Corpus linguistics and the study of nineteenth-century fiction

Mahlberg, Michaela

Citation for published version (Harvard):
Mahlberg, M 2010, 'Corpus linguistics and the study of nineteenth-century fiction', *Journal of Victorian Culture*, vol. 15, no. 2, pp. 292-298. <https://www.tandfonline.com/doi/abs/10.1080/13555502.2010.491667>

Journal of Victorian Culture
Vol. 15, No. 2, August 2010, 292–298

Routledge
Taylor & Francis Group

DIGITAL FORUM

Corpus Linguistics and the Study of Nineteenth-Century Fiction

Michaela Mahlberg

The increasing number of digital resources opens new routes for research in a variety of fields. The range of methods and approaches for the creation and exploitation of electronic data also benefits from cross-disciplinary approaches. In this contribution I outline some of the opportunities that corpus approaches offer for the study of nineteenth-century fiction. I begin with a brief overview of some basic concepts in corpus linguistics before focusing on corpus stylistics and presenting examples referring to key words, suspensions and clusters. In addition to showing the potential of these approaches, I also want to make the point that quantitative research can only provide valuable insights when it is linked to qualitative analysis.

Corpus linguistics is a field of linguistics that came into being relatively recently with its existence closely linked to developments in computing. Corpus linguistics studies language on the basis of samples of naturally occurring language. These samples are stored electronically in what is called a 'corpus'. Corpora may contain written language, for example from newspapers, textbooks, leaflets, essays written by language learners, as well as transcriptions of spoken language, from casual conversations, radio broadcasts, TV shows, lectures and so on. Corpus projects also include the compilation of video corpora and sign language corpora. Additionally, electronic collections of texts that were not designed as a corpus in the first instance (such as archives or the Web) can be used for corpus linguistic studies. So far, corpus linguistics seems to have had the greatest impact in the field of lexicography, where corpora can help identify the words and meanings to be included in a dictionary.[1]

Corpus software is used to quantify linguistic phenomena and display data so that the researcher can investigate linguistic patterns. I start with a simple example to illustrate some basic principles. Figure 1 shows 25 concordance lines for the search word 'pocket', retrieved with *WebCorp*.[2] This search engine makes it possible to access the Web as a corpus and retrieve concordances for a search word. In corpus linguistics, a 'concordance' is a display format that shows a search word with a specified amount of co-text to its left and to its right. A crucial feature of a quantitative approach to language is the observation of repetitions. Such repetitions

1. John Sinclair, ed., *Looking up. An Account of the COBUILD Project in Lexical Computing* (London: HarperCollins, 1987).
2. *WebCorp*, Research and Development Unit for English Studies, Birmingham City University (1999–2010) <http://www.webcorp.org.uk/> [accessed online April 2010].

ISSN 1355-5502 print/ISSN 1750-0133 online

DOI: 10.1080/13555502.2010.491667
http://www.informaworld.com

```
1             Stricken New Star staff could pocket £10m bonus A group of
2   News > Business > Business News Bankers pocket £2.7bn bonuses as record
3 Video Community Jobs Business Consultants pocket $20bn of global aid Tweet
4              League Man City Middlemen to pocket £28m from Kaka deal Agents
5                Expat Health Expats out of pocket after Government cuts their benefits
6                    sports hit hard in the pocket as they prepare for 2012
7                     1 eft £900,000 out of pocket by collapse of new box
8                       could be hit in the pocket by 'Chairman's Pledge' Burnley will
9  Technology Reviews Gadget Inspectors: HD pocket camcorders put to the test
10       Mobile Phones Phone makers launch 'pocket computers' to compete with iPhone
11   Second world war Neville Chamberlain's pocket diary goes on display in
12                     customers £1m out of pocket FocusClothing, a scam website claiming
13                News > World The World: A pocket history By Christopher Lloyd Friday,
14                The pound in your holiday pocket is taking a hammering If
15                  How to make a Sheffield pocket knife Still at the cutting
16        6 February 1998; Lockable folding pocket knife is a bladed article
17                Beware: the pound in your pocket may not be worth as
18              iPods and games consoles as pocket money soars More than 20
19              that parents have cut their pocket money by £2 a week
20                      13, 2009 The end of pocket money as we know it
21          have been awarded their biggest pocket money rise on record this
22                 could leave Tata out of pocket Rhys Blakely in Bombay The
23               display for first time The pocket watch of Titanic passenger John
24                     years at sea A silver pocket watch which lay at the
25             HMRC leaves taxpayers out of pocket with delays on refunds Sunday,
```

Figure 1. A concordance with twenty five examples of 'pocket' retrieved with *WebCorp*.

