually modified to be more applicable for AmE. In its final form, CEAL is more or less comparable with the CLMET, but criteria-wise these corpora are not in their final form as similar to each other as e.g. LOB, BLOB, F-LOB and CLOB are to the Brown, B-Brown, Frown and Crown corpora.

The material for CEAL was retrieved from text repositories with open access and/or public domain material. The majority of the texts come from Project Gutenberg, with additional material from the American Studies Commons, Oxford Text Archive, and one text from the Bible Bulletin Board. The method used to collect the texts was very rudimentary and time-consuming, but also inclusive: all the English texts in the databases were manually browsed through, and suitable texts were then downloaded. While the bulk of the texts were available in plain text format, a handful of the texts only available in XML/SGML or PDF were converted manually to plain text using freely available text conversion tools.

The text selection criteria were the following: the author of the text was born and educated on the new continent (or the author had a significant cultural/political influence there; see the discussion in Höglund and Syrjänen 2010: 439–440), only full texts were included, and the word count per author was set to a maximum of 350,000 words. However, the total number of words for most of the authors in the corpus is less than 200,000. In addition, the relationship between the authors' birth year and the publication year(s) of their work was controlled in the same manner as in the CLMET (see Figure 1). This was done primarily to maximize the linguistic differences between the three subperiods. The effect of this method can be seen in Figure 2; most of the texts tend to concentrate towards the latter parts of the three subperiods. Also, in the figure texts whose publication spans multiple years have been marked according to their starting year of publication. Notably, each of the lines represents a publication year and not an individual author or a text, so the word count in this graph is occasionally higher than the per-author word limit of 350,000.

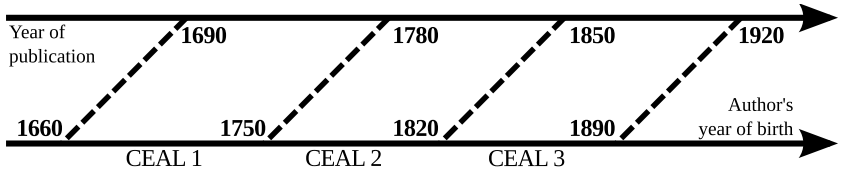

Figure 1: CEAL subperiods and publication year / author's birth year relationship

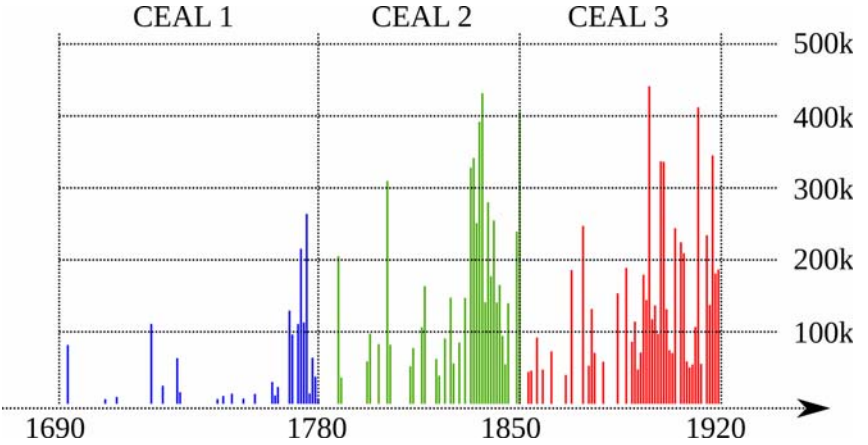

Figure 2: Chronological distribution of the texts in CEAL, with each line representing word count per publication year; texts whose publication spans multiple years are marked according to the starting year

