Clusters, key clusters and local textual functions in Dickens[1]

Michaela Mahlberg[2]

Abstract

The paper argues that corpus linguistics can make useful contributions to the descriptive inventory of literary stylistics. The concept of local textual functions is employed as a descriptive tool for the stylistic analysis of a corpus of texts by Charles Dickens. It is suggested that clusters, i.e. repeated sequences of words, can be interpreted as pointers to local textual functions. The focus is on five-word clusters and five functional groups are identified: Labels, Speech clusters, As If clusters, Body Part clusters and Time and Place clusters. The analysis draws on the identification of key clusters comparing the Dickens corpus with a corpus of nineteenth-century fiction, it identifies links to literary criticism and it gives specific attention to the group of Body Part clusters to illustrate the functional variation of clusters.

1. Introduction

> 'Wery pleasant,' rejoined Mr. Weller. 'Wery pleasant and comformable.'
>
> The precise meaning which Mr. Weller attached to this last-mentioned adjective, did not appear; but, as it was evident from the tone in which he used it that it was a favourable expression, Mr. Pickwick was as well satisfied as if he had been thoroughly enlightened on the subject.
>
> (*Pickwick Papers*)

Examples like the one above are often referred to by literary critics when they discuss the unique style of Charles Dickens or talk about the 'World of Dickens' with its entertaining characters and their idiosyncratic behaviour. The way in which we meet these characters and their worlds is through language. With a focus on the language of Charles Dickens, in this article I

[1] I am grateful to two anonymous reviewers for their helpful critical comments on an earlier version of this article.

[2] School of English, Modern Languages Building, University of Liverpool, Liverpool, L69 7ZR, United Kingdom
Correspondence to: Michaela Mahlberg, *e-mail*: m.mahlberg@liverpool.ac.uk

aim to explore the ways in which corpora might help us study the language of literary texts. In Section 2, I look at general issues of corpus linguistic theory; Section 3 contains an overview of clusters in a Dickens corpus and in Section 4, I compare Dickens with other nineteenth-century authors to identify key clusters. In Section 5, I suggest a classification of clusters into functional groups. Section 6 takes a wider look at these functions of clusters and identifies links to literary criticism. Section 7 then focusses on one of these groups, namely Body Part clusters. The article emphasises the need for a qualitative approach and deals with the interpretation of surface features identified by the computer. It will be argued that local classifications of linguistic phenomena provide a useful tool for the analysis of literary texts.

2. Corpora, literary texts and descriptive tools

Although corpus linguists have only recently developed an interest in what may be called 'corpus stylistics', the use of computers for the study of literary texts is not new. In a survey of projects that mainly cover the work of literary scholars, Hockey (2000: 67 ff.) presents examples of studies that go back to the 1960s. However, computer-assisted approaches to the analysis of literary works are not necessarily regarded as corpus linguistic approaches. While corpus linguistics aimed, initially, to create large text collections, in literary studies the computer was used to investigate individual texts or small amounts of data. As corpus linguistics has developed, small and purpose-built corpora have received more attention. At the same time, corpus linguistic approaches to the analysis of literary texts are becoming increasingly popular (e.g., Lawson, 2000; Adolphs and Carter, 2002; Culpeper, 2002; Hori, 2004; Semino and Short, 2004; Stubbs, 2005; Adolphs, 2006; Scott, 2006; Starcke, 2006; see also the overview in McEnery *et al*., 2006: 113–16). 'Corpus stylistics' does not only appear as the title that Semino and Short (2004) chose for their book describing an innovative approach to discourse representation, but it has come to refer to a whole new and growing field of study. Still, corpus approaches to literature and literary style are only in an early stage of development. The full potential of corpus linguistic methodology for literary stylistics is yet to be exploited and both philosophical and practical barriers need to be overcome, as Wynne (2006) emphasises. The link between corpus linguistics and literary stylistics is also discussed further in Mahlberg (forthcoming a).

Criticism of corpus stylistics sometimes reminds one of the suspicions held by many literary critics against the quantification of features of literary texts. However, criticism of corpus stylistics also seems to refer to issues concerning stylistics in general such as the selectiveness of linguistic features under investigation. A major disadvantage of the use

of computers for the study of literature can be seen in the nature of the tools, as described by Miall (1996):

> the gap is still immense between what readers can do effortlessly, and what a computer can do. Scholars interested in calling on a computer to aid their research are limited to a very narrow range of possible operations, and such operations still fall largely outside the mainstream work of literary scholarship.
>
> (Miall, 1996: online)

The exploration of ways in which computers can be used to study literary texts links in with fundamental questions relating to how corpus linguistic methodology can be applied in the study of language. For some, corpus linguistics is still mainly a methodology, for others, corpus linguistics requires its own theoretical context, an opposition that can be covered by the terms 'corpus-based' vs. 'corpus-driven' (see Tognini-Bonelli, 2001). Within the scope of this paper I cannot enter a detailed discussion of the relationship between theory and methodology in corpus linguistics, which I have done elsewhere (Mahlberg, 2005), but I want to summarise briefly what I regard as the main points of a corpus theoretical framework. These points are present to various degrees, and sometimes with differing implications, in the work of Sinclair (2004), Teubert (1999, 2005), Hunston and Francis (2000), Stubbs (2001), Tognini-Bonelli (2001) and Hoey (2005), and other authors that can be seen to belong to the neo-Firthian tradition. However, the way in which theoretical arguments are organised and fundamental claims identified is crucial to build a theoretical context. Therefore, the following three points outline what I see as the pillars of a corpus theoretical framework. These points can provide a bridge to literary stylistics and contribute to further developing the notion of corpus stylistics:

1) language is a social phenomenon;
2) meaning and form are associated; and,
3) a corpus linguistic description of language prioritises lexis.

If the focus is on language as a form of social interaction, it is possible to view meaning as use (point 1). Thus, meaning is observable through the repeated patterns of usage which are evident in corpora. Meaning has a subjective dimension of an individual's language experience and meaning has a social dimension that is shared within the discourse community. Thus we can distinguish between individual styles and conventional patterns and norms. The co-occurrence patterns that we find in corpora reflect the association of meaning and form (point 2). The link between meaning and form is the basis for a corpus linguistic description of meaning. Corpus linguistics describes meaning from a bottom-up point of view (point 3). The focus is on formal aspects of meaning that are

observable as patterns of lexical items. This focus on lexical items requires descriptive categories that are less general than systematic distinctions, which are meant to account for the whole of a language. So in a corpus theoretical approach, a grammar is seen as a set of generalisations about the behaviour of words in texts. Such a grammar is less general than a description of syntactic phenomena, but such a grammar can also account for the relationship between lexical and textual properties. It is best organised in a flexible way, as a 'flexible grammar' with local categories of description (for more detail on issues outlined in this paragraph see Mahlberg, 2005, especially Chapters 2 and 8).

In a corpus theoretical context, the application of corpus methodology to the study of literary texts can be described as 'corpus stylistics'. Corpus stylistics investigates the relationship between meaning and form. Thus it is similar to both stylistics and corpus linguistics (see Mahlberg, forthcoming a). Whereas stylistics pays more attention to deviations from linguistic norms that lead to the creation of artistic effects, corpus linguistics mainly focusses on repeated and typical uses, as these are what the computer can identify. What corpus linguistics says about norms is that words tend to co-occur and form habitual collocations. Corpus linguists have been arguing that collocation is a phenomenon that is more pervasive than established grammatical descriptions of English seem to suggest, and descriptive tools have been introduced to characterise this pervasiveness. Most notably, there is Sinclair's (2004) concept of the lexical item with the categories of 'collocation', 'colligation', 'semantic preference' and 'semantic prosody' to describe increasingly abstract co-occurrence patterns around a fixed core. Innovative descriptive categories that have been developed in the field of corpus linguistics can also be used in literary stylistics. Applications of the concept of semantic prosodies, for instance, are illustrated by Adolphs and Carter (2002). Sinclair (2004) stresses that a language description cannot claim to be adequate if it cannot be applied to the study of literature, too:

> Literature is a prime example of language in use; no systematic apparatus can claim to describe language if it does not embrace the literature also; and not as a freakish development, but as a natural specialization of categories which are required in other parts of the descriptive system. Further, the literature must be describable in terms which accord with the priorities of literary critics.
>
> (Sinclair, 2004: 51)

The descriptive tool that is illustrated in the present article is the concept of 'local textual functions'. Local textual functions are 'textual' as they describe the functions of words (or combinations of words) in text, and they are 'local' in that they do not claim to capture general functions, but functions specific to a (group of) text(s) and/or specific to a (group of) lexical item(s). Local textual functions are part of a bottom-up description

of meanings in texts (see point 3, above). In this paper, the group of texts under investigation is a corpus of texts by Charles Dickens and a corpus of nineteenth-century novels. A starting-point for the description of local textual functions is the identification of clusters with the help of *WordSmith Tools* (Scott, 2004). The main focus is not on technical issues alone, but on corpus theoretical issues such as the interpretation and classification of features on the textual surface. The discussion also links to issues in literary criticism.

