A corpus of Late Modern English texts*

Hendrik De Smet
University of Leuven

1 Introduction

It has on occasion been observed that the Late Modern English period is the most neglected period in the history of the English language (Rydén 1984; Denison 1998: 92). Interestingly, however, this is not only true as far as descriptive efforts are concerned, but also at the methodological basis of linguistic research. Symptomatic of a certain neglect of anything beyond the 17th century is the fact that the *Helsinki Corpus*, until now the most important electronic corpus for the study of the history of English, takes its final cut-off point in 1710. The apparent neglect is, in a way, surprising, since the Late Modern English period is a very well-documented one, and is much more easily accessible to the speaker of Present-Day English than – say – the Middle English period. It is only natural that more recent corpora have begun to fill the gap between Early Modern English and the present day, especially as it has become increasingly clear that historical change can often be tracked over relatively short time spans in the form of shifting frequencies of use (see e.g. Mair 2000; Nevalainen and Raumolin-Brunberg 2003). Thus, the *Lampeter Corpus* covers the transition from Early to Late Modern English (Siemund and Claridge 1997); the *ARCHER Corpus* covers the entire period from Late Modern to Present-Day English (Biber et al. 1994); the *Corpus of Late Modern English Prose* is representative of the latter half of the 19th and the beginning of the 20th centuries (Denison 1994); and more corpora could be added to this list.

The purpose of the present paper is to contribute to the study of Late Modern English by exploring an additional means of gathering and investigating Late Modern English language data. In particular, large amounts of Late Modern English data are available on the World Wide Web through, for instance, the *Project Gutenberg* or the *Oxford Text Archive*. The texts are often in the public domain and can, therefore, be freely downloaded and used for all kinds of non-commercial purposes, including linguistic ones. In this paper, I present a corpus of Late Modern English, compiled on the basis of texts drawn from the *Project*

Gutenberg and the *Oxford Text Archive*. For ease of reference, I will refer to the corpus as the *Corpus of Late Modern English Texts* (CLMET), but the reader should be reminded that the corpus is not exactly a fixed body of texts in the same way conventional corpora of English are; the corpus can be extended or reduced at wish, and similar – though not necessarily identical – corpora can be compiled without much effort by anyone who is interested in the study of Late Modern English. The corpus presented here is what I consider an acceptable and useful offshoot of a continual attempt to open up the rich resources of the Internet to historical linguistic research.

In what follows, I will discuss the make-up of the corpus as it has been compiled by myself (section 2); discuss some of its advantages and disadvantages (section 3); and briefly illustrate the potential of the corpus by surveying some of the research in which it has already been used (section 4).

2 Corpus make-up

The CLMET has been entirely compiled on the basis of texts from the *Project Gutenberg* and the *Oxford Text Archive* and covers the period from 1710 to 1920. It is subdivided into three sub-periods of 70 years each, i.e. 1710–1780; 1780–1850; and 1850–1920. On the notion that a corpus is a principled collection of texts (Sinclair 1992), the process of data collection has been guided by four principles.

First, the texts included within one sub-period of the CLMET are written by authors born within a correspondingly restricted time-span. This is schematically represented in Figure 1. The purpose of this measure is to increase the homogeneity within each sub-period – and accordingly, to decrease the homogeneity between the sub-periods. Historical trends should, as a result, appear somewhat more clearly. An additional advantage is that no author can be represented in two subsequent sub-periods of the corpus. A slight disadvantage is that the work of some authors is lost for inclusion in the corpus. To give an example, by birth the Victorian novelist George Eliot (1819-1880) belongs to the second sub-period of the corpus, but because all of her work falls within the third sub-period of the corpus by its date of publication, none of it could be excerpted for the corpus.