The genre distribution between the three subperiods is somewhat uneven. The original aim was to collect mostly fiction texts in order to make CEAL as comparable with the CLMET as possible, but due to historical realities, it proved not to be possible. Whereas for CEAL2 and CEAL3 there is relatively much fictive material available in the text repositories, for the CEAL1 period, spanning between the late 17th century and 1780, there is not much fiction around that would be native to the American continent, as the compilation criteria stipulated. Apparently, the main problem is not so much the availability of the online material (although much less AmE material is available from the CEAL 1 period than the other periods), but the fact that, at that time, people on the new conti-

nent were hardly writing any fiction of their own. Consequently, almost all of the texts currently included in CEAL1 are political writings, correspondence, journals, and religious texts. In CEAL2 the amount of fiction rises considerably; approximately 42 per cent of the texts are fictive prose. However, still more than half of the texts in CEAL2 are (auto)biographies, political and religious texts, journals, and other non-fiction texts. Lastly, CEAL3 consists almost solely of fictive texts; non-fiction covers only about 7 per cent of the subcorpus.

Regarding the timespan of CEAL2 and especially CEAL3, there was much material available in the text repositories, and we wanted to include many different styles of texts: colloquial style (e.g. David Crockett), complex style (e.g. Henry James), and texts that represent different vernaculars (e.g. F. Norris's *McTeague*; Grey's *Riders of the Purple Sage*), and texts by both male and female authors. The aim was to include as large a variety of styles as possible within the limits of the corpus. Also, if there were several books available by one author, we chose the author's early as well as late works.

The texts that are available in the online repositories from which the texts for CEAL were retrieved have not been uploaded with linguistic research in mind. This means that there is no guarantee that the texts follow the original spelling and form of the source text precisely. For some texts there are multiple versions available, with varying levels of accuracy; for instance, some digitized texts are accompanied by a transcriber's notes section showing the changes made to the original text during digitization. When compiling the corpus, whenever there was an opportunity to make a choice between two or more editions of the same work, we chose the one that we judged to be closest to the original (see the case of Franklin's letters below).

Some editing was done to the texts during the compilation process. For example, in some cases we included only parts of the larger collections of texts that were available by taking only the texts by a certain author from a larger work covering several writers. This was done, for instance, with *The Diplomatic Correspondence of the American Revolution* series of books, from which we only included correspondence by specific authors that fit the compilation criteria (e.g. from volume 7 only John Jay's correspondence was included). Since especially the first subcorpus includes a considerable number of small text sources with potentially overlapping passages (such as separate works covering letters from the same author), we also checked the subcorpora for duplicate texts or redundant passages using an automatic *ad hoc* approach, which involved replacing line breaks from each text with spaces and slicing the resulting text into 50-character strings, each of which was searched from all the other texts in the subcorpus. This approach worked surprisingly well, helping us identify, for

example, various letters by Benjamin Franklin that were included in two volumes of his writings (*Memoirs vol. 2* and *Letters*) but edited in a different way. The duplicate texts were compared to pictures of Franklin's original writings found online, and we decided to keep the letters in *Letters*, as they seemed to be closer to the originals (for example, they retained Franklin's original spelling of the past participle: *suppos'd, concern'd*). Of course, this means that the rest of *Memoirs*, both volumes, which are more heavily edited, are still included in the corpus. Keeping this in mind, we would not recommend using the corpus for research on linguistic phenomena that are susceptible to editing, such as spelling, punctuation, and so forth. On the other hand, we see that other phenomena such as grammar, syntax, vocabulary and semantics are not as likely to have been altered by editors, and are better subjects of investigation using CEAL.

Table 2 lists the texts that are included in CEAL. The table also includes the publication year of each text, the names of the authors and their years of birth and death, and the word counts for both individual texts and authors.[2]

Table 2: Contents of CEAL

CEAL1 (1690-1780)

Author	Birth-Death	Year	Title	Word count	Words/ author
Lyon, Lemuel; Haws, Samuel	-	1758-1775	The Military Journals of two private soldiers	13816	
Adams, John; Carmichael, William; Deane, Silas; Franklin, Benjamin; Lee, Arthur; Lovell, James; Morris, Robert.	-	1775-1780	The Diplomatic Correspondence of the American Revolution, Vol. 9	16173	
Deane, Silas; Franklin, Benjamin; Harrison, Benjamin; Hayward, Thomas; Hooper, William; Lee, Arthur; Lee, Richard Henry; Livingston, Philip; Lovell, James; Morris, Robert; Whipple, William; Witherspoon, John	-	1776-1779	The Diplomatic Correspondence of the American Revolution, Vol. 1	66818	

