Michaela Mahlberg. *Corpus stylistics and Dickens's fiction.* New York: Routledge 2013, xiv+221. (ISBN 978-0-415-80014-3)

Reviewed by Jane Demmen (Lancaster University)

1. Overview

Corpus Stylistics and Dickens's Fiction brings together some of Michaela Mahlberg's notable contributions to the field of corpus stylistics in recent years (e.g. 2007, 2009). In this book, Mahlberg demonstrates the ways in which corpus linguistic computer software can be brought to bear on digitised collections (or 'corpora') of literary texts, in order to find language patterns which would be difficult, or even impossible, to observe through close reading and manual analysis. Such language patterns can be indicative of language features that are of interest in linguistic analyses of, for example, character construction, authorial style, or fictional worlds. The analyses in this volume provide some new, empirically-based perspectives on the ways in which Charles Dickens constructed some of his most well-known characters, adding to existing discussions in literary critical research. Additionally, Mahlberg's work reveals language patterns in Dickens's novels which have hitherto gone unobserved. As Mahlberg emphasises (p. 22), corpus stylistic studies "do not automatically trigger a revolution in literary criticism. More important is how corpus findings corroborate, modify, and complement findings of literary critics". Mahlberg's research focuses particularly on the analysis of the 'local textual functions' of recurrent word sequences (explained further in sections 2 and 3 below). Her book adds to the growing body of work which provides evidence of the added analytical value that can be gained from examining longer units of language than single lexical items in linguistic explorations of literary texts (see also, for example, Stubbs 2005; Fischer-Starcke 2009).

The main aims of the book are (1) to carry out a systematic investigation of a feature already noted as important in Dickens's writing (the construction of memorable characters through the repetition of particular phrases), by focusing on the local textual functions of repeated sequences of words, and (2) to demonstrate how corpus linguistic methods can help to facilitate the analysis of literary texts. In chapter 1, Mahlberg introduces what has become known as 'corpus stylistics'. Chapter 2 concerns how the fictional world of literary texts, and in particular character construction, can be explored using corpus data, and chapter 3 explains the type of quantitative data used in the study (word clusters). The grouping of the quantitative data into functional categories is discussed in chapter 4, with some brief examples. Chapters 5, 6, 7 and 8 each focus on the function of a particular category of word cluster in more detail, and conclusions about the study are given in chapter 9. I will now review each chapter in more detail, and end with some general remarks about the book.

2. Background: Chapters 1-2

Chapter 1 provides a succinct but comprehensive explanation of the emergence of 'corpus stylistics', and the contribution that studies in this field have made by providing empirically-based perspectives on well-known literary texts. Mahlberg begins by explaining the interface between corpus linguistics ("the study of language on the basis of collections of computer-readable texts") and literary stylistics ("the study of style in literary texts") (p. 7). She outlines what corpus methods bring to stylistic analysis, particularly through the important concept of 'foregrounding' in stylistics (pp. 8-11). Foregrounded language is that which stands out to the reader, i.e. which has psychological prominence, through repetition and/or deviation from some kind of linguistic norm (see also Leech & Short 2007: 23, 39-40). Corpus linguistic software

tools can automatically identify and count words and word sequences which are repeated in texts, and generate lists of linguistic items according to frequency and/or statistical significance. The researcher can survey these lists with a view to finding lexical patterns. Lexical patterns are related to meanings and the mental organisation of language, notably through words working together in relationships of co-occurrence such as ‘collocations’ (pp. 15-16; see also, for example, Hori 2004). The researcher can use lexical patterns in the corpus data as a guide to finding language which may be foregrounded in the texts, though as Mahlberg emphasises, not all items which are statistically unusual, or deviant, are also psychologically prominent (a point also made by Leech & Short 2007 and Leech 2008). Therefore, corpus-generated results can be taken to be “tendencies of language use” (p. 10), and of particular help in highlighting language patterns which are subtle and which may go unnoticed through manual analysis (p. 11). Mahlberg also explains that her study is ‘data-driven’, meaning that the data generated with the corpus tools governs what she analyses in terms of local textual function in the texts (i.e. the analysis is not based on any pre-conceived list of items). The meaning of ‘local textual function’ in this study is explained in relation to general functions (such as those of lexical bundles), and encompasses “functions specific to a (group of) text(s) and/or specific to a (group of) lexical item(s)” (p. 17). This is supported by a further operational definition in chapter 4.

