3 Starting with the Texts: Corpora, Clusters, and Lexical Bundles

INTRODUCTION

In corpus linguistics, corpora are usually very large collections of texts or text samples. The present study is different from more mainstream corpus work in the sense that it focuses on several texts by one author and also aims to pay attention to individual texts and text extracts. Beginning with a description of the textual basis for this study (Section 3.1), the chapter provides some initial examples of 'clusters' (Section 3.2). Clusters can be defined in terms of purely formal criteria, but it is their meanings and functions that are the central interest of the present study. Different approaches exist to related phenomena (Section 3.3), and it is specifically the concept of the 'lexical bundle' that needs to be related to and distinguished from the way in which clusters are used in the present study (Section 3.4). Section 3.4.1 will look in particular at the lexical bundles that can be found in fiction and Section 3.4.2 at the more text-specific functions of longer clusters.

3.1 TEXTS AND CORPORA

The main corpus for this study contains 23 texts by Charles Dickens. This corpus is referred to as DCorp and Table 3.1 lists its content. DCorp contains all of Dickens's 15 novels (in italics in the table); it also includes 7 stories and sketches ("The Battle of Life", "A Christmas Carol", "The Cricket on the Hearth", "The Chimes", "The Haunted Man and the Ghost's Bargain", "Sketches by Boz", "The Uncommercial Traveller") and one text that can be classified as nonfiction ("American Notes for General Circulation"). Table 3.1 also contains the abbreviations for Dickens's works that are used in this book. For the selection of texts I followed Hori (2004). When I began my initial research on Dickens's language, it appeared useful to have the option of a reference point of previous studies. The results of the present study will place the focus of the discussion on the novels, as will become clear from the findings outlined in Chapter 4. Two other collections of texts are used as reference corpora. A corpus of 19th century novels (from now on referred

Table 3.1 Texts in DCorp (novels in *italics*)

AN	American Notes for General Circulation
BH	*Bleak House*
BL	The Battle of Life
BR	*Barnaby Rudge*
CC	A Christmas Carol
CH	The Cricket on the Hearth
Chi	The Chimes
DC	*David Copperfield*
DS	*Dombey and Son*
ED	*The Mystery of Edwin Drood*
GE	*Great Expectations*
HT	*Hard Times*
HM	The Haunted Man and the Ghost's Bargain
LD	*Little Dorrit*
MC	*Martin Chuzzlewit*
NN	*Nicholas Nickleby*
OCS	*The Old Curiosity Shop*
OT	*Oliver Twist*
OMF	*Our Mutual Friend*
PP	*The Pickwick Papers*
SB	Sketches by Boz
TTC	*A Tale of Two Cities*
UT	The Uncommercial Traveller

to as '19C'), containing 29 novels by 18 authors (see Appendix 1), and a corpus referred to as '37N' that contains only one text per author. 37N partly overlaps with 19C and also includes one text by Charles Dickens, his novel *Bleak House* (see Appendix 2). When *Bleak House* is excluded from 37N, the corpus is referred to as 36N. DCorp and 19C comprise approximately 4.5 million words each, and 37N amounts to about 6 million words.

All texts come from Project Gutenberg. Although Project Gutenberg texts have been used by various researchers (e.g., Manning & Schütze 1999; Barnbrook 1996; Stubbs 2005; Römer 2006), the quality of the texts can be viewed critically. Berglund et al. (2004: 14) even go as far as to say: "the policy of creating non-specific editions aimed at the general reader makes the use of Project Gutenberg texts unreliable to the serious arts and humanities scholar". What makes the Project Gutenberg texts vulnerable to criticism is the fact that the eBooks are created by a system based on contributions from volunteers and it is possible that one eBook is the result of the collation of work by different people. The main advantage is that the books are freely available and come as plain text files. As Project Gutenberg allows its users to distribute the texts, there are also various other websites that make texts available on the basis of Project Gutenberg material or that have links with Project Gutenberg (see, e.g., eBooks@Adelaide).[1] With the proliferation of eBooks that can now be read on convenient hand-held eBook readers, the choice of texts for corpus stylistics will be ever increasing. In addition to issues of copyright, format of the data, and costs, questions about the quality of texts remain, and it would require a separate study to evaluate different sources, specifically in relation to scholarly editions.[2]