Mather, Cotton	1663-1728	1693	The wonders of the invisible world	81687	98008
		1706	The Negro Christianized	6527	
		1710	Theopolis Americana	9794	
Beverley, Robert Jr.	1673-1722	1722	The history of Virginia	86808	
Edwards, Jonathan	1703-1758	1731-1750	Selected Sermons of Jonathan Edwards	63404	90603
		1732	Christian Charity	16263	
		1747	True Saints	10936	
Chauncy, Charles	1705-1787	1745	Marvellous Things done by the right Hand and holy Arm of God in getting him the Victory	6733	
Franklin, Benjamin	1706-1790	1722-1726	Boston and London – Writings 1722–1726	24296	145655
		1726-1757	Letters	25107	
		1771-1790	Memoirs (1)	35553	
		1771-1790	Memoirs (2)	60699	
Mayhew, Jonathan	1720-1766	1750	A Discourse concerning Unlimited Submission and Non-Resistance to the Higher Powers	14237	
Woolman, John	1720-1772	1774	The Journal of John Woolman	79276	
Adams, Samuel	1722-1803	1770-1773	The Writings of Samuel Adams (2)	129423	304475
		1773-1777	The Writings of Samuel Adams (3)	111026	
		1778-1802	The Writings of Samuel Adams (4)	64026	

Otis, James Jr.	1725-1783	1764	The Rights of British Colonies Asserted and Proved	30527	42316
		1765	A Vindication of the British Colonies	11789	
Leacock, John	1729-1802	1776	The Fall of British Tyranny	22447	
Sherwood, Samuel	1730-1783	1776	The Church's Flight into the Wilderness	14432	
Stocking, Abner	1730-1806	1775	An interesting journal of Abner Stocking of Chatham, Connecticut	9104	
Rogers, Robert	1727-1795	1766	Ponteach	23571	
Keteltas, Abraham	1732-1798	1777	God Arising	9823	
Washington, George	1732-1799	1754	The Journal of Major George Washington	7284	
Noble, Oliver	1733/4-1792	1775	Some Strictures upon the Sacred Story Recorded in the Book of Esther	11007	
Adams, John	1735-1826	1774-1780	Familiar Letters of John Adams and His Wife Abigail Adams, During the Revolution	80949	
Henry, Patrick	1736-1799	1775	Give Me Liberty or Give Me Death	1223	
Deane, Silas	1737-1789	1776-1779	The Diplomatic Correspondence of the American Revolution, Vol. 1	56986	
Paine, Thomas	1737-1809	1776	Common Sense	21667	78681
		1776-1780	The Writings of Thomas Paine (1)	57014	

Carmichael, William	1739-1795	1776-1780	The Diplomatic Correspondence of the American Revolution, Vol. 9	15635
Drayton, William-Henry	1742-1779	1776	A Charge, on the Rise of the American Empire	8991
Jefferson, Thomas	1743-1826	1775-1780	Memoirs, Correspondences, and Miscellanies, from the Papers of Thomas Jefferson	75450
Adams, Abigail	1744-1818	1774-1780	Familiar Letters of John Adams and His Wife Abigail Adams, During the Revolution	55202
Jay, John	1745-1829	1779-1780	The Diplomatic Correspondence of the American Revolution, Vol. 7	33372
Green, Ezra	1746-1847	1777-1778	Diary of Ezra Green	4690
Hardenbergh, John L.	1748-1806	1779	The Journal of Lt. John L. Hardenbergh	4486
Dodge, John	1751-1800	1780	The Dodge Narrative	6212
TOTAL				**1484463**

CEAL2 (1781-1850)