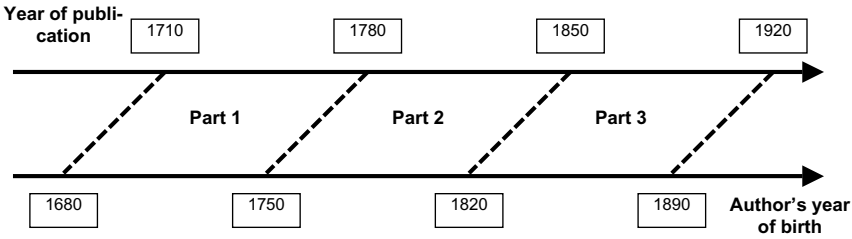

Figure 1: Corpus sub-periods

Second, all authors are British and are native speakers of English. The purpose of this measure is evident: it puts some (moderate) restriction on dialectal variation. The specific choice for British authors should facilitate comparison of the data from the CLMET to data from other historical corpora and from the large corpora of Present-Day English, which are mostly corpora of British English. Nevertheless, it should be pointed out here that the Internet could be used as a rich resource for other varieties of English as well, especially American English.

Third, any one author can only contribute a restricted amount of text to the corpus. The idea is, obviously, to avoid thwarting of the data by the idiosyncrasies of individual authors. The maximum amount of text per author is 200,000 words. This may seem a rather liberal cut-off point when compared to the maximum of 10,000 words per text in the *Helsinki Corpus* (Kytö 1996), but it should be pointed out that the problem of idiosyncratic language use is also counteracted by excerpting a large variety of authors, especially if all authors provide roughly the same amount of text. In that respect, the cut-off point could be laid at 200,000 words per author, because for many Late Modern English authors at least half of that amount of text is fairly easily available – especially for the second and third sub-period of the corpus.

Fourth, some attention has gone to insuring variation in terms of text genre and authorial social background. The texts found on the *Project Gutenberg* and the *Oxford Text Archive* have been collected and made publicly accessible on the Internet for other reasons than their linguistic interest, and are, partly as a result of that, typically literary, formal texts, mostly written by men who belonged to the better-off layers of 18th and 19th century English society. To counteract this bias, I have deliberately favoured non-literary texts over literary ones and texts from lower registers over texts from higher registers, whenever a choice could be made among the texts produced by a particular author. Further, I have paid

some special attention to including texts written by women authors. However, in spite of these efforts, it will be evident that the corpus continues to be biased to literary texts written by higher class male adults.

The application of the four principles just described has yielded the list of texts that is rendered in Table 1, and that constitutes the CLMET as it stands today. Table 1 specifies for each sub-period the authors, the amount of text they contribute, the specific works used, and their date of publication. The indication '(s)' signals that only part of a particular work has been selected for inclusion in the corpus.

Table 1: Contents of the CLMET

Author	**Title and year of first publication**		**No. of words**
Gay, John (1685–1732)	1728	*The Beggar's Opera*	17,427
Pope, Alexander (1688–1744)	1733–34	*An Essay on Man*	46,995
Chesterfield, Philip Dormer Stanhope (1694–1773)	1746–71	*Letters to his Son* (s)	199,819
Fielding, Henry (1707–54)	1749	*The History of Tom Jones, a Foundling* (s)	100,242
—	1751	*Amelia* (s)	99,569
Johnson, Samuel (1709–84)	1740–41	*Parliamentary Debates* (Vol. 1) (s)	163,695
—	1759	*Rasselas, Prince of Abyssinia*	37,070
Fielding, Sarah (1710–68)	1749	*The Governess; or, The Little Female Academy*	50,708
Hume, David (1711–76)	1739–40	*A Treatise of Human Nature* (s)	113,935
—	1751	*An Enquiry Concerning the Principles of Morals*	48,245
—	1779	*Dialogues Concerning Natural Religion*	35,972