The approach that the present study takes to the texts in the corpora is a practical one. The texts will not be perfect, but are sufficient for the purpose of the study. Misprints or other mistakes in the texts will affect word counts, but as the approach is a combination of quantitative and qualitative analysis, where the focus lies on the qualitative functional interpretation, the effect is not detrimental. The theoretical claims of this book are not based on exact frequency counts. What is crucial is to have enough data as evidence for the existence of functions. We also have to keep in mind that any corpus will contain mistakes. In corpus linguistics, it is unlikely that anyone would consider proofreading a corpus containing several millions of words. The difference we see between corpus linguistic and corpus stylistic work is that we are dealing with texts by well-known authors. Such texts have a longer-lasting impact on a society and culture than most texts in a general corpus, and they have also received more scholarly attention. On the one hand, this is an advantage for corpus studies, as the body of literary criticism offers valuable information for the contextualisation of findings. On the other hand, the existence of different editions also leads to a variety of electronic texts where the source edition is not always straightforward to trace.

The texts for the major part of this study are used in the form of plain text files.[3] Chapter 9 will also refer to some alternative text formats. The main tool for the processing of the texts is the software *WordSmith Tools* version 5 (Scott 2008). Texts from DCorp will be referred to by their abbreviations, as shown in Table 3.1. When extended examples are quoted from DCorp, they will be referred to by book title and chapter number. In tables and figures, the texts from 19C will be referred to by the abbreviations specified in Appendix 1. In the discussion of examples from the reference corpora, author and title will be provided in full for ease of reference. No cross-checks were made with printed versions of the texts to provide page references.[4]

3.2 DEFINING AND RETRIEVING CLUSTERS—INITIAL EXAMPLES

The use of clusters to analyse literary texts illustrates how a corpus approach is at the same time less flexible and more systematic than a literary stylistic approach. By identifying clusters we can get a systematic overview of the frequency and distribution of repeated sequences of words across a text or corpus. At the same time, with clusters we only focus on a selection of words and meanings in the text/corpus. Depending on the length of clusters, this selection can be very small. What I am arguing in this book is that clusters can serve as starting points for further detailed analyses. They do not replace other forms of stylistic analysis but complement them. A 'cluster' is a repeated sequence of words (cf. also Scott 2004–6: 204). This is a straightforward concept providing an operational definition for automatic retrieval by a concordance package like *WordSmith Tools*. To retrieve the five-word clusters in *DC,* for instance, the computer begins with the first word in the text and identifies five-word sequences, checking each sequence against those that occurred previously to count the cluster tokens. Thus, the first sentence of *DC*

> Whether I shall turn out to be the hero of my own life, or whether that station will be held by anybody else, these pages must show.

yields the following sequences—taking into account only words up to the comma:

Whether I shall turn out
I shall turn out to
shall turn out to be
turn out to be the
out to be the hero
to be the hero of
be the hero of my
the hero of my own
hero of my own life

Whether the next sequence in this list is *of my own life or* or *or whether that station will* depends on the settings specifying whether punctuation is or is not to be included in the clusters. Further settings concern the cluster length and the minimum frequency with which a cluster has to occur to be considered in an analysis. If the clusters get longer, the number of sequences that occur repeatedly decreases. Table 3.2 shows the number of two-, three-, four-, and five-word clusters in *DC* that occur a minimum of 10 times. While there are

Table 3.2 Number of clusters in *DC* occurring at least 10 times

2-word clusters	3-word clusters	4-word clusters	5-word clusters
4441	1405	117	5

more than four thousand two-word clusters, there are only five five-word clusters occurring at least 10 times.