3. Studying Dickensian clusters

The approach described here begins with clusters as initial pointers to local textual functions. A cluster is 'a group of words which follow each other in a text' (see Scott 2004–6: 204). Another common term for such repeated sequences of words is 'lexical bundles' used by Biber *et al*. (1999: 990) to characterise, 'sequences of word forms that commonly go together in natural discourse'. Other terms are, for instance, 'recurrent word-combinations' (Altenberg, 1998), 'chains' (Stubbs and Barth, 2003), or 'n-grams'. Not only does the terminology to refer to such sequences vary, but also the way in which they are studied. Altenberg (1998) focusses on recurrent word-combinations in spoken English. He starts with a categorisation of the structural types, which he then relates to their pragmatic functions. Culpeper and Kytö (2002), on the other hand, begin with a functional classification of lexical bundles before they consider grammatical characteristics. Also, Culpeper and Kytö (2002) look at different text types (trial proceedings and drama comedies) and are interested in diachronic findings (their data covers a period from 1560 to 1760). The variation of lexical bundles across different groups of texts is also addressed by Stubbs and Barth (2003), and in particular by Biber *et al*. (1999). (For further examples, see also Biber *et al*., 2004; Conrad and Biber, 2005; or Tribble, 2006.) When corpus studies investigate lexical bundles or clusters the focus tends to be on general patterns that hold across a number of texts in a register or subcomponent of a corpus. Culpeper and Kytö (2002), for instance, only deal with three-word lexical bundles that occur at least ten times, and in at least three different texts. Similarly, Biber *et al*. (2004: 376) limit their study to four-word bundles that are used in at least five different texts 'to guard against idiosyncratic uses by individual speakers or authors'.

Whereas general corpus studies may want to disregard idiosyncrasies of individual texts, corpus stylistic studies can pay closer attention to the individual qualities of a specific text, as Starcke (2006) demonstrates by concentrating on clusters in Austen's *Persuasion*. Stubbs (2005) also looks at clusters in one specific text, Conrad's *Heart of Darkness*, and he complements his findings with comparative data from the British National Corpus (BNC). The approach presented here is different

from these two approaches to literature in that more than one text by the same author is taken into account and the comparative data is not from a general corpus, but from a fiction corpus. In contrast with more general approaches to clusters, however, this study is not limited to clusters that have to occur across a sufficient number of texts. In particular the 'Label' clusters that will be introduced below, show that features of individual texts receive attention. Unlike more general functional interpretations of clusters, a local approach is taken here. Clusters are interpreted as pointers to local textual functions. For the identification of local textual functions the selection of texts is one parameter of 'localness'. Another is the lexical units under investigation. Local textual functions are associated with specific patterns on the textual surface – the repeated sequences of words. Therefore, the functions are local in the sense that they are associated with specific clusters. On the other hand, the types of clusters that are identifiable depend on the texts in the corpus. With corpora made up of fiction, types of functions that are identified will be more local to fiction than applicable to texts in general.

The repetition of a sequence of words can be interpreted as a sign of its functional relevance, and, thus, as an initial indication of local textual functions.[3] Conrad and Biber (2005: 58) further hypothesise, 'that extremely common, fixed sequences of words are used as unanalysed chunks by speakers and writers, and therefore will have identifiable discourse functions in texts'. Here, I also allow for clusters that are specific to individual texts and are therefore less common. As a consequence, the descriptive categories of this approach are less generally applicable and more local to the corpus under investigation. It has also to be stressed that local textual functions are not seen as restricted to fixed sequences of words. A textual function may be associated with a fixed core of one or more words but there can still be variation around this core. This point cannot be followed up in the present article, but I have discussed an example elsewhere (Mahlberg, 2007), and new tools like ConcGram (see Cheng *et al*., 2006) will prove useful for further studies in this field. With the focus on fixed sequences, the present study aims to suggest a starting-point for the classification of clusters and the identification of local textual functions as stylistic features in fiction.

The corpora that I used for my studies are a Dickens Corpus containing twenty-three texts and about 4.5 million words, and a corpus of twenty-nine texts by eighteen authors from the nineteenth century, also containing about 4.5 million words. Most of the texts are novels, but there are also shorter stories such as *Sketches by Boz* or *The Chimes*. A list of the texts in the corpora is provided in Appendix A. The texts for the two corpora are taken from Project Gutenberg. Project Gutenberg texts do not have to conform to consistent standards for the preparation of electronic

[3] See also Mahlberg (2005) on functions of high-frequency nouns.

texts,[4] and the editions that are chosen as a basis for the electronic version may not measure up to standards of editorial scholarship. The reasons for using these texts – in spite of potential problems with their quality – are mainly practical. Not all e-books or electronic text collections are freely available, or in a format that can be readily used with standard corpus tools such as *WordSmith Tools* (Scott, 2004). In this study, any problems with the Project Gutenberg texts did not appear to be too damaging.[5]

The twenty-five most frequent three-, four- and five-word clusters in the Dickens corpus, computed by *WordSmith*, are shown in Table 1. Contracted forms are treated as single words, so 'I don't know' (three-word cluster number 6), for instance, is counted as a three-word cluster. The decision to include apostrophes in words is arguable[6] and, at this point, a purely practical decision. Clusters are mainly regarded as pointers to textual functions so that a separation according to the exact numbers of words is less relevant than functional similarities between clusters.

Table 1 illustrates how the frequency of clusters decreases as their length increases. Biber *et al*. (1999: 992) point out that three-word clusters (or 'lexical bundles') are extremely common, because they are 'a kind of extended collocational association'. Longer clusters are, by contrast, 'more phrasal in nature and correspondingly less common' (Biber *et al*., 1999: 992). From Table 1 we can also see that longer clusters are more likely to be restricted to specific texts, a point that is also made by Stubbs and Barth (2003: 76). Of the twenty-five top three-word clusters, twenty occur in all twenty-three texts. Of the four-word clusters, only four occur in all texts, and none of the five-word clusters occur in all twenty-three texts. Different factors play a role in accounting for the distribution of clusters. The three-word cluster 'said Mr Pickwick' (number 21) occurs in one text only, but it is still relatively frequent, which is because Mr Pickwick is a character with a strong and lively presence throughout the whole of the *Pickwick Papers*. The five-word cluster 'the Father of the Marshalsea' (number 25) in *Little Dorrit* is also linked to a single character. As a unit, the cluster functions as a name, but it only occurs forty-five times. In general, leaving aside for a moment those clusters that contain names (or are names), we can say that

[4] General information on the creation of Project Gutenberg texts can be found at: http://www.gutenberg.org/howto/spd-howto. For a survey of free e-Books see Berglund *et al*. (2004).

[5] This study has been conducted in the context of a wider project on the investigation of clusters in the Dickens corpus and in 19C, where issues about the quality of the Gutenberg texts that are encountered in the course of the research are documented. The figures in this paper have to be seen as initial results. In the course of further research, typos or inconsistencies in the Gutenberg texts might be discovered and amended, and the continuous development of *WordSmith* might also affect details of quantitative information.