Sterne, Laurence (1713–68)	1759–67	*The Life and Opinions of Tristram Shandy* (s)	158,135
—	1768	*A Sentimental Journey through France and Italy*	42,249
Walpole, Horace (1717–97)	1735–48	*Letters* (Vol. 1) (s)	162,799
—	1764	*The Castle of Otranto*	36,171
Smollett, Tobias George (1721–71)	1751	*The Adventures of Peregrine Pickle* (s)	99,421
—	1771	*The Expedition of Humphrey Clinker* (s)	100,675
Smith, Adam (1723–90)	1766	*An Inquiry into the Nature and Causes of the Wealth of Nations* (s)	200,667
Reynolds, Joshua (1723–92)	1769–76	*Seven Discourses on Art*	39,563
Burke, Edmund (1729–97)	1770	*Thoughts on the Present Discontents*	30,386
—	1775	*On Conciliation with America*	26,883
Goldsmith, Oliver (1728–74)	1766	*The Vicar of Wakefield*	63,730
—	1773	*She Stoops to Conquer*	22,962
Gibbon, Edward (1737–94)	1776	*The Decline and Fall of the Roman Empire* (Vol. 1) (s)	199,087
TOTAL 1710–1780			2,096,405

Inchbald, Elisabeth (1753–1821)	1796	*Nature and Art*	47,126
Burns, Robert (1759–96)	1780–96	*The Letters of Robert Burns*	124,247
Wollstonecraft, Mary (1759–97)	1792	*Vindication on the Rights of Woman*	86,670
—	1796	*Letters on Norway, Sweden, and Denmark*	48,219
—	1798	*Maria*	45,428

Beckford, William (1760–1844)	1783	*Dreams, Waking Thoughts, and Incidents*	80,746
Malthus, Thomas (1766–1834)	1798	*An Essay on the Principle of Population*	54,451
Edgeworth, Maria (1767–1849)	1796–1801	*The Parent's Assistant*	168,182
Hogg, James (1770–1835)	1824	*The Private Memoirs and Confessions of a Justified Sinner*	84,166
Owen, Robert (1771–1858)	1813	*A New View of Society*	34,124
Southey, Robert (1774–1843)	1813	*Life of Horatio Lord Nelson*	96,781
—	1829	*Sir Thomas More*	39,124
Austen, Jane (1775–1817)	1796–1817	*Letters to her Sister Cassandra and Others* (s)	77,989
—	1811	*Sense and Sensibility* (s)	61,546
—	1813	*Pride and Prejudice* (s)	60,141
Lamb, Charles (1775–1834)	1807	*Tales from Shakespeare*	100,349
—	1808	*Adventures of Ulysses*	33,727
Smith, James (1775–1839), and Horace Smith (1779–1849)	1812	*Rejected Addresses*	28,759
Hazlitt, William (1778–1830)	1821–22	*Table Talk*	160,700
—	1823	*Liber Amoris*	30,911
Galt, John (1779–1839)	1821	*The Ayrshire Legatees*	50,072
—	1821	*Annals of the Parish*	65,613
De Quincey, Thomas (1785–1859)	1822	*Confessions of an English Opium-Eater*	38,839
Byron, George Gordon (1788–1824)	1810–13	*Letters 1810–1813*	110,243
Marryat, Frederick (1792–1848)	1841	*Masterman Ready*	99,705

Carlyle, Thomas (1795–1881)	1837	*The French Revolution* (s)	200,251
Shelly, Mary Wollstonecraft (1797–1851)	1818	*Frankenstein*	75,082
Bulwer-Lytton, Edward (1803–73)	1834	*The Last Days of Pompeii*	151,692
Borrow, George Henry (1803–81)	1842	*The Bible in Spain* (s)	199,199
Ainsworth, William Harrison (1805–82)	1843	*Windsor Castle*	117,072
Darwin, Charles (1809–82)	1839	*The Voyage of the Beagle* (s)	199,777
Kinglake, William (1809–91)	1844	*Eothen, or Traces of Travel Brought Home from the East*	89,058
Gaskell, Elizabeth (1810–65)	1848	*Mary Barton*	160,888
Thackeray, William Makepeace (1811–63)	1847–48	*Vanity Fair* (s)	200,907
Dickens, Charles (1812–70)	1841	*Barnaby Rudge* (s)	78,226
—	1843	*A Christmas Carol in Prose*	28,673
—	1848	*Dombey and Son* (s)	93,352
Brontë, Emily (1818–48)	1847	*Wuthering Heights*	116,760
Brontë, Anne (1820–49)	1847	*Agnes Grey* (s)	50,133
—	1848	*The Tenant of Wildfell Hall* (s)	150,730
TOTAL 1780–1850			3,739,657