In chapter 2, Mahlberg explains her rationale for investigating character construction in Dickens’s novels using recurrent word sequences as data. She argues that “clusters can be interpreted as textual building blocks for fictional worlds. These building blocks have particular functions in the creation of characters” (p. 26). Mahlberg discusses how distinctive individuals are constructed both through the

language in the novels and the cognitive processes of the reader, with reference to literary critical research into character construction in Dickens's work (e.g. Rosenberg 1996; John 2010), and theories of characterisation and text worlds (e.g., respectively, Culpeper 2001; Gavins 2007). At the end of chapter 2 Mahlberg briefly introduces the five main functional categories into which she groups her word cluster data: 'Labels', 'Speech clusters', 'Body Part clusters', 'As If clusters' and 'Time and Place clusters'. She explains that these groupings are assigned "by taking account of features on the textual surface: clusters with similar formal features are grouped together" (pp. 40-41). These categories are discussed further below in the course of chapters 4 to 8. As indicated above, the functional categories are derived from the data itself, rather than from any existing theoretical frameworks of language function; for further discussions of theories of language function, and particularly their application to stylistic analyses, see for example Leech (2008). The discussions in chapter 2 not only contextualise Mahlberg's work, but also broaden its potential relevance and interest to audiences beyond linguistics.

3. Methodology and data: Chapters 3-4

Mahlberg provides further detail of her corpora and her data in chapter 3. Her main resource is a 4.5 million-word corpus of 23 works by Charles Dickens, comprising 15 novels, 7 stories and sketches, and one text described as non-fiction ('American Notes for General Circulation'). Mahlberg explains that the latter (which seems something of a misfit compared to the other texts), is included because she follows Hori's (2004) text selection (p. 42). Mahlberg also uses two contemporaneous 'reference corpora' of nineteenth-century fiction in order to make comparisons with her Dickens corpus. The first is similar in size to the Dickens corpus (4.5 million words), comprising 29

nineteenth-century novels written by 18 authors, and the second is slightly larger (approximately 6 million words), comprising a single text by each of 37 authors. All of the texts are sourced from the website *Project Gutenberg*. Comparing a single-author corpus with a larger, multi-author contemporaneous corpus of the same text-type is a conventional approach in corpus stylistics, which helps the researcher to gauge whether or not a particular language feature is part of an author's individual style or characteristic of a particular genre and/or historical period. (See also, for example, Leech's 2008: 162-178 analysis of Virginia Woolf's short story *The Mark on the Wall*, and Fischer-Starcke's 2009 analysis of Jane Austen's novel *Pride and Prejudice*).

Mahlberg uses the popular corpus linguistic software *WordSmith Tools* (Scott 1996-2013) to extract her word cluster data. She explains that the operational definition of a 'cluster' is simply a recurrent sequence of words identified by the programme's algorithms (pp. 44-47). Cluster output from this programme and others like it has been used as the basis for studies of recurrent word combinations of varying structures and definitions, which are not necessarily compatible with one another. Mahlberg makes this clear through a brief survey of studies which focus on, for example, 'chunks', 'n-grams' and 'lexical bundles' (pp. 48-51). Drawing on the research of Biber *et al.* (1999), Manning & Schütze (1999), O'Keeffe *et al.* (2007) and Wray (2008), amongst others, Mahlberg explains that the main idea underlying the extraction and analysis of recurrent sequences of words is that language is argued as being to some extent 'formulaic'. In other words, it is stored in pre-assembled units or 'chunks' which are mentally loaded with semantic, pragmatic and discoursal information, and which can be deployed to increase the speed and efficiency of language processing (with the addition of situation-specific language as required). Mahlberg devotes a section of chapter 3 to evaluating the

respective advantages of examining lexical bundles or word clusters in her data. Lexical bundles are recurrent word sequences which meet particular criteria of frequency and distribution across groups of texts, which have been shown to be useful in profiling registers and text-types (see e.g. Biber *et al.* 1999: 993-1024; Biber 2006).