show relationships between patterns and meanings. To help the identification of patterns, a concordance can be sorted in various ways. The examples above are sorted to the first word on the right of 'pocket'. Among the patterns of 'pocket' are its co-occurrences with amounts of money as in lines 1 to 4 illustrating a verb meaning ('pocket $20bn of global aid'). Other patterns are combinations of 'pocket' with 'knife', 'money' or 'watch'. There are also examples illustrating a meaning of 'pocket' as referring to something small enough to fit into a pocket: 'pocket camcorders' or 'computers', or something small and simple as in 'a pocket history'. If the concordance lines were sorted to the left of 'pocket', the pattern 'out of pocket' would also become more clearly visible (cf. lines 5, 7, 12, 22, 25). The analysis of concordances is only one of the methods used in corpus linguistics. Other methods include the comparison of word frequencies across corpora, annotating corpora with further information to retrieve specific grammatical patterns, and applying statistical tests to assess the significance of frequency information.[3]

Although these concordance lines were retrieved with *WebCorp*, they are not a totally random sample. *WebCorp* allows users to specify a domain and the above search was limited to UK broadsheet newspapers.[4] Newspapers are a valuable source to study concepts and meanings that are relevant to a society at any given point in time. Concordance line 2, for instance, is from the headline 'Bankers Pocket £2.7bn Bonuses as Record Profits Announced' of an article about Barclays' profits in *The*

3. For an introduction, see Mike Scott and Christopher Tribble, *Textual Patterns. Key Words and Corpus Analysis in Language Education* (Amsterdam: John Benjamins, 2006).
4. *Times Online* <http://www.timesonline.co.uk>; *Telegraph.co.uk* <http://www.telegraph.co.uk>; *Guardian.co.uk* <http://www.guardian.co.uk>; *The Observer* <http://observer.co.uk>; *The Independent* <http://www.independent.co.uk/>.

Independent from 16 February 2010.[5] Newspapers in general and more specifically focused selections of articles can add to the investigation of cultural key words or issues of social relevance.[6] With readily available general corpora for present day English as well as the vast resources provided by the Web, it is easy to see how corpus linguistics can play an important role for disciplines such as sociolinguistics or cultural studies.

Today newspapers automatically come in electronic format and are even published on the Web. With the focus on the nineteenth century, however, similar resources are less readily available. When digitized archives of nineteenth-century newspapers hold images of newspaper pages these can be searched with the tools provided by the archive, but are less accessible by standard concordance tools. The *Times Archive* <http://archive.timesonline.co.uk/>, for instance, contains 200 years of *The Times* newspaper (from 1785 to 1985). It allows users who have paid for a subscription to search the archive and view images of newspaper pages where occurrences of the search word are highlighted. However, as the concordance above should have shown, patterns become more clearly discernible when the immediate co-texts around a word can be easily compared. As Tognini-Bonelli points out: '[a] corpus, examined at first in KWIC [Key Word in Context] format with the node word aligned in the centre, is read vertically, scanning for the repeated patterns present in the co-text of the node'.[7] Databases and archives, in contrast, tend to be designed for users who are interested in individual texts. With the help of optical character recognition (OCR) software, the images of the texts can be turned into plain text format. Plain text documents can then be processed by concordance tools such as *WordSmith Tools*.[8] The *Times Archive* allows users to view the plain text formats but, because of access rights, large amounts of plain texts are not readily available for download.[9] Additionally, for older texts inaccuracies may occur because of archaic typefaces or source material that is damaged.[10] Such inaccuracies have implications for the generation of quantitative data. If the word 'pocket' appears, for instance, as 'p&cket' in a text, it would not be listed in a concordance for 'pocket'. The corpus tool would treat the two forms as separate words.

5. Russell Lynch, 'Bankers Pocket £2.7bn Bonuses as Record Profits Announced', *The Independent*, 16 February 2010 <http://www.independent.co.uk/news/business/news/bankers-pocket-pound27bn-bonuses-as-record-profits-announced-1901311.html> [accessed online April 2010].
6. See the study on 'sustainable development' in Michaela Mahlberg, 'Lexical Items in Discourse: Identifying Local Textual Functions of *Sustainable Development*', in Michael Hoey, Michaela Mahlberg, Michael Stubbs and Wolfgang Teubert, *Text, Discourse and Corpora. Theory and Analysis* (London: Continuum, 2007), pp. 191–218.
7. Elena Tognini Bonelli, *Corpus Linguistics at Work* (Amsterdam: John Benjamins, 2001), p. 3.
8. Mike Scott, *WordSmith Tools*, *Version 5* (Liverpool: Lexical Analysis Software Ltd, 2008).
9. It is possible to negotiate access to some of the data for research purposes, as a PhD student of mine has done.
10. See also the 'Note about plain text quality' that the *Times Archive* displays at the top of plain text versions <http://archive.timesonline.co.uk/tol/archive/> [accessed online April 2010].