Hughes, Thomas (1822–96)	1857	*Tom Brown's Schooldays*	105,982
Freeman, Edward Augustus (1823–92)	1888	*William the Conqueror*	57,067
Yonge, Charlotte Mary (1823–1901)	1873	*Young Folk's History of England* (s)	51,339
—	1865	*The Clever Woman of the Family* (s)	74,807

—	1870	*The Caged Lion* (s)	77,241
Collins, William Wilkie (1824–89)	1859–60	*The Woman in White* (s)	96,398
—	1868	*The Moonstone* (s)	101,932
Huxley, Thomas Henry (1825–95)	1894	*Discourses*	95,883
Blackmore, Richard Doddridge (1825–1900)	1869	*Lorna Doone, A Romance of Exmoor* (s)	202,593
Bagehott, Walter (1826–77)	1867	*The English Constitution*	97,933
—	1869	*Physics and Politics*	56,554
Meredith, George (1828–1909)	1870	*The Adventures of Harry Richmond* (s)	97,677
—	1895	*The Amazing Marriage* (s)	98,235
Booth, William (1829–1912)	1890	*In Darkest England and the Way out*	126,065
Rutherford, Mark (1831–1913)	1893	*Catherine Furze*	67,367
—	1896	*Clara Hopgood*	48,987
Carroll, Lewis (1832–98)	1865	*Alice's Adventures in Wonderland*	26,699
—	1871	*Through the Looking Glass*	29,639
—	1889	*Sylvie and Bruno*	65,579
Butler, Samuel (1835–1902)	1880	*Unconscious Memory* (s)	51,231
—	1903	*The Way of All Flesh* (s)	74,069
—	1912	*Note-Books* (s)	76,734
Abbott, Edwin (1838–1926)	1884	*Flatland*	33,805
Pater, Walter Horatio (1839–94)	1885	*Marius the Epicurean* (Vol. 1)	56,847
—	1886–1890	*Essays from 'The Guardian'*	24,020

—	1896	*Gaston de Latour, An Unfinished Romance*	38,212
Hardy, Thomas (1840–1928)	1873	*A Pair of Blue Eyes* (s)	101,665
—	1874	*Far from the Madding Crowd* (s)	100,100
Grossmith, George (1847–1912), and Weedon Grossmith (1852–1919)	1894	*The Diary of a Nobody*	42,276
Gosse, William Edmund (1849–1928)	1907	*Father and Son, A Study of Two Temperaments*	79,185
Haggard, Henry Rider (1856–1925)	1887	*She*	111,944
Gissing, George (1857–1903)	1891	*New Grub Street* (s)	94,810
—	1893	*The Odd Woman* (s)	101,691
Jerome, Jerome K. (1859–1927)	1889	*Three Men in a Boat*	67,445
—	1909	*They and I*	70,125
Hope, Anthony (1863–1933)	1894	*The Prisoner of Zenda*	54,157
—	1898	*Rupert of Hentzau*	83,351
Kipling, Rudyard (1865–1936)	1894	*The Jungle Book*	51,162
—	1897	*Captains Courageous*	53,452
Wells, Herbert George (1866–1946)	1888	*The Time Machine*	32,507
—	1897	*The War of the Worlds*	60,308
—	1902–03	*Mankind in the Making*	103,549
Bennett, Arnold (1867–1931)	1902	*The Grand Babylon Hotel* (s)	51,852
—	1908	*The Old Wives' Tale* (s)	149,599