Author	Birth-Death	Year	Title	Word count	Words/ author
Hamilton, Alexander; Madison, James	1755-1804; 1751-1836	1787-1788	Federalist Papers	183825	
Waterhouse, Benjamin	1754-1846	1816	A Journal of a Young Man from Massachusetts	106372	
Biggs, William	1755-1827	1788	Indian Captivity of William Biggs	10673	
Marshall, John	1755-1835	1804-1807	The Life of George Washington	121068	
Tyler, Royall	1757-1826	1787	The Contrast	21438	
Webster Foster, Hannah	1758-1840	1797	The Coquette	58999	
Monroe, James	1758-1831	1817-1824	State of the Union Address	42400	
Weems, Parson	1759-1825	1805	Weems' Life of General Francis Marion	82258	
Daggett, David (Jonathan Steadfast)	1764-1851	1804	Count the Cost	11009	
James, William Dobein	1764-1830	1821	A Sketch of the Life of Brig. Gen. Francis Marion	62500	
Low, Samuel	1765-?	1788	The Politician Out-Witted	25753	
Dunlap, William	1766-1839	1798	André	14846	
Jackson, Andrew	1767-1845	1829-1836	State of the Union Address	85257	
Harris, Thaddeus Mason	1768-1842	1841	Biographical Memorials of James Oglethorpe	96251	
Brown, Charles Brockden	1771-1810	1798	Wieland; or the Transformation	82682	165794
			Jane Talbot	83112	

Dodge, David Low	1774-1852	1812	War Inconsistent with the Religion of Jesus Christ	52289	
Meriwether, Lewis; Clark, William	1774-1809; 1770-1838	1804	History of the Expedition under the Command of Captains Lewis and Clark, vol. 1	177581	
Tucker, George (Joseph Atterley)	1775-1861	1827	Voyage to the Moon	55992	
Dunham, Jacob	1779-?	1850	Journal of Voyages	62688	
Webster, Daniel	1782-1852	1817-1845	Select Speeches of Daniel Webster	121161	
Irving, Washington	1783-1859	1836	Astoria	161117	
Crockett, David	1786-1836	1834	Narrative of the Life of David Crockett	53788	
Grandy, Moses	1786?-?	1842	Narrative of the Life of Moses Grandy	13907	
English, George Bethune	1787-1828	1813	The Grounds of Christianity	77560	158026
		1822	A Narrative of the Expedition to Dongola and Sennaar	38938	
		1824	Five Pebbles from the Brook	41528	
Leslie, Eliza	1787-1858	1833	Pencil Sketches	173598	
Seaver, James E.	1787-1827	1824	A Narrative of the Life of Mrs. Mary Jemison	49188	
Cooper, James Fenimore	1789-1851	1826	The Last of the Mohicans	147644	324934
		1840	Pathfinder	177290	
Fisk, Willbur	1792-1839	1835	Calvinistic Controversy	80798	
Withers, Alexander Scott	1792-1865	1831	Chronicles of Border Warfare	147303	

Goodrich, Samuel Griswold	1793-1860	1844	Lives of Celebrated Women	94337	
Schoolcraft, Henry Rowe	1793-1864	1839	Algic Researches vol. 1	53481	105856
		1839	Algic Researches vol. 2	52371	
Bryant, William Cullen	1794-1878	1834-1850	Letters of a Traveller	105722	
Drake, Benjamin	1795-1841	1838	The Great Indian Chief of the West	71466	155896
		1841	Life of Tecumseh, and of His Brother the Prophet	84430	
Kennedy, John P.	1795-1870	1838	Rob of the Bowl	69745	
Ames, Nathaniel	1796-1835	1835	An Old Sailor's Yarns	84541	
Prescott, William H.	1796-1859	1837	History of the Reign of Ferdinand and Isabella, the Catholic, vol. 1	152918	
Ware, William	1797-1852	1836	Zenobia	158593	
Alcott, William Andrus	1798-1859	1835	The Young Man's Guide	85939	158199
		1836	The Young Mother	72260	
Beecher, Catherine Esther	1800-1878	1842	A Treatise on Domestic Economy	126969	
Seward, William H.	1801-1872	1849	Life and Public Services of John Quincy Adams	109257	
Caruthers, William A.	1802-1846	1834	The Cavaliers of Virginia, vol. 1	53637	112529
		1834	The Cavaliers of Virginia, vol. 2	58892	
Child, Lydia Maria	1802-1880	1833	An Appeal in Favor of that Class of Americans Called Africans	84883	