[6] Biber *et al*. (1999) count contracted forms as a single word in their description of lexical bundles. By contrast, William Fletcher's PIE tool follows the BNC tagging and tokenises contractions and possessives with apostrophes as separate units, see http://pie.usna.edu/POScodes.html.

	three-word clusters	F	T	four-word clusters	F	T	five-word clusters	F	T
1	out of the	1,209	23	as if he were	452	21	as if he had been	90	20
2	as if he	1,157	23	as if he had	302	22	his hands in his pockets	90	20
3	there was a	1,091	23	at the same time	289	22	in the course of the	88	18
4	it was a	1,050	23	in the course of	263	21	what do you mean by	73	18
5	one of the	1,005	23	as if it were	249	23	as if it were a	72	18
6	i don't know	1,001	23	in the midst of	234	23	the opposite side of the	70	19
7	that he was	959	23	in a state of	232	22	a quarter of an hour	66	17
8	the old man	941	20	for the first time	231	23	at the bottom of the	66	19
9	that it was	852	23	with an air of	225	18	on the part of the	65	17
10	that he had	808	23	i don't know what	224	23	what do you think of	64	16
11	he had been	783	22	i beg your pardon	222	20	in the middle of the	62	19
12	i am not	720	23	on the part of	222	20	with his hands in his	60	20
13	would have been	692	23	it would have been	218	22	as if it had been	58	18
14	i am sure	677	23	said the old man	214	13	at the top of the	56	18
15	what do you	677	23	what do you mean	208	21	i beg your pardon sir	56	16
16	if he had	670	23	in a low voice	200	20	on the opposite side of	54	17
17	i have been	640	23	up and down the	178	20	on the other side of	53	17
18	it was not	635	23	the top of the	171	20	at the end of the	52	15
19	at the door	624	23	as if they were	169	21	as a matter of course	51	15
20	it would be	605	22	i have no doubt	168	18	as much as to say	50	16
21	said mr pickwick	586	1	if he had been	168	22	the other side of the	50	17
22	in the same	559	23	i should like to	162	22	up and down the room	48	13
23	if he were	550	22	in the way of	162	19	with the air of a	46	14
24	there was no	527	23	for a long time	161	20	as if he were a	45	16
25	might have been	517	23	as if she were	152	21	the father of the marshalsea	45	1

F = frequency of cluster, T= number of texts in which the cluster occurs

Table 1: Top twenty-five three-, four- and five-word clusters in Dickens

the longer the cluster, the more likely it is to link to a particular text. The maximum cluster length that *WordSmith* worked with when this study was done was eight words. Among the eight-word clusters we find ‘the young man of the name of Guppy’ from *Bleak House*, or ‘the monotony of bells

and wheels and horses' feet' from *Dombey and Son*, which occur exclusively in the respective novels. There are also clusters that are longer than eight words, which can be identified by concordancing eight-word clusters. The following sequence of twelve words is used by Mr Bagnet in *Bleak House* and characterises his relationship with his wife: 'But I never own to it before her. Discipline must be maintained.' Such phrases have been noticed and discussed by literary critics and the computer can help to trace and analyse them systematically.

Shorter clusters that are more flexible and more frequent are more difficult to characterise. One option is to classify clusters according to structural criteria. Following Biber *et al*.'s (1999: 1001 ff.) classification of clusters, Table 1 contains, for instance, clause fragments consisting of a subject pronoun followed by a verb ('I don't know what', four-word cluster number 10), noun phrase expressions ('the opposite side of the', five-word cluster number 6), prepositional phrase expressions ('at the same time', four-word cluster number 3), *etc*. In their study, Biber *et al*. (1999) focus on frequent repetitions in large amounts of data and look at the distribution of lexical bundles across different registers. The present study deals with much smaller quantities of data in a corpus of texts of a similar type, and focusses on functional criteria. Still, a comparison of features across texts provides useful insights. The following sections will look at clusters in Dickens compared to the 19C corpus. In common with the Dickens corpus, the 19C corpus contains about 4.5 million words, which allows some initial comparisons without the need for normalised figures.

4. Frequent clusters and key clusters

In the following, I concentrate on five-word clusters. Five-word clusters are still flexible enough to occur across a number of different texts in the Dickens corpus; however, at the same time their frequencies are sufficiently manageable to allow for a detailed analysis (the cut-off point of five is arbitrary). In the Dickens corpus we find 4,904 different types of five-word clusters that occur a minimum of five times compared to 3,409 clusters in the 19C corpus. Table 2 presents the top twenty-five five-word clusters in both corpora. The table shows how the most frequent clusters in the two corpora overlap. The thirteen clusters in bold appear among the top twenty-five in both lists.

In general word-frequency lists, high frequency words tend to be grammatical words. In the cluster lists we also find sequences of grammatical words such as 'as if he had been' or 'as if it were a'. Furthermore, both lists illustrate clusters that function as the beginning of noun phrases or prepositional phrases such as, 'the other side of the', 'in the middle of the' and 'at the end of the'. It can be argued that although these clusters contain nouns, which are usually regarded as content words, the clusters function primarily as predetermining elements for what follows.

The high-frequency clusters in Table 2 are mainly time and/or place clusters. Most of them are parts of longer expressions, but there are also examples that are more structurally complete ('a quarter of an hour', 'up and down the room'). Other clusters in Table 2 reflect the fact that the texts contain speech ('I beg your pardon sir', 'I am sorry to say'). It is also noticeable that both lists contain clusters beginning with *as if*. In common with Speech clusters, As If clusters appear to reflect the fact that the texts in the corpora are fictional, but we also find clusters that are common in English in general. For instance, the cluster 'at the end of the', is frequent

	Dickens			*19C*		
Rank		*Freq.*	*Texts*		*Freq.*	*Texts*
1	**as if he had been**	90	20	**a quarter of an hour**	106	25
2	his hands in his pockets	90	20	**at the end of the**	83	24
3	**in the course of the**	88	18	**the other side of the**	81	21
4	what do you mean by	73	18	**in the middle of the**	79	21
5	as if it were a	72	18	**in the course of the**	72	20
6	the opposite side of the	70	19	in the direction of the	72	15
7	**a quarter of an hour**	66	17	**on the other side of**	56	21
8	**at the bottom of the**	66	19	the other end of the	55	19
9	on the part of the	65	17	**as if it had been**	54	17
10	**what do you think of**	64	16	i should like to know	53	19
11	**in the middle of the**	62	19	at the back of the	51	16
12	with his hands in his	60	20	for the first time in	50	16
13	**as if it had been**	58	18	**what do you think of**	50	17
14	**at the top of the**	56	18	**as if he had been**	49	18
15	i beg your pardon sir	56	16	**as a matter of course**	47	12
16	on the opposite side of	54	17	as if she had been	46	13
17	**on the other side of**	53	17	**at the bottom of the**	45	20
18	**at the end of the**	52	15	**up and down the room**	43	16
19	**as a matter of course**	51	15	at the door of the	40	17
20	as much as to say	50	16	**at the top of the**	39	15
21	**the other side of the**	50	17	by the side of the	38	16
22	**up and down the room**	48	13	for a minute or two	36	12
23	with the air of a	46	14	i am sorry to say	36	18
24	as if he were a	45	16	and at the same time	35	14
25	the father of the marshalsea	45	1	for the first time since	35	14

Table 2: Top twenty-five five-word clusters in Dickens and the 19C Corpus

in present-day general English across different registers (see, for example, Biber *et al*., 1999: 1013, 1017) and one of the patterns in which we find the high-frequency noun *end*. In order to obtain precise information on clusters that can be regarded as 'mainstream' in nineteenth-century English, further information from a corpus of non-fiction would be needed.

A useful starting-point for a functional analysis of clusters in Dickens compared to 19C is the identification of 'key clusters'. *WordSmith* calculates key clusters in the same way as 'key words'. Key words are based on the comparison of 'simple' word lists, i.e., word lists containing

words, not clusters. Corresponding to the calculation of key words (see Scott 2004–6: 118), *WordSmith* calculates the keyness of a cluster by looking at:

- the frequency of the cluster in Dickens
- the number of five-word clusters in Dickens (i.e., cluster tokens, not types)
- the frequency of the cluster in 19C
- the number of five-word clusters in 19C

and cross-tabulates these. This comparison generates 'positive' key clusters, i.e., clusters that occur more often in Dickens than would be expected by chance in comparison with 19C as a reference corpus, and 'negative' key clusters, i.e., clusters which occur less often than would be

Rank	*Cluster*	*D Freq.*	*D Texts*	*19C Freq.*	*19C Texts*	*Keyness*
1	his hands in his pockets	90	20	13	8	65.07
2	the father of the marshalsea	45	1	0	0	62.60
3	the person of the house	37	3	0	0	51.47
4	do me the favour to	32	12	0	0	44.52
5	as if he would have	41	15	2	2	43.62
6	what do you mean by	73	18	15	11	41.92
7	with his hands in his	60	20	12	7	35.17
8	go so far as to	24	13	0	0	33.39
9	i beg your pardon sir	56	16	11	9	33.27
10	how do you find yourself	23	11	0	0	31.00
11	as if he were a	45	16	7	4	31.19
12	hands in his pockets and	40	17	5	4	31.16
13	with his hand to his	31	10	2	2	30.80
14	on the part of mr	34	11	3	3	30.62
15	who had by this time	22	10	0	0	30.61
16	the lady of the caravan	22	1	0	0	30.61
17	on the top of his	21	12	0	0	29.21
18	the old man with a	21	6	0	0	29.21
19	on the part of the	65	17	18	13	28.49
20	how do you do mr	29	11	2	2	28.28
21	as if he were going	32	12	3	2	28.19
22	captain gills said mr toots	20	1	0	0	27.82
23	upon my word and honour	25	8	1	1	27.68
24	beg your pardon sir said	25	6	1	1	27.68
25	as if it were a	72	18	23	12	26.77

Table 3: Top twenty-five five-word key clusters based on a comparison of the Dickens corpus with 19C

expected in comparison with 19C. Key clusters can be calculated with the *KeyWords* tool or in *WordList* with the 'comparing wordlists' function. To calculate the keyness in Table 3 the *WordList* function has been used because it compares all the clusters in both lists. So the four steps

described above take into account all clusters in both lists. The *KeyWords* tool, by contrast, is uni-directional and focusses on words/clusters in a smaller text compared with a larger reference corpus. It does not provide information on words that occur only in the reference corpus but not in the smaller text. This procedure can be problematic when the texts are of similar size. Table 3 shows the top twenty-five key clusters resulting from this comparison.