The settings and cut-off points for the analysis of clusters obviously depend on the corpus under analysis. The novel *DC* comprises about 350,000 words and has 64 chapters. In terms of the general corpora that are used in corpus linguistics these are tiny numbers and consequently, the number of clusters based on such a corpus is relatively small. Corpora can contain hundreds of millions of words and the resources provided by the World Wide Web have pushed the limits even further. Still, in this book I want to argue that with the focus on a very specific corpus and individual texts, the analysis of clusters can provide useful starting points for textual analysis. Table 3.3 shows the 20 most frequent two-, three- and four-word clusters in *DC*. The most frequent two-word clusters are combinations of grammatical words: *in the, of the, I was, that I*. The first lexical word occurs in cluster number 14 *my aunt* and then cluster number 20 is *said Mr.* These two-word clusters noticeably reflect features of the text. David Copperfield's aunt Betsey plays an important role in his life and the novel is a first-person narration, so the cluster *my aunt* comes up. The cluster *said mr* is a reflection of the text being a narrative. We find further clusters with names of characters such as the three-word clusters *said Mr Peggotty* and *said Mr Micawber,* and the four-word cluster *Mr and Mrs Micawber.* In addition to the rank (N) and the frequency of a cluster (F), Table 3.3 also shows the number of chapters (Ch) in which a cluster occurs. Of the two-word clusters, 11 occur in all of the chapters in *DC,* whereas of the three- and four-word clusters none occurs in all chapters. The four-word cluster *my dear Mr Copperfield* is limited to six chapters only. Its distribution relates to the fact that the cluster is always part of the speech of Mrs Micawber.

Table 3.2 shows that there are fewer longer clusters occurring a minimum of 10 times. Table 3.4 illustrates that these longer clusters are also more restricted to specific chapters. Table 3.4 further highlights how different clusters can overlap. All 13 instances of *the man with the wooden* overlap with the cluster *man with the wooden leg;* the 14 occurrences of the latter cluster are due to one occurrence of *that cruel man with the wooden leg*. The clusters *dear Copperfield said Mr Micawber* and *my dear Copperfield said Mr* are occurrences of the same six-word sequence *my dear Copperfield said Mr Micawber.*

These initial examples show that clusters for a single novel yield relatively small numbers of instances, but at the same time they point to specific features of the text. In a large general corpus of many millions of words, names and forms of address of specific characters in a novel would not show among the very frequent clusters and also might not be of interest to the linguist. Among the clusters for *DC,* we find examples that are frequent across texts, but also in the language more generally, e.g., the most frequent four-word cluster *I don't know what* in *DC* is a cluster that is also frequent in present-day conversation (cf. Biber et al. 1999: 1002). The way in which clusters can be interpreted and compared across texts depends on various factors. The next section looks at some general issues about the frequencies and functions of clusters.

Table 3.3 Top 20 of the two-, three-, and four-word clusters in *DC*

N	2-word	F	Ch	3-word	F	Ch	4-word	F	Ch
1	in the	1558	64	said my aunt	219	20	i don't know what	35	28
2	of the	1368	64	i could not	128	50	i have no doubt	33	25
3	i was	961	61	there was a	127	50	in the course of	33	24
4	that i	931	64	that i was	126	49	i could not help	32	20
5	i had	894	64	i don't know	125	47	i am sure i	31	22
6	to be	867	64	out of the	124	54	a good deal of	28	23
7	it was	864	62	said mr peggotty	110	12	as if he were	28	19
8	i am	823	63	it was a	109	47	as if it were	28	24
9	to the	755	64	i am sure	104	48	it would have been	26	19
10	and i	724	64	said mr micawber	103	11	as if i had	25	19
11	in a	715	62	if i had	98	44	for a little while	25	18
12	i have	712	64	that i had	91	39	for a long time	24	20
13	on the	691	63	i had been	86	42	in a state of	24	19
14	my aunt	671	40	that it was	85	44	if i had been	23	16
15	to me	623	64	as if he	79	43	my dear mr copperfield	23	6
16	with a	620	64	i am not	78	37	as well as i	22	16
17	of my	553	63	i have been	74	37	i beg your pardon	21	14
18	when i	553	63	a good deal	71	40	in the midst of	21	17
19	at the	541	64	i should have	71	40	mr and mrs micawber	21	10
20	said mr	518	48	my aunt and	67	24	as if he had	19	15