Emerson, Ralph Waldo	1803-1882	1841 1850	Essays, First Series Representative Men	74165 57862	132027
Hawthorne, Nathaniel	1804-1864	1837	Twice-told Tales	146175	
Bird, Robert M.	1806-1854	1837	Nick of the Woods	132653	
Spooner, Lysander	1808-1887	1845	The Unconstitutionality of Slavery	55065	
Poe, Edgar Allan	1809-1849	1833-44	The Works of Edgar Allan Poe — Volume 1	81434	174249
		1834-49	The Works of Edgar Allan Poe — Volume 2	92815	
Ingraham, Joseph Holt	1809-1860	1839 1839	Captain Kyd, vol. 1 Captain Kyd, vol. 2	69772 69600	139372
Fuller, Margaret	1810-1850	1843-1850	At Home and Abroad	164994	
Mayhew, Ira	1814-1894	1850	Popular Education	145712	
Bibb, Henry	1815-1854	1849	Narrative of the Life and Adventures of Henry Bibb, an American Slave	53062	
Halleck, Henry Wager	1815-1872	1846	Elements of Military Art and Science	139888	
Eastman, Mary	1818-1887	1849	Dacotah	77094	
Melville, Herman	1819-1891	1850	White-Jacket	139497	
TOTAL				**5740042**	

CEAL3 (1851-1920)

Author	Birth-Death	Year	Title	Word count	Words/ author
Adams, William Taylor (Oliver Optic)	1822-1897	1854	The Boat Club	46104	233016
		1856	Now or Never	47660	
		1858	Poor and Proud	47669	
		1895	Across India	91583	
Coffin, Charles Carleton	1823-1896	1887	My Days and Nights on the Battle-field	65736	163939
		1895	Daughters of the Revolution and their Times	98203	
Curtis, George William	1824-1892	1853	The Ptiphar Papers	44273	88625
		1856	Prue and I	44352	
Warner, Charles Dudley	1829-1900	1872	Backlog Studies	52034	138242
		1889	That Fortune	86208	
Jackson, Helen Hunt	1830-1885	1884	Ramona	153588	
Davis, Rebecca Harding	1831-1910	1861	Life in the Iron-Mills	14707	109268
		1861	Margret Howth	58191	
	1897		Frances Waldeaux	36370	
Perry, Nora	1831-1896	1894	Hope Benham	64404	124772
		1895	A Flock of Girls and Boys	60368	
Alcott, Louisa May	1832-1888	1868-1869	Little Women	186064	
Alger, Horatio Jr.	1832-1899	1866	Timothy Crump's Ward	40273	85532
		1887	The Store Boy	45259	
Talmage, Thomas De Witt	1832-1902	1872	The Abominations of Modern Society	53106	140140
		1895	Around the Tea-Table	87034	
Twain, Mark (Samuel Clemens)	1835-1910	1876	The Adventures of Tom Sawyer	70800	
Roe, Edward Payson	1838-1888	1872	Barriers Burned Away	142130	