Currently, the area of nineteenth-century studies where corpus linguistic methods are most readily applicable is the study of fiction. Compared to contemporary fiction, theoretically nineteenth-century fiction should pose even fewer problems for corpus studies because copyright has expired. However, commercial databases such as *Literature Online* <http://lion.chadwyck.co.uk/> limit access to subscribers and provide interfaces that are mainly designed to display texts so that they can be read horizontally, not vertically as in a concordance. A helpful resource for corpus work is *Project Gutenberg* <http://www.gutenberg.org>. The main advantage of the site is that the texts are freely available and come in plain text format. Each text can be downloaded as an individual file so users can collect their own corpora for use with corpus software. An issue to consider when using texts from *Project Gutenberg*, however, is that the ebooks are produced by volunteers and the quality of proofreading can vary.

The availability of texts is not the only point to consider when following a corpus approach. Corpus studies are usually interested in characteristics of language in general. When corpus linguists study the language of fiction, fiction tends to be treated as a register in comparison to other registers. Qualities of individual texts do not play an important role. However, at the interface of corpus linguistics and literary stylistics, researchers have started to become interested in work that can be called 'corpus stylistics'. Corpus stylistics employs methods and approaches of corpus linguistics and links them with concerns in literary stylistics and literary criticism.[11] Corpus linguistic methods that take into account statistical measures of significance may be difficult to apply to the study of individual texts, because there may simply not be enough data to make reliable claims. But, at the same time, as Short points out, 'analysing a long novel in close stylistic detail could take a lifetime'.[12] Thus the computer-assisted quantification of linguistic phenomena can be very helpful. It also makes it possible to compare individual texts against general reference corpora. Still, the value of applying corpus methods is defined through the links that can be made between quantitative findings and qualitative analysis.

A starting-point for the analysis of a novel is to begin by generating key words. In a recent article, Catherine Smith and I exemplify the key words method by comparing Jane Austen's *Pride and Prejudice* against a corpus of novels by eighteen nineteenth-century authors other than Austen.[13] The key words were retrieved with the corpus software *WordSmith Tools*.[14] Key words in *Pride and Prejudice* are words that are relatively more frequent in the novel than in the reference corpus made up of

11. Michaela Mahlberg, *Corpus Stylistics and Dickens's Fiction* (London: Routledge, forthcoming).
12. Mick Short, *Exploring the Language of Poems, Plays and Prose* (Harlow: Pearson Education, 1996), p. 255.
13. Michaela Mahlberg and Catherine Smith, 'Corpus Approaches to Prose Fiction: Civility and Body Language in *Pride and Prejudice*', in *Language and Style*, ed. by Beatrix Busse and Daniel McIntyre (Basingstoke: Palgrave, 2010), pp. 449–67.
14. Scott, *WordSmith Tools*, *Version 5*.

the novels by the other eighteen authors.[15] Key words can provide a first overview of a text, pointing to words that are potentially useful for more detailed analysis. One of the key words for *Pride and Prejudice* is the noun 'civility'. Smith and I look at the contexts in which 'civility' occurs with the help of a concordance, as in the example of 'pocket' above. We argue that the patterns that are revealed by the concordance can be related to Emsley's definition of 'civility'. For Emsley, civility is 'ideally the outward manifestation of real goodness, politeness based on respect, tolerance, and understanding', whereas in practice, it 'has a great deal to do with … maintaining social niceties even when one does not feel like being polite'.[16] To find 'civility' as a key word for *Pride and Prejudice* is in line with the fact that the novel deals with misunderstandings and misjudgements that result from a mismatch between outward civilities and true virtues of characters. Further steps in a key word analysis might be to compare patterns of the key word in the text under analysis with the meanings that are found for the word in the reference corpus; in this case, by looking at a concordance for 'civility' across the novels by the other eighteen authors. Another option for a more detailed analysis is to link the analysis of 'civility' to findings on body language in *Pride and Prejudice*. The importance of body language for the novel is highlighted, for instance, by Korte.[17]

For the study of body language it is not always straightforward to pin down meanings to linguistic units on the textual surface that can easily be identified by the computer. One of the approaches that Smith and I suggest for the study of body language is to look for 'suspensions'. A suspension is a span of (narrator) text which interrupts a span of quoted speech (or thought, or writing). The concept of the suspension is based on the work of Lambert who looks at the suspended quotation in Dickens, focusing on interruptions of five or more words.[18] While Lambert was not able to draw on corpus methods to find suspensions, Smith and I illustrate how corpus annotation and software can help to identify specific places in a novel.[19] We work with an electronic version of *Pride and Prejudice* that has been annotated for quoted speech and suspensions in particular. The annotation is based on quotation marks and other punctuation indicators and stored in XML (extensible Mark-up Language). Here is an example of a suspension highlighted in italics: '"And this,"