Table 3.4 Five-word clusters in *DC* occurring a minimum of 10 times

N	Cluster	F	Ch
1	man with the wooden leg	14	2
2	the man with the wooden	13	2
3	dear copperfield said mr micawber	10	6
4	my dear copperfield said mr	10	6
5	on the top of the	10	8

3.3 APPROACHES TO CLUSTERS AND FUNCTIONAL INTERPRETATIONS

The study of clusters has been approached in a number of ways and terminology varies for what in this book are called clusters. Another term for sequences of words is 'n-grams', with *n* specifying the number of words to appear together.[5] The term n-gram is not restricted to sequences of words, but can also refer to letters or phonemes. In computational linguistics, we come across n-grams, for instance, in the context of stochastic problems to predict the next word. N-gram models deal with probabilities of words occurring in sequences, working with the Markov assumption that only the words in the immediately preceding context affect the next word (cf. Manning & Schütze 1999: 192ff.). N-gram models are ideally trained on huge amounts of data to generate the best results. For the purpose of providing a textbook example, Manning and Schütze (1999: 195) use a corpus of Jane Austen's novels to build models and present the results. An example of a huge data set are the n-grams released by Google in 2006. These n-grams were retrieved from about one trillion word tokens of text from Web pages.[6]

In corpus linguistics, a term that has come to be used for sequences of words is the 'lexical bundle.' In their *Longman Grammar of Spoken and Written English,* Biber et al. (1999: 989) introduce the term 'lexical bundles' for "bundles of words that show a statistical tendency to co-occur". Relative frequencies are part of the definition of the concept, and quantitative criteria are key in identifying lexical bundles: "lexical bundles are the sequences of words that most commonly occur in a register" (Biber et al. 1999: 989; see also Biber 2006: 134, 173). Biber et al. (1999: 992) point out that three-word bundles are "a kind of extended collocational association, and they are thus extremely common", whereas "four-word, five-word, and six-word bundles are more phrasal in nature and accordingly less common". The operational definition that Biber et al. (1999: 992ff.) use requires that sequences of words must occur at least 10 times per million words in a register to qualify as lexical bundles. For the less frequent five- and six-word bundles a cut-off of at least 5 times per million words is used. In addition, lexical bundles must be spread across at least five different texts in the register "to exclude individual speaker/writer idiosyncrasies" (Biber et al. 1999: 993).

In their comparison of conversation and academic prose, Biber et al. (1999: 996) show, for instance, that lexical bundles consisting of a personal pronoun and a lexical verb phrase are common in conversation (e.g., *I don't know how, I don't think so, I would like to*), whereas in academic prose, patterns that include a preposition plus a noun phrase are more widely used (e.g., *as a result of, at the time of, in the form of*). The variation of lexical bundles across different groups of texts is also addressed, for instance, in Biber et al. (2004), Conrad and Biber (2005), and Biber (2006). Further studies investigating sequences of words in relation to specific groups or types of texts are Stubbs and Barth (2003), using the term 'chains' to refer to word sequences; Culpeper and Kytö (2002), focusing on data from 1560 to 1760; Oakey (2002), looking at academic writing; Mittmann (2004), comparing British and American spoken English; and De Cock (1998), investigating recurrent word combinations in native and nonnative speaker English.

In addition to the frequencies and distributions of clusters, functions that are associated with clusters play an important role. Altenberg (1998), for instance, looks at what he calls 'recurrent word-combination' in spoken English as represented by the London-Lund Corpus. He classifies the recurrent word-combinations according to a set of grammatical structures and interprets them functionally. He discusses, for instance, independent clauses in view of conventionalised speech acts and discourse strategies. Among his examples are a set of responses that include thanks, *thank you very much (indeed);* reassuring expressions that function as responses to apologies, *it's all right;* or 'agreements' supporting statements by the previous speaker, *(yes) that's right (yes)*. O'Keeffe et al. (2007) also concentrate on spoken English and use the term 'chunks' to refer to recurrent strings of words.