Bierce, Ambroce Gwinnett	1842-1914	1874	Cobwebs from an Empty Skull	52941	113005
		1893	Can Such Things Be?	60064	
James, Henry	1843-1916	1875	Roderick Hudson	131723	186074
		1911	The Outcry	54351	
Jewett, Sarah Orne	1849-1909	1879	Old Friends and New	58356	157031
		1890	Betty Leicester	56619	
		1896	The Country of the Pointed Firs	42056	
Bellamy, Edward	1850-1898	1887	Looking Backward	77836	183435
		1900	The Duke of Stockbridge	105599	
Chopin, Kate	1850-1904	1890	At Fault	57569	121562
		1893-1899	The Awakening and Selected Short Stories	63993	
Marden, Orison Swett	1850-1924	1896 1897	How to Succeed Architects of Fate	75355 100423	246016
		1916	The Victorious Attitude	70238	
Freeman, Mary E. Wilkins	1852-1930	1894 1900	Pembroke The Heart's High-way	79906 74315	154221
van Dyke, Henry	1852-1933	1895	Little Rivers	57968	169603
		1901	The Ruling Passion	54865	
		1907	Days Off	56770	
Rathborne, St. George	1854-1938	1893 1912	Miss Caprice Canoe Mates in Canada	55230 48456	151409
		1912	Chums in Dixie	47723	
Washington, Booker T.	1856-1915	1899	The Future of the American Negro	38334	114961
		1901	Up from Slavery: an Autobiography	76627	

Gilman, Charlotte Perkins	1860-1935	1911 1915	The Crux Herland	52556 52145	104701
Ottolengui, Rodrigues	1861-1937	1892 1898	An Artist in Crime Final Proof	71380 97466	168846
Porter, William Sydney (O. Henry)	1862-1910	1904 1907	Cabbages and Kings Heart of the West	63099 77807	140906
Stratton-Porter, Gene	1863-1924	1915	Michael O'Halloran	145728	
Davis, Richard Harding	1864-1916	1891 1902 1910 1916	Gallegher and Other Stories Ranson's Folly Once upon a Time The Man that Could Not Lose	47779 74271 54586 12917	189553
Phillips, David Graham	1867-1911	1904 1912	The Cost The Price She Paid	75487 96678	172165
Porter, Eleanor H.	1868-1920	1908 1916	The Turn of the Tide Just David	54332 59044	113376
Tarkington, Booth	1869-1946	1918	The Magnificent Ambersons	99511	
Norris, Frank	1870-1902	1899 1903	McTeague A Deal in Wheat	112623 48832	162922
Rice, Alice Hegan	1870-1942	1903 1909 1917	Lovey Mary Mr. Opp Calvary Alley	21758 50465 89119	161455
Churchill, Winston	1871-1947	1917	The Dwelling-Place of Light	138569	
Crane, Stephen	1871-1900	1895 1899 1899	The Red Badge of Courage Active Service The Monster and Other Stories	46203 79739 35747	161689
Dreiser, Theodore	1871-1945	1900	Sister Carrie	156282	

MacGrath, Harold	1871-1932	1899	Arms and the Woman	70704	107107
		1915	The Voice in the Fog	36403	
Grey, Zane	1872-1939	1906	The Spirit of the Border	88039	191736
		1912	Riders of the Purple Sage	103697	
Wright, Harold Bell	1872-1944	1907	The Shepherd of the Hills	74842	142190
		1919	The Re-Creation of Brian Kent	67348	
Cather, Willa	1873-1947	1913	O Pioneers!	55435	136755
		1918	My Antonia	81320	
Palmer, Frederick	1873-1958	1912	Over the Pass	115400	
London, Jack	1876-1916	1904	The Sea-Wolf	105651	177715
		1906	White Fang	72064	
Rinehart, Mary Roberts	1876-1958	1906	The Man in Lower Ten	64452	183728
		1919	Dangerous Days	119276	
Norris, Kathleen	1880-1966	1917	Martie the Unconquered	117605	
TOTAL				**6319792**	

3 Evaluation

In this section we will briefly assess CEAL in terms of its contents and, to some extent, usefulness for linguistic research. One challenge in the evaluation of CEAL is that there is no comparable corpus that could act as a point of reference. This means that, in order to evaluate the quality of CEAL, we can either compare it to corpora from other time periods and varieties (Section 3.1), explore its internal composition by cross-comparing the subcorpora (Section 3.2) or use it to study different phenomena in language and thereby indirectly judge its usefulness (Section 3.3). As the present paper is not a research paper but an introductory article to the corpus, the evaluations presented below are fairly straightforward.