Columns three and five respectively contain the frequency of a cluster in the Dickens corpus ('D Freq.'), and the frequency of the cluster in the 19C corpus ('19C Freq.'); the final column gives the keyness. The clusters are ordered according to their keyness. Table 3 also shows the number of texts in which a cluster is found ('D Texts', '19C Texts'). This information has been added for the interpretation of the results; it does not play a part in the calculation of keyness. An initial overview shows that there are three types of positive key clusters. First, there are clusters that are frequent in Dickens but do not occur in 19C, so their frequency is high. In this group we can distinguish between clusters that occur in different texts across the Dickens corpus, or those that appear in only one Dickens text, such as the second cluster in the list, 'the Father of the Marshalsea', which is restricted to *Little Dorrit*; cluster number 16, 'the lady of the caravan', which occurs only in *The Old Curiosity Shop*; and, cluster number 22, 'Captain Gills said Mr Toots', which is found only in *Dombey and Son*. Finally, there are key clusters that are frequent in Dickens and occur in 19C, too, but comparatively less frequently than in Dickens. An example is the first cluster in the list, 'his hands in his pockets', which occurs ninety times and in twenty texts in Dickens, compared to thirteen times and eight texts in 19C.

The number of key clusters that *WordSmith* generates depends on the choice of a significance value. The present key clusters were calculated with a *p*-value of 0.00001. The smaller the *p*-value, the fewer clusters are found, i.e., the clusters are statistically more significant. With the given *p*-value, *WordSmith* finds sixty-six positive and seven negative key clusters, with a positive keyness ranging from 19.86 to 65.07 and a negative keyness between –34.78 and –20.87. A negative key cluster is, for instance, 'the Prince of Little Lilliput', which occurs twenty-five times in *Vivian Grey* in 19C, but does not occur in Dickens. The way in which a keyness value is interpreted depends on the corpora that are to be compared and the questions to be answered (see, for instance, Scott, 2006: 63 ff., on the effects of different reference corpora). When we look at key clusters we deal with much smaller numbers than we would for key words, but, at the same time, the repetition of a sequence of words has significance in itself. However, the cut-off points in this study are exploratory. The *p*-value of 0.00001 was set because a smaller *p* of 0.000001 only produces thirty-one positive and three negative key clusters, which seemed too small a sample for the functional analysis. A bigger *p*-value of 0.0001 produces 136 positive key clusters and twenty-eight negative key clusters; and it was

found that the additional clusters did not contribute further, major functional groups, so the sixty-six key clusters were taken to provide an initial overview to complement the observations on the top frequency clusters in Table 2.

5. Groups of clusters

A classification of clusters can follow a number of different criteria. One option is a structural classification, taking into account the grammatical features of the clusters, as was hinted at in the previous section. Another option is to look at the functions that the clusters fulfil in texts. The groups of clusters described in this section were identified in a bottom-up fashion with no *a priori* assumptions. Furthermore, the groups are dynamic, i.e., the description developed in the course of the analysis as more detailed observations required adjustments of the criteria, and when more studies have been conducted further adjustment may become necessary. We will first look at functional groups of key clusters before the frequent clusters described in Table 2 are taken into account as well. The sixty-six positive key clusters fall into five groups:

1. Labels
2. Speech clusters
3. As If clusters
4. Body Part clusters
5. Time and Place clusters

The following provides a brief overview of each of the groups.

5.1 Labels

The clusters in the Labels group are, or contain, the names of characters, e.g., 'Mr Pickwick and his friends', or in the case of 'The Six Jolly Fellowship Porters', a name for a place: 'a tavern of a dropsical appearance' (*Our Mutual Friend*). There are also more general Labels such as 'man of the name of'. In Table 4, bold print is used to highlight general expressions, i.e., expressions that occur in more than one Dickens text, although they may not occur in 19C at all. Furthermore, the Labels group contains phrases that, superficially, do not look like a name or referring expression, but are still associated with one particular character or concept. These clusters occur in one text only. We find, for instance, clusters that are parts of Mr Snagby's phrase 'not to put too fine a point (up)on it' in *Bleak House*, and the phrase 'How not to do it' occurs in *Little Dorrit*

characterising a 'great political science'. We can observe the following relationship between quantitative and qualitative information: the clusters in the Labels group tend to be linked to particular characters and concepts – most of them occur only in one Dickens text, and not at all in 19C.

Number of types: 20	*D*	*19C*
the father of the marshalsea	45	0
the person of the house	37	0
the lady of the caravan	22	0
the old man with a	21	0
captain gills said mr toots	20	0
mr pickwick and his friends	19	0
my dear said the jew	19	0
gentleman in the white waistcoat	18	0
man of the name of	22	1
the gentleman in the white	17	0
the six jolly fellowship porters	16	0
my dear said mrs nickleby	16	0
mr winkle and mr snodgrass	16	0
how not to do it	16	0
my lovely and accomplished relative	16	0
young man of the name	15	0
put too fine a point	15	0
man with the wooden leg	15	0
not to put too fine	15	0
to put too fine a	15	0

Table 4: Labels

5.2 Speech clusters

The clusters in this group contain a first or second person pronoun or possessive, which is taken as an indication of interaction. Further analysis would be necessary to distinguish between different types of speech and thought presentation associated with these clusters. When clusters contain features of both Labels and Speech, the Labels category overrules Speech; an example is 'my dear said Mrs Nickleby' (in Table 4 above), which both contains a first person possessive and a name, and is thus classified as Label. In contrast with the Labels group, Speech key clusters are more likely to occur in 19C, too, and not only in Dickens.

Number of types: 14	*D*	*19C*
do me the favour to	32	0
what do you mean by	73	15
i beg your pardon sir	56	11
how do you find yourself	23	0
how do you do mr	29	2
upon my word and honour	25	1
beg your pardon sir said	25	1
what i am going to	29	3
am glad to see you	24	2
i am glad to see	29	4
what do you want here	15	0
how do you do sir	23	2
will you allow me to	23	2
you be so good as	19	1

Table 5: Speech clusters

5.3 As If clusters

The next group contains clusters that start with *as if*. One of the clusters in this group begins with *if*, but a concordance of the cluster shows that twenty-four of the twenty-six occurrences are part of the six-word clusters 'as if he were going to', and only two are examples of indirect speech ('inquired if…', 'asked Paul if…'). All of the As If key clusters occur in 19C as well as in Dickens.

Number of types: 5	*D*	*19C*
as if he would have	41	2
as if he were a	45	7
as if he were going	32	3
as if it were a	72	23
if he were going to	26	3

Table 6: As If clusters

5.4 Body Part clusters

Body Part clusters contain at least one noun referring to a part of the human body, and in most cases this is the noun *hand*.

Number of types: 9	*D*	*19C*
his hands in his pockets	90	13
with his hands in his	60	12
hands in his pockets and	40	5
with his hand to his	31	2
his head as if he	18	0
laying his hand upon his	22	1
the palms of his hands	17	0
his head on one side	30	4
her hand upon his shoulder	15	0

Table 7: Body Part clusters

5.5 Time and Place clusters

The clusters in the fifth group contain a nominal time or place expression with or without a preposition. Among the key clusters there are no time expressions, as Table 8 shows, but the top frequency clusters (Table 2) include time expressions so the functional group Time and Place will cover both types. There are clusters in this group where the noun can not be clearly classified as either time or place ('after a great deal of'), but the cluster as a whole functions as a time or place expression.