Unlike Altenberg (1998), O'Keeffe et al. (2007: 64) stress that they are interested in chunks primarily because of their "pragmatic integrity and meaningfulness regardless of their syntax". Similarly, Biber (2006: 174) highlights that lexical bundles generally do not represent structurally complete units, but serve as "discourse framing devices: they provide a kind of frame expressing stance, discourse organization, or referential status, associated with a slot for the expression of new information relative to that frame". Biber (2006) is particularly interested in the link between the distribution of lexical bundles and their discourse functions. Focusing on university registers, Biber (2006), building on Biber et al. (2004), outlines a taxonomy of discourse functions, with lexical bundles being characterised by three primary discourse functions: stance expressions (*I don't know if, I'm not going to, it is possible to*), discourse organisers (*want to talk about, if we look at, one of the things*), and referential expressions (*or something like that, have a lot of, the nature of the*). Biber (2006) shows, for instance, how classroom teaching combines characteristics of conversation and textbooks/academic prose.

Similar to Biber (2006), O'Keeffe et al. (2007: 71) interpret 'chunks' in terms of 'frames' that are "grammatically incomplete strings [. . .] to which new, unpredictable content can be attached". Their pragmatic categories include discourse marking (*and then, you know what I mean, at the end of the day*), face and politeness (*do you think, I was going to say*), and vagueness and approximation (*a couple of, and things like that*). O'Keeffe et al. (2007: 71, 75) emphasise "the all-pervasiveness of interactive meaning-making in everyday conversation and the degree to which speakers constantly engage with each other on the interactive plane". With their focus on conversation, O'Keeffe et al. (2007) are mainly interested in pragmatic categories characterising the speaker-listener world as opposed to the content-world (see also McCarthy & Carter 2004). So, a cluster like *in the middle of the* that falls into the category of Biber's (2006) referential bundles, is for O'Keeffe et al. (2007) potentially explainable through repeated events in the content-world. Another way of interpreting functions of lexical bundles is suggested by Bastrukmen and Nesi (2006), who are interested in the potential of lexical bundles to create cohesion.

Studies of lexical bundles, chunks, or formulaic sequences, as they are sometimes called, provide ample evidence for the relationship between the frequency and the pragmatic or discourse functions of the sequences under investigation. Repeated patterns occur repeatedly because they are associated with functions that are relevant to the communicative needs of a specific register. In addition, prefabricated phrases increase processing efficiency by being memorised as single units. Although formulaic choices occur both in speech and writing, their prominent place in spoken grammar can be linked to the online production pressure of speech. Evidence for the relationship between the repetition and frequencies of sequences of words and the processing effort they require is provided, for instance, by Conklin and Schmitt (2008), Bannard and Matthews (2008), and Arnon and Snider (2010). Wray (2002, 2008) points out that there is a link between the functions of formulaic expressions and their role in relieving processing pressure. Repeated patterns not only reduce effort for the speaker but also for the hearer. They can therefore be used to manipulate the hearer's actions, knowledge, and belief in order to promote the self. Formulaic expressions in greetings, forms of address, and linguistic markers of social identify exemplify means by which the speaker can influence the hearer's perception of them (Wray 2008: 69).

The frequency of lexical bundles or chunks distinguishes them from idioms. Because of their low frequencies, most idioms cannot be considered as lexical bundles (Biber 2006: 134). Some idioms and fixed formulas are occasionally used in fiction, e.g., *what on earth, how do you do, a piece of cake*. Such expressions are not frequent in conversation, but are more commonly used in fictional dialogue (Biber et al. 1999: 1025). The fact that most lexical bundles are not idiomatic can be seen as one of the reasons why repeated sequences of words only started to receive attention through corpus linguistic work. The focus of previous research on structurally complete units may also have contributed to lexical bundles being overlooked (cf. Biber 2006: 135).

Both quantitative and qualitative information contributes to the characterisation of clusters across registers and within a specific register. This also applies to fiction. However, the approach of a study can be shaped by different aims. Hoover (2002) is an example of a study that shows the advantage of including frequent word sequences (more specifically, sequences of two words) in multivariate analyses for the characterisation of authorial style and authorship attribution. For Hoover's purposes, three-word clusters are not frequent enough and do not offer enough information (cf. Hoover 2002: 162). Starcke (2006), in contrast, focuses on three-word clusters in one novel only, where the clusters do provide useful information.