As noted above in section 2, Mahlberg makes clear earlier in the book that the parameters of 'local textual functions' in her study extend to "functions specific to a (group of) text(s) and/or specific to a (group of) lexical item(s)" (p. 17, chapter 2). While some of the recurrent word combinations in her data could fit the definition of a lexical bundle, through occurring in multiple texts, others may be text-specific, and so she uses a more localised definition for 'word clusters'. In chapter 3 Mahlberg presents the most frequently-occurring four-word units in the larger of her two reference corpora, distinguishing clusters which qualify as lexical bundles from others which do not, and linking some of the lexical bundle functions with those in Biber's (2006) study. Mahlberg experiments with different lengths of clusters, and finds that five-word clusters yield a manageable number of results to analyse from the corpora, most of which have functions at the level of locality that she is interested in.

Mahlberg's basic taxonomy of local textual functions is set out in chapter 4. She determines the functional categories based on a sample of 56 five-word clusters which occur with relatively high frequency in fiction by Dickens, when it is compared to each of the two reference corpora of other contemporaneous fiction. She does this using the 'keyness' tool in *WordSmith Tools*, which measures word or cluster frequencies in two texts or sets of texts using a statistical test. As Mahlberg points out, this means that the sample clusters "point to tendencies of usage that are characteristic of Dickens" (p. 63) compared to the other authors. Therefore, the local textual functions which determine

the functional categories are those which Dickens made use of relatively more than his contemporaries. Mahlberg notes that the same functional categories were also determined in her (2007) study, not only from key clusters but also from frequent clusters in Dickens's work and other nineteenth century fiction. That the categories are particularly oriented to Dickens's style serves the main purpose of the book in analysing language strategies which Dickens particularly made use of in characterisation. (Chapters 5 to 8 are based mainly on data from the Dickens corpus, not the other contemporaneous corpora).

Mahlberg defines the categories of local textual functions in her data as follows (p. 66):

- 'Labels' clusters include a name, and/or "are part of an expression that is used in a way similar to a name", and/or "occur in only one of the texts in the corpora under investigation", e.g. *mr pickwick and his friends*, *how not to do it.*
- 'Speech' clusters "contain a first-or second-person pronoun or possessive", e.g. *how do you do sir*, *upon my word and honour*.
- 'Body Part' clusters "are defined by the presence of a body part noun", e.g. *his hands in his pockets*, *his head on one side*.
- 'As If' clusters "are identified through the occurrence of *as if*", e.g. *as if he would have, as if he were a.*
- 'Time and Place' clusters "contain a nominal time or place expression and/or a preposition indicating that the cluster functions as a time or place expression", e.g. *at the upper end of*, *after a great deal of*.

There is also a miscellaneous ('d') category into which clusters not fitting any of the above local textual functions are grouped (p.67).

Developing a functional category system which is consistent in its inclusion criteria, yet flexible enough to accommodate the variety of items which occur in a dataset, is not easy. Mahlberg explains some of the difficulties, such as the inevitability of encountering results which have multiple functions. Where a cluster would qualify for inclusion in more than one category, Mahlberg places it in the highest appropriate category on the above list (for example, a cluster containing a body part noun and *as if* would be classified as a Body Part cluster) (p. 67). I expand on the definitions of the categories in the course of discussing chapters 5 to 8, in which Mahlberg provides further details. Continuing in chapter 4, Mahlberg notes that all five-word clusters which occur at least five times in one text are included in the study (p. 68), establishing the operational boundaries of 'local textual function'. Fewer qualifying clusters are present in the other contemporaneous fiction, although Mahlberg finds that all the functional categories are represented in the other contemporaneous fiction. In other words, the local textual functions which are more prevalent in Dickens's work are not exclusive to it. Mahlberg notes that clusters in the Time and Place group have more general functions than those in the other groups, and from chapter 5 onwards she excludes this category and focuses on just the other four cluster categories.