Number of types: 5	*D*	*19C*
on the top of his	21	0
at the upper end of	23	1
on the opposite side of	54	15
after a great deal of	16	0
the opposite side of the	70	26

Table 8: Time and Place clusters

In addition to these five groups, there are thirteen clusters that do not fit into either of the groups. Among them, some could be regarded as borderline categories, for instance, 'on the part of Mr' or 'on the part of the' are borderline Time and Place, but for the purposes of this study I will not provide more detailed classifications. Returning to the top twenty-five clusters in Table 2, we see that the Time and Place category collects most examples, whereas Labels make up the biggest group for the key clusters. When we compare Dickens and 19C with regard to the top twenty-five clusters, for 19C most clusters fit into the Time and Place group, there are also examples of Speech and As If clusters, but there are no Labels or Body

Part clusters. Still, the key clusters illustrate that there are examples of Body Part clusters in 19C, and the negative key clusters contain Labels ('the Grand Duke of Reisenburg' in *Vivian Grey* or 'Sir Percival and the Count' in *The Woman in White*).

The five functional groups have been identified on the basis of key clusters and top frequency clusters. A point for further work is to investigate to which extent the five categories can be applied to cover all the five-word clusters in individual texts (see Mahlberg, forthcoming a, b). Although the numbers on which the present classification is based are relatively small, there are several reasons why the groups seem to provide a useful starting-point. First, frequencies for five-word clusters are not high, so if we account for the most frequent clusters a number of important patterns will be covered. Secondly, from the point of view of literary stylistics or criticism, clusters that are functionally interesting are those that mark patterns which are specific to individual texts. Such clusters are accounted for by the Labels group. Thirdly, the observations are based on a number of texts and take comparisons into account, which ensures that the categories are not only restricted to the situation in one particular text. Fourthly, as we will see in the following, links between the functional groups and arguments put forward by literary criticism lend support to the classification. But first it is useful to look at the local nature of the categories with the help of Figure 1.

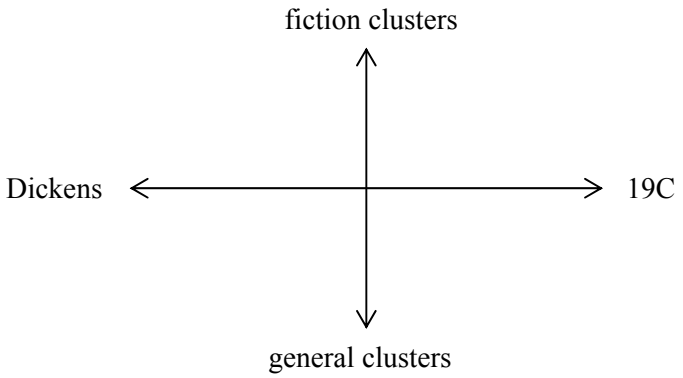

Figure 1: Dimensions of variation of clusters as pointers to local textual functions

The fact that the five groups of clusters capture both key clusters and frequent clusters shows different aspects of localness. Clusters that are key are relatively more frequent in the Dickens corpus than in 19C. The way in which keyness for clusters is interpreted needs closer investigation and it seems best to describe the use of clusters in terms of tendencies on a continuum. At one end, we have clusters that are more specific to Dickens, or even specific to one particular Dickens text. Here the columns in Table

3 that provide information on the distribution of clusters across texts help to specify the interpretation of keyness in Dickens. At the other end, we have clusters that are more specific to 19C: these clusters are negative key clusters with regard to Dickens. The *x*-axis in Figure 1 illustrates this relationship. For the continuum of clusters we also have to add a further dimension, illustrated by the *y*-axis in Figure 1: there are clusters that seem to be more typical of fiction compared to a more general variety of nineteenth-century texts. Although we are more likely to find Time and Place clusters that occur in fiction as well as across a variety of texts, Body Part, Labels, As If and Speech clusters are categories that seem to leave more room for fiction-specific patterns.

Crucial to this study is the role that key clusters play in the identification of the local textual functions associated with the five groups of clusters. On the one hand, when we only look at clusters in a single text the categories for classification might be more specific to the text, i.e., more local, but it might be more difficult to then relate them to a broader framework for the analysis of literary texts. On the other hand, if we want to account for a larger number of clusters across a number of texts the functions we identify become more general and less local. Therefore, key clusters provide a starting-point to identify local textual functions in Dickens. The functions can be found to some extent in the works of other nineteenth-century writers and their relevance in an individual Dickens text can also vary. In particular, the group of Time and Place clusters shows that some Time and Place clusters can be key whereas others are among the top frequency clusters in both Dickens and 19C. Therefore, the five groups of clusters provide a broad indication of sets of functions. These functions can then be more, or less, characteristic of (or key to) a specific text or groups of text. As Figure 1 illustrates, the dimensions of interpretation for clusters have to be seen in relation to the selection of texts under investigation. It will be a task for future work to add detail to frequency measures for key clusters. The present study now focusses on a more detailed discussion of the type of functions that the key clusters helped to identify. Since key clusters are the extreme points in a broad group of similar clusters, the following discussion just uses the term ‘clusters’ to underline the broad nature of a functional group of clusters.

6. Functions of clusters and literary criticism

The basic functions of the five groups of clusters are: characterising people, places and things in the stories (Labels), expressing interaction between the characters (Speech clusters), describing looks and movements (Body Part clusters), creating textual worlds by comparison and contrast (As If clusters), and locating and relating actions in time and place (Time and Place). I cannot discuss all five groups in detail due to limited space. So, in this section, I shall concentrate on the characteristics of Labels and As If

clusters, as two groups that strongly point to links in literary criticism, and then focus in more detail on Body Part clusters in Section 6.

What is important to all groups is that to get a more detailed picture, it is necessary to look at concordances of the clusters and their functions in broader textual contexts. Take the cluster 'the old man with a', for instance. It is classed as a Label, because it contains a noun that refers to a character. Out of context, we may assume that *with* introduces a postmodifier that provides details characterising the man. However, a concordance reveals that in the majority of the twenty-one examples the prepositional phrase accompanies the act of speaking, as in the following two examples:

(1) 'Pretty, pretty, pretty!' said the old man with a clap of his hands.
(*Our Mutual Friend*)

(2) 'IF!' exclaimed the old man, with a look of excessive contempt.
(*Pickwick Papers*)

The functional patterns of 'the old man with a' are linked to the fact that this cluster occurs in six different texts and is not a label for one character only.

The present approach is corpus-driven, but it is clear that links to literary criticism and previous stylistic analysis play an important role, also with regard to the social context and the reception of literature. The Labels group links in with well-known observations on the language of Dickens. As Quirk (1961) points out, phrases to individualise characters are striking in Dickens, and, 'the use of this well-established dramatic device was an obvious desideratum to a writer who worked by means of serial publication, since it provided the reader with a most immediate means of recall and identification' (Quirk, 1961: 20 ff.). Examples of such individualising phrases in the Labels group are Mr Snagsby's favourite phrase mentioned above, or the cluster 'my lovely and accomplished relative' used by Cousin Feenix in *Dombey and Son*. Examples from the groups of Speech and Body Part clusters can also function to individualise characters. However, in terms of the key clusters none of the Speech clusters and none of the Body Part clusters are limited to a single novel. Individualising functions for clusters of these groups become evident when we analyse clusters in a single text, as we will see below for the example 'his head on one side'.

Another link to literary criticism is suggested by the As If clusters. Literary critics have pointed out particular functions of the 'fanciful as if'. 'It generally takes the form of the invention of some improbable but amusing explanation of the appearance or behaviour of one of the characters in a novel' (Brook, 1970: 33). One of Brook's (1970: 32) examples is from *Little Dorrit* where a woman is described as being 'in such as tumbled condition altogether, that it seemed as if it would be an act

of kindness to iron her'. According to Brook (1970: 34), the fanciful comparison is not dependent on the presence of *as if*, but can also appear in other forms, as in an example from *Our Mutual Friend* where Mrs Wilfer talks to her husband about their daughter '*with a lofty air of* never having had the least co-partnership in that young lady: of whom she now made reproachful mention as an article of luxury which her husband had set up entirely on his own account and in direct opposition to her advice'.