3.1 Comparison with other corpora

Evaluating the overall representativeness of CEAL in relation to other corpora is not without difficulties. If the CEAL data looks different from the other corpora, it might be either (a) because AmE 1690–1920 is simply different, or (b) because CEAL is unbalanced and/or biased. In order to minimize this issue when comparing with other corpora, we use corpora that are as close as possible to CEAL time- and variety-wise, and/or parameters that have been observed to remain more or less stable across time periods and varieties. The concordancer that was used for the searches was the freely available AntConc (Anthony 2014).

First, the most frequent words in CEAL were compared with the most frequent words in the CLMETEV, and the Brown and Frown corpora. The CLMETEV was chosen because it covers the same time span as CEAL, and Brown and Frown were chosen because they represent American English. The results are presented in Table 3:

Table 3: Twenty most frequent words in CEAL, the CLMETEV, and Brown + Frown

	CEAL		**CLMETEV**		**Brown + Frown**	
rank	**word**	**freq. pmw**	**word**	**freq. pmw**	**word**	**freq. pmw**
1	the	64545	the	59588	the	66186
2	of	35847	of	35863	of	34374
3	and	33202	and	33068	and	28471
4	to	28518	to	28845	to	26220
5	a	21654	a	20958	a	22605
6	in	17896	in	17823	in	20898
7	i	13337	i	14883	that	11411
8	that	12804	that	12669	is	9721
9	he	11368	it	11686	for	9366
10	it	11248	was	10801	was	8987
11	was	10919	he	10615	he	8787
12	his	9197	his	8870	it	8648
13	with	8275	as	8358	as	7323
14	as	7931	is	8244	with	7163
15	you	7677	with	8205	on	6833

16	for	7589	for	7987	his	6379
17	is	7300	you	7195	be	5785
18	had	6618	not	6894	by	5252
19	her	6433	be	6833	at	5113
20	be	6410	had	6564	i	4828

As can be seen in Table 3, the lists are generally very similar, sharing most of the words, and with the six most frequent tokens given in the same order. The proportions seem to be similar as well. Of course it has to be borne in mind that Brown and Frown include a variety of text types, while CEAL and the CLMETEV mostly cover fiction, personal letters and so forth, which, for instance, explains the relatively high frequency of the first person singular pronoun *I* in them. Nevertheless, these results indicate that CEAL, as a whole, consists of what might be referred to as 'typical' language, without any major divergences.

3.2 Comparison across the subcorpora

In addition to the overall word frequency, we wanted to see the situation in the three subcorpora. The wordlists from CEAL1, CEAL2 and CEAL3 are presented in Table 4:

Table 4: most frequent words in CEAL subcorpora

rank	**CEAL1**	**CEAL2**	**CEAL3**
1	the	the	the
2	of	of	and
3	and	and	to
4	to	to	of
5	in	a	a
6	a	in	i
7	that	that	in
8	i	it	he
9	it	was	was
10	be	i	that
11	is	his	it
12	as	with	you

13	for	is	his
14	have	as	her
15	with	he	she
16	this	for	had
17	by	be	with
18	they	by	for
19	their	which	as
20	not	on	but

Table 4 displays more variation in the top-20 lists of words than Table 3, which might be expected, considering the different proportions of fiction and non-fiction texts in the three parts of the corpus. Nevertheless, the lists feature mostly the same words as the lists in Table 3, which is again seen as a sign of what might be considered 'typical' language.