15. For further examples of key words analyses see, Mike Scott, 'Key Words of Individual Texts', in Mike Scott and Christopher Tribble, *Textual Patterns. Key Words and Corpus Analysis in Language Education* (Amsterdam: John Benjamins, 2006), pp. 55–72; Jonathan Culpeper, 'Keyness: Words, Parts-of-Speech and Semantic Categories in the Character-Talk of Shakespeare's *Romeo and Juliet*', *International Journal of Corpus Linguistics*, 14.1 (2009), 29–59.
16. Sarah Baxter Emsley, *Jane Austen's Philosophy of the Virtues* (Basingstoke: Palgrave Macmillan, 2005), pp. 90–91.
17. Barbara Korte, *Body Language in Literature* (Toronto: University of Toronto Press, 1997).
18. Mark Lambert, *Dickens and the Suspended Quotation* (New Haven: Yale University Press, 1981).
19. The claims made by Lambert for the distribution of suspensions across Dickens's fiction can be tested with computer methods, as argued by Michaela Mahlberg, Catherine Smith and Matthew Brook O'Donnell, 'Investigating Speech in Fiction Using an XML-Annotated Corpus', Paper presented at ICAME 30, Lancaster, UK, 27–31 May 2009.

cried Darcy, as he walked with quick steps across the room, ''is your opinion of me! . . .''. The annotation makes it possible to extract all suspensions that occur in a text and then, in a further step, analyse them qualitatively to find functional categories. Suspensions are a useful place to look for body language, as they contain information accompanying speech. When people speak there is always body language involved. Although in novels the extent to which body language is made explicit varies, contexts of speech provide useful starting-points to find examples.

A fundamental principle in corpus linguistics is the assumption that repetition of a pattern points to functional relevance of the pattern. One option to find patterns in corpora is 'clusters' – repeated sequences of words, such as 'at the end of the'. The parameters for the length and minimum frequency of a cluster vary according to the research question. While I cannot go into much methodological detail here, I want to highlight the need to link corpus findings with qualitative approaches. I have employed the analysis of clusters for the study of texts by Charles Dickens focusing on clusters of the length five, i.e. repetitions of five words.[20] The findings suggest that groups of clusters can be interpreted as building blocks of fictional worlds. One group of clusters contains what I call 'labels', that is clusters that either contain names or expressions used to refer to characters – 'inspector bucket of the detective', 'I believe said mr dombey', 'man with the wooden leg' – or clusters that do not contain names but are specific to individual texts, such as 'i expect a judgment shortly' from *Bleak House* where it is spoken by Miss Flite. Labels overlap to some extent with what has been identified in the literature as habitual phrases that aid characterization.[21] Another group of clusters are body part clusters which contain a body part noun such as 'back' in the cluster 'with his back to the' (example (1) below), or 'hands' in 'his hands in his pockets' (example (2)). Clusters in this group point to examples of body language. They are found on a functional continuum ranging from contextualizing functions to illustrating habitual and exaggerated behaviour.

(1) In the friendliest manner he is making himself quite at home *with his back to the* fire, executing a statuette of the Colossus at Rhodes. (*Our Mutual Friend*)
(2) 'Di-rectly, sir,' said the coachman, with *his hands in his pockets*, looking as much unlike a man in a hurry as possible. (*Sketches by Boz*)

The analysis of functions of clusters can provide links to the approach to the externalization of character put forward by Juliet John.[22] John shows how Dickens's methods of characterization draw on the methods of popular melodrama. She discusses the relevance of the ostension of the private and the depiction of transparent character. John argues that Dickens depicts emotions in exaggerated ways and in his narrative prose gestures and actions add to externalize character, whereas

20. See Mahlberg, *Corpus Stylistics and Dickens's Fiction*.
21. See G.L. Brook, *The Language of Dickens* (London: Andre Deutsch, 1970).
22. Juliet John, *Dickens's Villains. Melodrama, Character, Popular Culture* (Oxford: Oxford University Press, 2001).

the mind only takes a marginal place. The analysis of clusters can contribute detail on the linguistic means used for this mode of characterization.

The examples of key words, suspensions and clusters illustrate only three of the options to approach literary texts with the help of corpus methods. Ultimately, the choice of method depends on the text under analysis and the overall aim of the study. Although in this article the examples have to remain brief, they show that the application of corpus techniques to the study of literary texts has to combine quantitative and qualitative analyses to provide useful insights. Thus, literary scholars may profit from corpus methods, and through the engagement with literary criticism, corpus linguists may be able to develop more specific corpus resources and methods.

Michaela Mahlberg
University of Nottingham
michaela.mahlberg@nottingham.ac.uk