In literary criticism, striking examples can easily receive attention. With the help of corpus linguistic tools and descriptive categories such striking examples can be seen as part of a bigger picture. It is possible to provide comprehensive accounts of repeated forms and compare related occurrences. We see, for instance, that different As If structures can be distinguished. Whereas 'as if he had been' and 'as if it had been' are frequent in both Dickens and other 19C fiction, the key clusters show examples of As If that are relatively more frequent in Dickens. A more detailed account of the patterns that follow *as if* can then add detail to the functions (see Mahlberg, forthcoming b). A cluster analysis can also help to achieve a more systematic overview by finding other forms of comparison in addition to the As If clusters. In Table 2, we find the clusters 'as much as to say' and 'with the air of a' that fulfil functions similar to As If clusters. In example (3) 'with the air of a' introduces an amusing comment. In example (4) there is also a comment, but of a less striking type, that may not be classified as 'fanciful'. In example (5) 'as much as to say' translates a cough into direct speech and thus adds the narrator's comment to what happens in the situation. The comment with 'as much as to say' is similar to the comment 'as if she didn't quite mean that, but rather the contrary' two lines before in the same example (5), where Mrs Joe's speech is commented on by the narrator.

(3) … and offered Mr. Pickwick a pinch of snuff *with the air of a* man who had made up his mind to a Christian forgiveness of injuries sustained.

(*Pickwick Papers*)

(4) 'Beg your pardon, sir,' said the stranger, 'bottle stands—pass it round—way of the sun—through the button-hole—no heeltaps,' and he emptied his glass, which he had filled about two minutes before, and poured out another, *with the air of a* man who was used to it.

(*Pickwick Papers*)

(5) 'Mrs. Joe,' said I, as a last resort, 'I should like to know – if you wouldn't much mind – where the firing comes from?'

'Lord bless the boy!' exclaimed my sister, *as if she didn't quite mean that*, *but rather the contrary*. 'From the Hulks!'

'Oh–h!' said I, looking at Joe. 'Hulks!'

Joe gave a reproachful cough, *as much as to say*, 'Well, I told you so.'

(*Great Expectations*)

On the whole, the As If group collects clusters that compare or comment on actions and situations in a story, and thus contribute to the creation of a textual world. Within the group some examples can be more striking and closer to what critics have identified as the fanciful *as if* than others. Moreover, the five-word clusters that fall into this group are not the only ways of expressing As If functions. In example (5), the *as if* introduces a construction that is not a frequent five-word cluster: the sequence 'as if she didn't quite' occurs just once in the Dickens corpus. The As If group illustrates that a description of local textual functions can characterise different functional facets within a broad group that is identified on the basis of repetition on the textual surface. The following section will discuss this point further by looking at examples of Body Part clusters.

7. Body Part clusters

The broad function of Body Part clusters is to describe the appearance and behaviour of characters. Examples of Body Part clusters can also be covered to some extent under the headings of body language or non-verbal communication, as outlined in Korte (1997). The discussion here will only occasionally link to descriptions of body language in literature. The focus is on categories that are based on the clusters under investigation. Both in Dickens and in 19C there are two basic functions of Body Part clusters: a 'contextualising' function and a 'highlighting' function. The borderline between the two is not clear-cut but fuzzy. Examples of the contextualising function occur together with other activities, often speech, which are more central to the story, as in the examples of 'his hands in his pockets' below (6 to 8). The contextualising function is often found with a prepositional phrase with *with* (as in examples 6 and 8), so the grammatical function coincides with the circumstantial meaning that the cluster contributes to the narration. There is, however, no one-to-one relationship between grammatical function and function in the narration.

(6) 'Ha ha ha!' laughed the Doctor thoughtfully, with *his hands in his pockets*. 'The great farce in a hundred acts!'

(*The Battle of Life*)

(7) 'You see, Mr Richard,' said Brass, thrusting *his hands in his pockets*, and rocking himself to and fro on his stool, 'the fact is, …

(*Old Curiosity Shop*)

(8) ‘More than that, eh!’ retorted Mr. Jaggers, lying in wait for me, with *his hands in his pockets*, his head on one side, and his eyes on the wall behind me; ‘how much more?’

(*Great Expectations*)

When the clusters accompany another activity they tend to be unobtrusive and the activity or posture that they describe does not strike us as too extraordinary. It is possible to find several clusters with a contextualising function together, as in example (8), which illustrates ‘his hands in his pockets’, as well as ‘his head on one side’, and there is also a repeated four-word cluster, ‘his eyes on the’. Such a sequence of Body Part clusters receives more emphasis. Clusters that provide contextual information can support another activity when they illustrate a gesture that is typically associated with a particular situation. In example (9), ‘with his hand to his chin’ supports Mr Boffin’s thinking. The cluster ‘with his hand to his’ is also illustrated in example (10): Mr Smallweed has problems hearing so he puts his hand to his ear. If you put your hand on someone’s shoulder this can be an encouraging gesture, as in example (11), where Nicholas tries to comfort Smike. In example (12), ‘the palms of his hands’ is part of an expression describing nervousness, but we are not told directly that the Jew is nervous – instead his behaviour is described: he is rubbing the palms of hands *nervously*:

(9) ‘Let me see then,’ resumed Mr Boffin, *with his hand to his* chin. ‘It was Secretary that you named; wasn’t it?’

(*Our Mutual Friend*)

(10) ‘Eh? What do you say I have got of my own?’ asked Mr. Smallweed *with his hand to his* ear.

(*Bleak House*)

(11) ‘Hush!’ said Nicholas, *laying his hand upon his* shoulder. ‘Be a man; you are nearly one by years, God help you.’

(*Nicholas Nickleby*)

(12) ‘Regarding this boy, my dear?’ said the Jew, rubbing *the palms of his hands* nervously together. ‘The boy must take his chance …

(*Oliver Twist*)

Examples of the contextualising function show how a narration includes elements that add to the creation of a lively story. The characters are not only involved in conversation, mediation or activities that carry the story forward, but they also have human features that are less central to the story at a particular moment and that contribute to the ongoing characterisation. The description of such features and movements helps the reader to

visualise situations contained in the story. As Korte (1997: 189f.) points out, details of body language can add to the realism of a novel. For the contextualising function the important point is that the description does not strike the reader as particularly unusual. A habit or action is described that readers can visualise without problems and that they may not even be particularly aware of in the process of reading. The following examples of 'his hands in his pockets' show that the contextualising function is not restricted to the writing of Dickens, but is also found in the 19C corpus:

(13) Mr. Earnshaw vouchsafed no answer. He walked up and down, with *his hands in his pockets*, apparently quite forgetting my presence; and his abstraction was evidently so deep, and his whole aspect so misanthropical …

(*Wuthering Heights*)

(14) He sauntered on, with *his hands in his pockets*, humming the chorus of a comic song.

(*Armadale*)

(15) 'Oh, well,' said Lush, rising with *his hands in his pockets*, and feeling some latent venom still within him, 'if you have made up your mind!—only there's another aspect of the affair…

(*Daniel Deronda*)

The unobtrusive character of such Body Part clusters is also reflected by the fact that they seem to have common collocations or collocating clusters. For instance, to 'walk up and down' (example 13), is an activity that is also found in Dickens together with 'his hands in pockets' in situations where a character seems to be thinking about something. According to Korte (1997: 196) the description of practical actions to portray mental states has been a common device in literature. In the nineteenth century, body language was also increasingly used for characterisation and Dickens' notorious use of body language has received much attention (see also Korte, 1997: 135). The identification of body part clusters provides further evidence of the importance of body language in Dickens.