3.3 Practical application

The corpus has already been piloted in linguistic research; Höglund (2014) used the work-in-progress version of CEAL in his doctoral dissertation to study the development of the *tough* construction (*John is easy to please*) in the past three centuries. In Höglund's study, CEAL1, with its modest size of 1.5M words, proved to be too small to investigate the fairly infrequent *tough* construction, but the larger CEAL2 and CEAL3 provided enough material for the research. The CEAL data was compared to the data from the CLMETEV, and the comparison showed quite nicely, for instance, that the *tough* construction began to be used more first in BrE around the early 19th century, and that AmE followed the trend in the late 19th century. In addition, the passive *tough* construction (*John is easy to be pleased*) was observed to follow the same pattern but in the other direction: it was falling into disfavor earlier in BrE than in AmE (Höglund 2014: 76–77). This type of change seems to be in accordance with earlier observations about 'colonial lag' (see Rohdenburg and Schlüter (2009), especially chapters 1 by Hundt, 5 by Schlüter, and 19 by Rohdenburg and Schlüter), and would thus validate the reliability of the results and indicate that CEAL represents AmE quite well at least in this respect. However, more studies of different types are of course needed in order to properly evaluate the usefulness of CEAL as a linguistic resource.

4 Conclusion

The present paper has introduced a new resource for linguistic research that has not been previously available: the *Corpus of Early American Literature* (CEAL). Before CEAL, corpus material from the 18th century written on the new continent was hardly available. The corpus covers a little over 200 years of texts, a period that includes the formation in the 18th century of the variety now known as American English, as well as its further development in the 20th century. That is, CEAL makes it possible to trace the development and formation of American English as a variety in its own right. Another advantage of CEAL is that the data can be fairly easily compared to the CLMETEV, which includes British English, as the two corpora are more or less comparable in terms of timespan, size, and compilation criteria.

It is our belief that the vast amount of text material that has been released through online repositories and endeavors for open access is massively beneficial for the study of English, which continuously needs larger and more specialized corpora with which to shed light on linguistic phenomena. All the material made available online has also made the compilation of corpora such as CEAL more accessible for linguists in general. In its current form, CEAL remains a fairly modest corpus when it comes to size, but one which helps form a more complete picture of the history and development of American English.

Availability and disclaimer

The corpus introduced in this article, CEAL, can be compiled by anyone by using the instructions and the text repositories mentioned, and downloading the texts in the content list. CEAL can also be obtained more easily by contacting one of the compilers. The usual disclaimers apply: the compilers are not responsible for the accuracy of the individual texts, and it is solely the responsibility of the user of the corpus to evaluate the authenticity of the texts and the suitability of the corpus for her/his research.

Notes

1. The project was related to and a part of Höglund's doctoral thesis (2014), which was funded by the Finnish Academy of Science and Letters, Jutikkala Fund. The compilers of the corpus are grateful for the comments received from Maarit Piipponen, David Robertson and Markku Salmela, of the University of Tampere, on the contents of the corpus. The final deci-

sions on what to include in the corpus, and all the possible mistakes therein, are solely the authors' responsibility.

2. A more detailed table with notes on the texts and sources is available from the compilers.

References

Anthony, Laurence. 2014. AntConc (Version 3.4.4) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/.

Davies, Mark. 2010–. *The Corpus of Historical American English: 400 million words, 1810–2009*. Available online at http://corpus.byu.edu/coha/.

De Smet, Hendrik. 2005. A Corpus of Late Modern English. *ICAME Journal* 29: 69–82.

Höglund, Mikko. 2014. "Self-discipline strategies were easy to design but difficult to adhere to." A usage-based study of the Tough Construction in English. Doctoral dissertation, University of Tampere.

Höglund, Mikko and Kaj Syrjänen. 2010. Towards a corpus of early American literature: On the challenges of compiling a comparable diachronic corpus. In I. Moskowich-Spiegel Fandiño, B. Crespo García, I. Lareo Martín and P. Lojo Sandino (eds.). *Language windowing through corpora*, 429–442. A Coruña: Universidade da Coruña.

Kučera, Henry and W. Nelson Francis. 1967. *Computational analysis of Present-day American English*. Providence, RI: Brown University Press.

Kytö, Merja. 1994. Towards a corpus of early American English. In M. Kytö, M. Rissanen and S. Wright (eds.). *Corpora across the centuries*, 33–39. Amsterdam and Atlanta: Rodopi.

Rohdenburg, Günter and Julia Schlüter (eds.). 2009. *One language, two grammars? Differences between British and American English*. Cambridge: Cambridge University Press.