Galsworthy, John (1867–1933)	1904	*The Island Pharisees*	70,492
—	1906	*The Man of Property*	110,623
Churchill, Winston (1874–1965)	1899	*The River War, An Account of the Reconquest of the Sudan*	126,807
Chesterton, Gilbert Keith (1874–1936)	1912	*What's Wrong with the World*	60,318
—	1914	*The Wisdom of Father Brown*	71,935
Forster, Edward Morgan (1879–1970)	1905	*Where Angels Fear to Tread*	49,988
—	1908	*A Room with a View* (s)	49,518
—	1910	*Howards End* (s)	100,510
TOTAL 1850–1920			3,982,264

3 Advantages and disadvantages

In addition to being freely available, I believe the corpus outlined above has two main advantages. First, the corpus is highly manipulable; texts can be added to or excluded from the corpus, or can be expanded or reduced in size with a simple text browser – all at wish. The most important consequence of this is that the corpus can continue to grow, as new texts are drawn from the Internet. Second, the corpus is fairly large. As shown in the previous section, it comprises slightly less than ten million words. This means that in terms of size the CLMET belongs somewhere in between the traditionally small historical corpora of English, such as the *Helsinki Corpus*, and the synchronic 'monster' corpora of Present-Day English, such as the *British National Corpus*. Consequently, while it is presumably too small for lexicographic purposes, the corpus is large enough for the study of relatively infrequent syntactic patterns, or borderline phenomena between grammar and the lexicon, such as lexico-grammatical patterning, grammaticalisation, and lexicalisation – all of which are of interest in current linguistic theory.

At the same time, it is important to recognise some of the disadvantages of the corpus. One problem is that the corpus make-up is evidently not ideal. As already remarked above, the corpus is biased both sociolinguistically and in terms of genre and register, which makes it unfit for any fine-grained sociolinguistic analysis. However, as long as a sociolinguistic analysis is not the purpose

of one's research, this may not be a fundamental problem, if (and only if) the sociolinguistic make-up of the corpus remains more or less consistent over the different sub-periods – which seems to be the case for the CLMET. In addition, if the corpus is further extended, it may, among other things, become possible to make diachronic comparisons between British and American English, so that a coarse kind of sociolinguistic research comes within the range of what the corpus can do. Against this optimism it must be pointed out that, although a sociolinguistic bias is, perhaps, not a problem as such, the particular tendency for the CLMET to be largley made up of formal writings by highly schooled (and linguistically self-conscious) authors is unfortunate, because these are exactly the type of texts where one expects language change to be kept at a tight leash.

Another, rather different problem of the CLMET is that the exact bibliographical history of the corpus texts is often highly unclear. Internet sources tend to provide no specification as to which version of a text lies at the basis of its electronic edition, who the intermediate editors have been, and what they might have done to the original text. It is clear from occasional editorial footnotes and modernised spellings that the texts scanned in for electronic publication are often late 19th or early 20th century editions of earlier prints or manuscripts. For this reason, the corpus had better not be used for the study of phenomena that might lightly attract editorial interventions – for example, matters of punctuation, spelling-related issues such as the alternation between *a* and *an* in the indefinite pronoun, or anything that might be seen by an editor as a production error. On the other hand, it seems unlikely that an editor should introduce radically new constructions into a text – for instance, a finite instead of a non-finite clause – or that editorial intervention could have any bearing on the timing of semantic developments within specific words or constructions.

4 Research

Eventually, the value of a corpus is measured by what it can do. In this respect, it is useful to briefly discuss some of the research in which the CLMET described in this paper has been or is being used. It must be added that in most cases the data drawn from the corpus have been complemented with data from other, conventional corpora, or from the *Oxford English dictionary.* As will be clear from the following survey, the CLMET has so far been mainly, and most successfully, used in studies involving qualitative change in the history of English, and has been less extensively 'tested' when it comes to quantitative studies of language change.