When we look at the clusters in context we can see how the contextualising function differs from the highlighting function. In the following example, the narrator comments on the circumstantial information thereby giving it more emphasis; the expression 'looking as much unlike a man in a hurry as possible' is similar to comparsions with As If clusters:

(16) '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*)

A point made with the help of a Body Part cluster can be even more complex and stretch over several sentences, as in the following example from *Barnaby Rudge*. The cluster 'his hands in his pockets' is part of the description of Joe's miserable situation. Although Joe is given a hard time by John, he stays surprisingly calm:

> (17) In short, between old John and old John's friends, there never was an unfortunate young fellow so bullied, badgered, worried, fretted, and brow-beaten; so constantly beset, or made so tired of his life, as poor Joe Willet. This had come to be the recognised and established state of things; but as John was very anxious to flourish his supremacy before the eyes of Mr Chester, he did that day exceed himself, and did so goad and chafe his son and heir, that but for Joe's having made a solemn vow to keep *his hands in his pockets* when they were not otherwise engaged, it is impossible to say what he might have done with them.
>
> (*Barnaby Rudge*)

The previous two examples show how Body Part clusters can not only provide contextual information but also help to highlight a particular point of the description. The highlighting function also accounts for Body Part clusters that are associated with a particular character. Let us consider the cluster 'with his hand to his', which occurs thirty-one times in Dickens. A concordance of the cluster shows that for eleven of these cases the final *his* is followed by *forehead*, and the cluster 'with his hand to his forehead' occurs eight times linked to Twemlow in *Our Mutual Friend*. Similarly, 'his head on one side' occurs thirty times altogether in Dickens, of which eight occurrences are found in *David Copperfield*, where the cluster is linked five times to Mr Chillip. Below is an extract from *David Copperfield*:

> (18) He was the meekest of his sex, *the mildest* of little men. He sidled in and out of a room, to take up the less space. He walked as softly as the Ghost in Hamlet, and more slowly. He carried *his head on one side*, partly in modest depreciation of himself, partly in modest propitiation of everybody else. It is nothing to say that he hadn't a word to throw at a dog. He couldn't have thrown a word at a mad dog. He might have offered him one gently, or half a one, or a fragment of one; for he spoke as slowly as he walked; but he wouldn't have been rude to him, and he couldn't have been quick with him, for any earthly consideration.
>
> Mr. Chillip, looking *mildly* at my aunt with *his head on one side*, and making her a little bow, said, in allusion to the jewellers' cotton, as he softly touched his left ear: 'Some local irritation, ma'am?'
>
> (*David Copperfield*)

In this short passage, 'his head on one side' occurs twice. In the first occurrence, the cluster is grammatically more prominent than in the second: 'his head' is the object of the main verb, whereas in the second occurrence the whole cluster is part of a prepositional phrase. Furthermore, the first occurrence is part of a longer description of Mr Chillip and the fact that he carried his head on one side is commented on by the narrator, David, who sees Chillip's habit as a reflection of both 'modest depreciation of himself' and 'modest propitiation of everybody else'. Then, in the next paragraph, Mr Chillip speaks, and his speech is accompanied by a description of his body language, picking up on the previous characterisation with the adverb *mildly* that refers to 'the mildest' in the first line, and the repetition of 'his head on one side'. Thus on the one hand, the second occurrence of the Body Part cluster takes a less central role with regard to the action of the story at this particular point: it accompanies speech. On the other hand, the characterisation that is associated with it is stressed through repetition within a fairly short passage. Mr Chillip's habit is also emphasised when we see how other characters perceive him and notice his way of having his head on one side (example 19). Although David's aunt cannot remember Chillip's name correctly, she still recalls (as the reader will probably, too) his habit of having his head on one side:

(19) 'That little man of a doctor, with *his head on one side*,' said my aunt, 'Jellips, or whatever his name was, what was he about? …
(*David Copperfield*)

To summarise, we can describe the functions of Body Part clusters as follows: Body Part clusters can provide contextual information that accompanies the description of a situation or activity which is more central to the story. Body Part clusters can also be a central part of a description and can highlight habits or behaviour of a character. Differences between these functions are a matter of degree and can depend on a variety of factors. We have looked at grammatical features, repeated links to a specific character and comments by the narrator. The examples from *Wuthering Heights*, *Armadale* and *Daniel Deronda* illustrate that the contextualising function is not only limited to Dickens. We can also find examples of the highlighting function in 19C. The cluster 'his head on one side' occurs four times in 19C; all four examples are from *The Mill on the Floss* and three of them are associated with Mr Tulliver. Similarly, three of the thirteen examples of 'his hands in his pockets' are from *Armadale*, and all three are linked to the character Allan. There are also examples where the narrator comments on a character's behaviour, as in the example above from *Wuthering Heights* (13), where *apparently* and *evidently* signal an interpretation of the behaviour described. In the examples below, 'in a dubiative manner' (20) and 'as if he would' (21) introduce similar comments:

(20) 'No,' said Tom, opening his pocket-knife and holding it over the puff, with his head on one side *in a dubitative manner*.
(*The Mill on the Floss*)

(21) 'I cannot!' Sir Michael lifted his hand *as if he would* command his nephew to be silent, but that imperious hand dropped feeble and impotent at his side.
(*Lady Audley's Secret*)

This analysis of examples leads to the following interpretation of the keyness of Body Part clusters. Body Part clusters offer a variety of textual functions. Although functions that we identify in Dickens might also be present in the writing of other nineteenth-century authors, it seems that Dickens makes more extensive use of the variety of such functions. The numbers of clusters are comparatively low and have to be interpreted with caution, but they seem to indicate tendencies. In addition to overall frequencies a closer look at the distribution across texts is also important. Dickens' preference for Body Part clusters seems to be supported by the following observation: the highest number of occurrences for which a cluster is linked to a specific character is three in 19C; in Dickens, however, we have already seen the example of Twemlow with eight repetitions. When we interpret numbers, there are several considerations that need further investigation. One reason why Body Part key clusters are more frequent in Dickens than in 19C could be related to the number of male and female characters in the texts under investigation, and gender differences in body language. For the cluster with the highest keyness, for instance, no examples of a corresponding female form ('her hands in her pockets') were found in either of the corpora. The dress conventions of the time may play a part here, too. However, the example below shows that the question is not a straightforward distinction between male and female:

(22) Dolly nodded and smiled, and feeling in her pockets (there were pockets in those days) with an affectation of not being able to find what she wanted, which greatly enhanced her importance, …
(*Barnaby Rudge*)

8. Conclusion

It is assumed that literary texts have individual qualities that contribute to their literariness, even though 'literariness' is not an absolute quality (see Carter, 2004: 69). In literary stylistics, linguistic tools are used to describe these qualities. In this paper, I have argued that the linguistic tools suggested by corpus linguistics are applicable to literary texts, too, and can therefore broaden the descriptive inventory of literary stylistics. Local textual functions were presented as a descriptive category illustrating links

between corpus theoretical arguments and issues that play a part in literary stylistics and criticism. The emphasis that, in the corpus theoretical framework, is given to local descriptions is also crucial to the study of literature. A corpus approach can provide additional detail on formal features, as we have seen with the As If clusters, for instance. By accounting exhaustively for specific linguistic forms on the textual surface, a corpus approach can also compare sets of texts. Here, the key clusters view texts by Charles Dickens in relation to other nineteenth-century literature. However, clusters, as features on the textual surface, are mainly seen as pointers to local textual functions. With the help of key clusters, functional groups were identified that provide a local view on functions in Dickens. The fact that the key clusters are more frequent in Dickens than in 19C can also be interpreted in the sense that other sets of local textual functions might be identifiable to describe works by nineteenth-century writers other than Dickens. Furthermore, the five groups of local textual functions were taken as broad groups within which some clusters can fulfil more striking and more noticeable functions than others. The functional group of Time and Place clusters is the group that most clearly shows the continuum between key clusters and clusters that are frequent in general. The functional variation within a group was further illustrated for the example of Body Part clusters by identifying the 'contextualising' and the 'highlighting' functions.

Such observations support the point that literariness is not an absolute quality. Such observations also show that corpus approaches can complement approaches in literary criticism. The latter tend to focus on striking features or specific examples, whereas the former can contribute to accounting for a range of features and functions, and relationships between these. It is important to note, however, that functions associated with clusters can also be realised by a variety of related patterns. Further studies have to investigate more flexible patterns and different patterns with similar functions. A discussion of patterns around the core 'his hands…pockets', for instance, can be found in Mahlberg (2007). The functional groups of five-word clusters discussed in this paper can be used as an initial tool for the analysis and comparison of literary texts. The categories presented here were developed in a bottom-up way to capture the data under investigation. More detailed studies of individual texts are needed to refine the categories, to specify different levels of localness, and to discuss links to literary criticism in more detail (for a description of Labels in *Bleak House* see Mahlberg, forthcoming a; on As If patterns in *Great Expectations* see Mahlberg, forthcoming b). Issues for further investigation are also quantitative questions of cut-off points and significance values, but also the relationship between clusters and the length of a text.

References

Adolphs, S. 2006. Introducing Electronic Text Analysis. A Practical Guide for Language and Literary Studies. London: Routledge.

Adolphs, S. and R. Carter. 2002. 'Point of view and semantic prosodies in Virginia Woolf's *To the Lighthouse*', Poetica 58, pp. 7–20.