4. Analyses: Chapters 5-8

Chapter 5 concerns Speech clusters (those featuring a first- or second-person pronoun or a possessive). Drawing on Carter & McCarthy's (2006: 835) discussions of cluster functions, Mahlberg identifies four functions amongst the Speech clusters in the Dickens corpus (pp. 76-82), which are summarised below:

- 'Negotiating information', e.g. *you don't mean to say, I don't know how it, what do you think of*. Mahlberg argues that "[t]he interpersonal meanings associated

with such clusters relate to the monitoring of shared or new knowledge, as well as hedging and expressing various degrees of certainty about the world" (p. 77).

- 'Turn-taking', e.g. *I was going to say, I should like to ask, I want to speak to.* Mahlberg finds that these items "function to elicit responses including forms of backchannelling from the listener or to hand the turn over from one speaker to another" (pp. 78-79).
- 'Politeness formulae', e.g. *I beg your pardon sir, will you have the goodness.* These clusters have a function in conveying thanks, apologies or greetings (for example) (pp. 80-81).
- 'First-person narration', e.g. *I could not help observing*, which are items that convey the narrator's view (pp. 81-82).

Mahlberg points out that function is context-dependent. She illustrates a series of confrontations between characters with different Speech clusters from the Dickens corpus (e.g. *what have you got to say, what do you want here*) (pp. 83-90). She follows this with longer analyses of confrontations in extracts from two novels, *Pickwick Papers* and *Oliver Twist*. Mahlberg links some of her findings in this chapter to linguistic stylistic theory of speech and thought presentation (Semino & Short 2004). For example, she illustrates that the cluster *I felt that I was* is used in indirect thought presentation, which gives insights into the viewpoint of the narrator (p. 82). In these analyses Mahlberg demonstrates the way that language in fiction is multi-layered, telling the story to the reader whilst also building up character impressions in the reader's mind through strategies of recurrent word sequences.

Chapter 6 concerns Body Part clusters, i.e. those featuring a body part noun. Mahlberg adapts Korte's (1997) framework for analysing body language in literary texts

and applies it to her Dickens data. This chapter is an excellent demonstration of how a systematic and careful analysis of language patterns that surface in empirical data can complement and provide additional insights into language style features which have been noted by literary critical scholars. Mahlberg explains that Korte's (1997) approach takes in both the real-world understanding of body language a reader is likely to have, as well as the ways that body language is used in literary texts, and that it can be adapted to suit texts from different historical periods. Mahlberg also links Korte's work to other theories of characterisation, and makes the important observation that the meaning of body language is culturally and historically situated (pp. 101-109). In a detailed analysis of the cluster *with his back to the*, which occurs in the Dickens corpus and the other nineteenth-century fiction, Mahlberg finds a pattern of descriptions of male characters taking a 'fireplace pose' (pp. 111-114). Linking this with Korte's (1997: 212) work, Mahlberg explains that this is a typical pose in which men are represented in this period (in art as well as literature), because "[t]he fireplace is a prominent area of the room and standing in front of it puts men in a prominent and powerful position" (p. 112). There are also cases in Mahlberg's data where women are represented in this pose. These are fewer, however: a pattern indicating that a woman standing in front of the fire is unusual, and therefore of interest in the novels.

Mahlberg goes on to discuss further types of distinctive body language which surface from the patterns in her data, including 'gaze behaviour' (through clusters featuring the noun *eyes*) and 'touch behaviour and authentication' (exemplified by *his hand upon his shoulder*). Mahlberg argues that some of the body language in fiction provides an 'authenticating' effect, i.e. it confirms to the reader that real-world behavioural norms exist in the fictional world (pp. 117-118), whereas other body

language stands out as unusual and provides 'thematic cues' which draw the reader's attention. For example, the cluster *her apron over her head* occurs in more than one of the texts in the Dickens corpus, and indicates a type of body language which "may refer to hiding from something" (p. 119). Its recurrence six times in the novel *Little Dorrit*, however, particularly characterises the maid Affery Flintwich, for whom "the apron serves as a kind of shelter" and helps conveys her "worried and confused state of mind" (pp. 122-123). Mahlberg explains that her findings encompass

> not only the idiosyncratic body language that is commonly discussed by critics but also and maybe equally importantly the body language that is less noticeable by the reader and therefore does not seem to have received much attention in the literature (p. 100).