De Smet (2005) and De Smet and Cuyckens (2004; forthcoming) have used a slightly extended version of the CLMET to investigate changes in the English system of verbal complementation. These include semantic changes, such as the semantic development of the construction '*like* + *to*-infinitive' from a volitional to a habitual construction; and syntactic changes, such as the emergence and spread of *for…to*-infinitives from Early Modern to Present-Day English. They have also used the corpus to study the impact of entrenchment or routinisation on the long-standing competition between infinitives and gerunds as verbal complements in English.

Breban (forthcoming) has made use of the CLMET in her work on adjectives of comparison such as *similar, comparable, other, different*, etc. In particular, she has used the corpus to document changes in the function these adjectives fulfil within the noun phrase, tracking developments from more lexical attribute uses to more grammatical post-determiner and classifier uses.

Vanden Eynde (2004), finally, has used data from the CLMET to investigate historical developments in so-called *edge*-noun constructions. Such constructions – e.g. *on the edge of, on the verge of, on the brink of* – show a trend to develop from purely lexical constructions indicating location at the edge of something to aspectual constructions expressing the imminent occurrence of an event.

5 Conclusion

The study of the history of the English language can, I believe, only benefit from exploiting the extensive amounts of Late Modern English data available from Internet sources such as the *Project Gutenberg* or the *Oxford Text Archive*. In this paper I have therefore proposed a more systematic, or principled, way of doing so, offering a first version of a corpus of Late Modern English based entirely on material freely available from the Internet. It is evident that the corpus described in the preceding sections has its disadvantages, and, in many respects, it cannot stand the comparison with some of the so-called 'small but beautiful' corpora already available for the study of the history of English. On the other hand, given its size, the corpus may still complement the smaller corpora. As pointed out above, the corpus lends itself best for the study of lexico-grammatical phenomena that are somewhat less frequent, and for which smaller corpora tend not to provide sufficient data. In this sense, the corpus could be seen as an electronic counterpart to the vast quotation databases used by the traditional grammarians of the early 20th century. It is hoped, in any case, that a

more systematic use of Internet data could further the study of the allegedly most neglected period in the history of the English language.

Availability

Anyone interested can compile a corpus from the texts available through the *Project Gutenberg*, the *Oxford Text Archive*, or any other electronic archiving project whose texts are publicly accessible. They can, obviously, follow the principles outlined here, or choose to apply a different set of principles. However, the version of the CLMET described in this paper can also be obtained in a less time-consuming manner, by contacting the author of the paper.

Hendrik De Smet
Department of Linguistics
Blijde-Inkomststraat 21
B-3000 Leuven, Belgium
Tel.: 0032 16 32 47 72
hendrik.desmet@arts.kuleuven.ac.be

Note

* I gratefully acknowledge the financial support from the University of Leuven for the project (OT/2003/20/TBA) that allows me to work as a researcher at the Department of Linguistics. I would also like to thank those who have shown interest in my work and have commented on it: the members of the Functional Linguistics research unit at the linguistics department of the University of Leuven – in particular, Hubert Cuyckens – and the participants of the 25th ICAME Conference in Verona in May 2004.

References

Biber, Douglas, Edward Finegan and Dwight Atkinson. 1994. ARCHER and its challenges. Compiling and exploring a representative corpus of historical English registers. In U. Fries, G. Tottie, and P. Schneider (eds.). *Creating and using English language corpora*, 1–14. Amsterdam: Rodopi.

Breban, Tine. Forthcoming. The grammaticalization of the English adjectives of comparison. A diachronic case study. To appear in R. Facchinetti and M. Rissanen (eds.). *Corpus linguistic studies in diachronic English*. Bern: Peter Lang.

Denison, David. 1994. A corpus of Late Modern English prose. In M. Kytö, M. Rissanen, and S. Wright (eds.). *Corpora across the centuries. Proceedings of the First International Colloquium on English Diachronic Corpora,* 7–16. Amsterdam: Rodopi.

Denison, David. 1998. Syntax. In S. Romaine (ed.). *The Cambridge history of the English language.* Vol. 4. *1776–1997,* 92–329. Cambridge: Cambridge University Press.