Altenberg, B. 1998. 'On the phraseology of spoken English: The evidence of recurrent word-combinations' in A.P. Cowie (ed.) Phraseology. Theory, Analysis, and Applications, pp. 101–22. Oxford: Oxford University Press.

Berglund, Y., A. Morrison, R. Wilson and M. Wynne. 2004. Online. An investigation into free eBooks.
http://ahds.ac.uk/litlangling/ebooks/report/FreeEbooks.html
(last accessed: December 2005).

Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan. 1999. Longman Grammar of Spoken and Written English. Harlow: Longman.

Biber, D., S. Conrad and V. Cortes. 2004. '*If you look at* ..: Lexical bundles in university teaching and textbooks', Applied Linguistics 25 (3), pp. 371–405.

Brook, G.L. 1970. The Language of Dickens. London: Andre Deutsch.

Carter, R. 2004. Language and Creativity. The Art of Common Talk. London: Routledge.

Cheng, W., C. Greaves and M. Warren. 2006. 'From n-gram to skipgram to concgram', International Journal of Corpus Linguistics 11 (4), pp. 411–33.

Conrad, S. and D. Biber. 2005. 'The frequency and use of lexical bundles in conversation and academic prose' in W. Teubert and M. Mahlberg (eds) The Corpus Approach to Lexicography, Thematischer Teil von Lexicographica. Internationales Jahrbuch für Lexikographie 20, 2004, pp. 56–71. Tübingen: Niemeyer.

Culpeper, J. 2002. 'Computers, language and characterisation: An analysis of six characters in Romeo and Juliet' in U. Melander-Marttala, C. Östman and M. Kytö (eds) Conversation in Life and in Literature, pp. 11–30. Uppsala: Universitetstryckeriet.

Culpeper, J. and M. Kytö. 2002. 'Lexical bundles in Early Modern English dialogues. A window into the speech-related language of the past' in T. Fanego, B. Méndez-Naya and E. Seoane (eds) Sounds, Words, Texts and Change. Selected papers from 11 ICEHL, Santiago de Compostela, 7–11 September 2000, Vol. 2, pp. 45–63. Amsterdam: John Benjamins.

Hockey, S. 2000. Electronic Texts in the Humanities. Oxford: Oxford University Press.

Hoey, M. 2005. Lexical Priming. A New Theory of Words and Language. London: Routledge.

Hori, M. 2004. Investigating Dickens' Style. A Collocational Analysis. Basingstoke: Palgrave Macmillan.

Hunston, S. and G. Francis. 2000. Pattern Grammar. A Corpus-driven Approach to the Lexical Grammar of English. Amsterdam: John Benjamins.

Korte, B. 1997. Body Language in Literature. Toronto: University of Toronto Press.

Lawson, A. 2000. '"Die schöne Geschichte": a corpus-based analysis of Thomas Mann's *Joseph und seine Brüder*' in B. Dodd (ed.) Working with German Corpora, pp. 161–80. Birmingham: University of Birmingham Press.

Mahlberg, M. 2005. English General Nouns: a Corpus Theoretical Approach. Amsterdam: John Benjamins.

Mahlberg, M. 2007. 'Corpora and translation studies: textual functions of lexis in *Bleak House* and in a translation of the novel into German' in V. Intonti, G. Todisco and M. Gatto (eds) La Traduzione. Lo stato dell'arte. Translation. The State of the Art, pp. 115–35. Ravenna: Longo.

Mahlberg, M. Forthcoming a. 'Corpus stylistics: bridging the gap between linguistic and literary studies' in M. Hoey, M. Mahlberg, M. Stubbs and W. Teubert, Text, Discourse and Corpora. Theory and Analysis. London: Continuum.

Mahlberg, M. Forthcoming b. 'A corpus stylistic perspective on Dickens' *Great Expectations*' in M. Lambrou and P. Stockwell (eds) Contemporary Stylistics. London: Continuum.

McEnery, A., R. Xiao and Y. Tono. 2006. Corpus Based Language Studies. An Advanced Resource Book. London: Routledge.

Miall, D.S. 1996. Online. 'Representing and interpreting literature by computer', http://www.ualberta.ca/~dmiall/complit.htm (last accessed: August 2006), first published 1995 in the Yearbook of English Studies 25, pp. 199–212.

Project Gutenberg, 2003–2006, http://www.gutenberg.org/ (last accessed: July 2006)

Quirk, R. 1961. 'Some observations on the language of Dickens', A Review of English Literature 2 (3), pp. 19–28.

Scott, M. 2004. WordSmith Tools. Version 4.0. Oxford: Oxford University Press.

Scott, M. 2004–6. WordSmith Tools. Version 4.0. Manual. Oxford: Oxford University Press.

Scott, M. 2006. 'Key words of individual texts' in M. Scott and C. Tribble, Textual Patterns: Key Words and Corpus Analysis in Language Education, pp. 55–72. Amsterdam: John Benjamins.

Semino, E. and M. Short. 2004. Corpus Stylistics. Speech, Writing and Thought Presentation in a Corpus of English Writing. London: Routledge.

Sinclair, J. 2004. Trust the Text. Language, Corpus and Discourse. London: Routledge.

Starcke, B. 2006. 'The phraseology of Jane Austen's *Persuasion*: phraseological units as carriers of meaning', ICAME Journal 30, pp. 87–104.

Stubbs, M. 2001. Words and Phrases. Corpus Studies of Lexical Semantics. Oxford: Blackwell.

Stubbs, M. 2005. 'Conrad in the computer: examples of quantitative stylistics Methods', Language and Literature 14 (1), pp. 5–24.

Stubbs, M. and I. Barth. 2003. 'Using recurrent phrases as text-type discriminators. A quantitative method and some findings', Functions of Language 10 (1), pp. 61–104.

Teubert, W. 1999. Online. 'Corpus linguistics – a partisan view'. http://tractor.bham.ac.uk/ijcl/teubert_cl.html (last accessed: January 2004).

Teubert, W. 2005. 'My version of corpus linguistics', International Journal of Corpus Linguistics 10 (1), pp. 1–13.

Tognini-Bonelli, E. 2001. Corpus Linguistics at Work. Amsterdam: John Benjamins.

Tribble, C. 2006. 'English for academic purposes. Building an account of expert and apprentice performances in literary criticism' in M. Scott and C. Tribble. Textual Patterns. Key Words and Corpus Analysis in Language Education, pp. 131–59. Amsterdam: John Benjamins.

Wynne, M. 2006. 'Stylistics: Corpus Approaches' in K. Brown (ed. in chief) The Encyclopaedia of Language and Linguistics, pp. 223–26. Oxford: Elsevier.

Appendix A

Texts in the Dickens corpus:

American Notes, *The Battle of Life*, *Barnaby Rudge*, *Bleak House*, *A Christmas Carol*, *The Chimes*, *The Cricket on the Heath*, *David Copperfield*, *Dombey and Son*, *Great Expectations*, *Hard Times*, *The Haunted Man*, *Little Dorrit*, *Martin Chuzzlewit*, *The Mystery of Edwin Drood*, *Nicholas Nickleby*, *The Old Curiosity Shop*, *Oliver Twist*, *Our Mutual Friend*, *The Pickwick Papers*, *Sketches by Boz*, *A Tale of Two Cities*, *The Uncommercial Traveller*

Texts in the 19C corpus:

Jane Austen: *Persuasion*, *Emma*, *Pride and Prejudice*
Mary Elizabeth Braddon: *Lady Audley's Secret*
Anne Brontë: *Agnes Grey*
Charlotte Brontë: *The Professor*, *Jane Eyre*
Emily Brontë: *Wuthering Heights*
Edward George Bulwer-Lytton: *The Last Days of Pompeii*
Wilkie Collins: *The Woman in White*, *Armadale*
Antonina or, *the Fall of Rome*
Benjamin Disraeli: *Vivian Grey*
Sir Arthur Conan Doyle: *The Hound of the Baskervilles*
George Eliot: *Daniel Deronda*, *Middlemarch*, *The Mill on the Floss*
Elizabeth Gaskell: *North and South*, *Mary Barton*, *Cranford*
Thomas Hardy: *Tess of the D'Urbervilles*, *The Return of the Native*, *Jude the Obscure*
Mary Shelley: *Frankenstein*
Robert Louis Stevenson: *The Strange Case of Dr Jekyll and Mr Hide*
Bram Stoker: *Dracula*
William Makepeace Thackeray: *Vanity Fair*
Anthony Trollope: *The Small House at Allington*
Oscar Wilde: *The Picture of Dorian Gray*