Chapter 7 concerns clusters which feature *as if*. Mahlberg notes that they "mainly consist of function words and link two clauses", so they are not suited to the same kind of functional analysis as the other cluster types (p. 128). This does perhaps raise the question of whether *as if* merits a specific category of local textual function. However, Mahlberg points out that linguistic and literary scholars of Dickens's work have noted that *as if* is of interest, and she opts to analyse the words which co-occur with *as if* most frequently (i.e. the collocates). This step demonstrates that a variety of analytical angles may be necessary to explore corpus data fully and usefully, with the aid of multiple facilities in the corpus software tools. Mahlberg observes that *as if* is mainly used to introduce a comment by the narrator, and to identify what these comments tend to concern she analyses the left collocates of *as if* in detail. Examples include comments on the way characters look (e.g. "Ralph *looked* as if he did not quite understand the observation" (p. 142), or the way they speak (e.g. "'Oh, Mr Pickwick!' said Arabella, in a low *voice*, as if alarmed at the silence [...]" (p. 144). (Italics Mahlberg's own, indicating words identified as collocates by the corpus software tool).

The analysis of the largest category of clusters, Labels, is the focus of chapter 8. Mahlberg acknowledges that Labels may seem "a rather heterogeneous category that almost seems to collect what does not fit the other groups easily" (p. 152), and that there is some "fuzziness" in this group (p. 158). It does seem the least cohesive of the categories, mainly because the greatest proportion of member items are included on the basis of distribution only (occurring in just one text in the corpus). The other inclusion criteria for a cluster to be admitted to this category (that "it contains a name or is part of an expression that is used in a way similar to a name"; p. 152 and mentioned above) would also benefit from some further clarification (for example, whether terms of address, terms of reference, role titles, kinship titles, honorifics, pronouns and/or other items are included).

Mahlberg classifies the Labels clusters into six sub-groups. 'Reporting Speech Labels' include a reporting verb (e.g. *i suppose said mr pickwick*). 'Speech Labels' include a first-or second-person pronoun or possessive, which as Mahlberg notes is also the definition of a Speech cluster (the focus of chapter 5). Mahlberg explains that "Speech Labels are specific to a text whereas Speech clusters occur across different texts", and she argues that there is a functional difference in that "Speech Labels are more closely linked to the content of the text; they are associated with particular characters and contribute character information in a more striking way" (p. 155). 'Body Language Labels' feature a body part noun, as do the Body Part clusters in chapter 6, but the Labels either include a name as well or occur in just one text. Some of the Labels are character-specific, as with the Speech Labels, but not all. 'Attributes' are Labels which "refer to fictional characters by highlighting attributes instead of using proper names", e.g. *the keeper of the wine-shop*, *the lady of the caravan* (pp. 155-156).

‘Long clusters’ are Labels clusters which are related to at least three others through overlapping with one another (i.e. they are embedded in a longer cluster) (pp. 156-157). ‘Names of people and places’ are exactly that, e.g. *mr jarndyce of bleak house*, *the theatre royal drury lane*, which Mahlberg argues “are more text-specific than place references in the form of Time and Place clusters” (p. 158). Mahlberg carries out a detailed analysis of the character Rigaud in the novel *Little Dorrit*, whose evil nature is partly constructed through repeated references to his facial expressions, notably through the cluster *and his nose came down* (pp. 159-163). Mahlberg likens the Labels clusters “to what in the literature is variously described as ‘speech tics’ or ‘character tags’”, and argues that they can be linked to “more general body language or speech behaviour” (p. 163) (such as that which is shown through the main functional categories of Body Part and Speech clusters).