De Smet, Hendrik. 2005. *For...to*-infinitives as verbal complements in Late Modern and Present-Day English. Between motivation and change. Preprint 225. University of Leuven: Department of Linguistics.

De Smet, Hendrik and Hubert Cuyckens. 2004. A diachronic perspective on the variation between gerunds and infinitives as verbal complements. Paper presented at ICAME25 (25th Conference of the International Computer Archive of Modern and Medieval English), Verona (Italy), 19–23 May 2004.

De Smet, Hendrik and Hubert Cuyckens. Forthcoming. Pragmatic strengthening and the meaning of complement constructions. The case of *like* and *love* with the *to*-infinitive. To appear in *Journal of English Linguistics*.

Kytö, Merja. 1996. *Manual to the diachronic part of the Helsinki Corpus of English Texts. Coding conventions and lists of source texts.* University of Helsinki: Department of English.

Mair, Christian. 2002. Three changing patterns of verb complementation in Late Modern English. A real-time study based on matching text corpora. *English Language and Linguistics* 6: 105–131.

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

Rydén, Mats. 1984. The study of eighteenth century English syntax. In J. Fisiak (ed.). *Historical syntax,* 509–520. Berlin: Mouton Publishers.

Siemund, Rainer and Claudia Claridge. 1997. The Lampeter Corpus of Early Modern English Tracts. *ICAME Journal* 21: 61–70.

Sinclair, John. 1992. *Corpus, concordance, collocation.* Oxford: Oxford University Press.

Vanden Eynde, Martine. 2004. Edge-noun expressions as markers of imminence. A case of grammaticalization. Unpublished MA Thesis. University of Gent: Department of Linguistics.

The texts are often in the public domain and can, therefore, be freely downloaded and used for all kinds of non-commercial purposes, including linguistic ones. In this paper, I present a corpus of Late Modern English, compiled on the basis of texts drawn from the Project. Forskjellen mellom syklon og orkan. 69. ICAME Journal No. 29. Gutenberg and the Oxford Text Archive. For ease of reference, I will refer to the corpus as the Corpus of Late Modern English Texts (CLMET), but the reader should be reminded that the corpus is not exactly a fixed body of texts in the same way conventional corpora ... For example, the Corpus of Late Modern English Texts (CLMET) covers the period from 1710 to 1920 and com-prises 333 texts from ve different genres (De Smet et al., 2015). CLMET contains 34M words from British authors and is available under a Creative Commons license. The Corpus of Historical American English (COHA) contains more than 400M words of historical English from the 1810s until the 2000s (Davies, 2010). COHA is balanced by genre and time with the following four genres: Fiction, Magazine, Newspaper, and Non-Fiction.Â The CorunËœa Corpus of English Scientic Writing (Crespo-GarcÂ´Ã±a and Moskowich, 2015) is a collection of text sam-ples representing late Modern English scientic writing ex-cept medical texts. Late Modern English. Method. For the Late Modern period, a search of CLMETEV (De Smet 2005) was undertaken. Because the corpus is not parsed, it is not possible to search for any preposition following what. In order to investigate the development of what with, a number of different searches were undertaken. The first was a simple search for the string "what with". This returned a number of false positives of the type It's poisoned I don't know what with [COCA]; these were discounted. Each instance of the what with construction was classified based on the following complement,[1] as illustrated below: (22) NP complements. The following text gives a full introduction to English sounds, grammar, and vocabulary. It begins with a study of the distinctive sounds of English (phonology). It turns next to an analysis of the structure of English words and their classification (morphology) as well as the classification of English words and their grammatical modification. This is followed by an exploration of the meaning of English words (lexical semantics).Â Note that the source of each citation is designated by the name of the corpus and the genre in which it occurs (e.g. COCA:FIC designates an example taken from a fictional text in COCA). Students are encouraged to explore the corpus on their own, which requires free registration. A tutorial on its use is included on the COCA website.