5. Conclusions

Finally, in chapter 9, Mahlberg draws together the findings of her study and suggests some directions for further work, notably the possibility of using word cluster data extracted with corpus tools to explore ‘suspensions’ (pp. 168-175). This idea derives from Lambert’s concept of a ‘suspended quotation’, i.e. “a protracted interruption by the narrator of a character’s speech” of five words or more (Lambert 1981: 6).

Corpus Stylistics and Dickens’s Fiction makes a valuable contribution to the analysis of literary texts using corpus linguistic methods, and adds useful and interesting insights into Dickens’s character construction techniques. Mahlberg’s functional groupings of the cluster data are practical, and successful in enabling her to break new ground in the investigation of Dickens’s language style. There may be some issues of replicability of the functional categories in other studies (particularly the Labels

clusters, which Mahlberg herself notes as problematic). Mahlberg's book offers fresh perspectives and empirical evidence to support, extend and contextualise the way Dickens's characters are shaped by patterns of repeated language units. It will be of interest to scholars of corpus stylistics and/or phraseology, and to researchers from linguistic and literary disciplines with an interest in the works of Charles Dickens.

References

Biber, Douglas. 2006. *University language: A corpus-based study of spoken and written registers.* Amsterdam: Benjamins.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of spoken and written English*. London: Longman.

Carter, Ronald & Michael McCarthy. 2006. *Cambridge Grammar of English: A Comprehensive Guide: Spoken and Written English, Grammar and Usage*. Cambridge: Cambridge University Press.

Culpeper, Jonathan. 2001. *Language and characterisation. People in plays and other texts.* Harlow: Pearson Education.

Fischer-Starcke, Bettina. 2009. Keywords and frequent phrases of Jane Austen's *Pride and Prejudice*. A corpus-stylistic analysis. *International Journal of Corpus Linguistics* 14(4). 492-523.

Gavins, Joanna. 2007. *Text world theory: An introduction.* Edinburgh: Edinburgh University Press.

Hori, Masahiro. 2004. *Investigating Dickens' style: A collocational analysis*. Basingstoke: Palgrave Macmillan.

John, Juliet. 2010. *Dickens and mass culture*. Oxford: Oxford University Press.

Korte, Barbara. 1997. *Body language in literature*. Toronto: University of Toronto Press.

Lambert, Mark. 1981. *Dickens and the suspended quotation*. New Haven, CT: Yale University Press.

Leech, Geoffrey. 2008. *Language in literature. Style and foregrounding*. Harlow: Pearson Education.

Leech, Geoffrey & Mick Short. 2007. *Style in fiction: A linguistic introduction to English fictional prose.* 2nd Edn. Harlow: Pearson Education.

Mahlberg, Michaela. 2007. Clusters, key clusters and local textual functions in Dickens. *Corpora* 2(1). 1-31.

Mahlberg, Michaela. 2009. Corpus stylistics and the Pickwickian watering-pot. In Paul Baker (ed.) *Contemporary corpus linguistics*, 47-63. London: Continuum.

Manning, Chris & Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

O'Keeffe, Anne, Michael McCarthy & Ronald Carter. 2007. *From corpus to classroom: Language use and language teaching.* Cambridge: Cambridge University Press.

Project Gutenberg. http://www.gutenberg.org/ (19 August, 2013).

Rosenberg, Brian. 1996. *Little Dorrit's shadows: Character and contradiction in Dickens.* Columbia: University of Missouri Press.

Scott, Mike. 1996-2013. *WordSmith Tools*. Oxford: Oxford University Press/Lexical Analysis Software Ltd. http://www.lexically.net/wordsmith/index.html (19 August, 2013).
Semino, Elena & Mick Short. 2004. *Corpus stylistics: Speech, writing and thought presentation in a corpus of English writing.* London: Routledge.
Stubbs, Michael. 2005. Conrad in the computer: Examples of quantitative stylistic methods. *Language and Literature* 14(1). 5-24.
Wray, Alison. 2008. *Formulaic language. Pushing the boundaries.* Oxford: Oxford University Press.