To conclude this section, it remains to point out that clusters as uninterrupted sequences of words only capture a specific part of the phraseology of English. More flexible patterns have been studied, for instance, by Renouf and Sinclair (1991), Biber (2009), and Cheng et al. (2006). There are also corpus tools that specifically address the variability of word sequences and offer various options to identify, for instance, 'phrase-frames' (cf. Fletcher's Phrases in English site[7]) or 'concgrams' (cf. Greaves 2009, Scott 2008). While the focus of this study is on clusters, in the qualitative analyses, examples of more flexible patterns will also be discussed.

3.4 LEXICAL BUNDLES OR CLUSTERS FOR THE STUDY OF DICKENS'S FICTION?

In the present approach, the questions that are addressed by using clusters for stylistic analysis are somewhat different from the questions that guide the study of lexical bundles. On the one hand, the hypothesis underlying the present book is that patterns identified with the help of corpus methods can contribute new categories to the inventory of descriptive tools for literary stylistics. So this would mean identifying patterns that are shared across a number of texts, as are lexical bundles. On the other hand, the present approach also stresses that the analysis of individual texts and text extracts has an important place in stylistic analysis. I am therefore interested in local textual functions both at the level of register as well as at the level of text. From this point of view the lexical bundle approach is clearly different. Biber (2006) points out that there are sequences of words that occur repeatedly in a single text, but such sequences often do not represent lexical bundles. Lexical bundles have to be widely distributed across texts, whereas "[t]hese local repetitions reflect topical concerns of the discourse. In contrast, lexical bundles can be regarded as the more general building blocks that are used frequently by many different speakers/writers within a register" (Biber 2006: 174, endnote 4). The pervasive nature of lexical bundles is also reflected in Carter and McCarthy's (2006: 15) account of word clusters, i.e., "groups of words that often occur together" and that "can operate as the frequent and regular building blocks in the construction of meaning". The criteria that are used in this book to identify clusters and investigate local textual functions differ from the definition of lexical bundles in several respects. To show what the differences are and how they impact on a functional interpretation, Section 3.4.1 takes a look at lexical bundles in fiction. Section 3.4.2 concentrates on longer, more text-specific clusters.

3.4.1 Lexical bundles in fiction

Lexical bundles are defined by their length, their frequency cut-off, the minimum number of texts they have to occur in, and the requirement that they

Table 3.5 Four-word lexical bundles in 37N

N	Lexical bundle	Freq.	Texts	% of Texts
1	*for the first time*	385	37	100.00
2	*at the same time*	306	35	94.59
3	*at the end of*	282	36	97.30
4	*in the midst of*	282	34	91.89
5	*i should like to*	275	32	86.49
6	for the sake of	260	33	89.19
7	*the end of the*	245	34	91.89
8	the rest of the	240	34	91.89
9	in the course of	234	32	86.49
10	(it would have been)	227	35	94.59
11	i do not know	205	31	83.78

do not cross turns or punctuation. As lexical bundles aim to capture linguistic features that are shared by different authors, in this section I use the corpus 37N, i.e., a corpus with 37 novels by 37 authors including Dickens (see Appendix 2). The focus is first on four-word bundles with a frequency cut-off of 40 per million words, following Biber (2006). The bundles are retrieved with *WordSmith Tools*, specifying in the settings that bundles 'stop at punctuation'. There are about 6 million words in 37N, so the cut-off of 40 per million translates into a cut-off of 243 for four-word bundles. With these settings there are seven different bundles, which are shown in Table 3.5. The table presents the clusters in frequency order, and I will return to clusters numbers 8 to 11 further on. Table 3.5 also provides information on the number of texts in which a lexical bundle occurs. All seven lexical bundles meet the criterion of distribution across at least five texts, the percentage of texts in which each occurs ranging from about 86 to 100.

In his comparison of lexical bundles across registers, Biber (2006) shows that the number of lexical bundles in different registers varies between about 20 and 80. Academic prose has the lowest number of different lexical bundles, conversation has about twice as many as academic prose, and classroom teaching in turn has about twice as many as conversation. The number for textbooks is in between academic prose and conversation. Comparing Biber's (2006) figures to the results for 37N, we find that fiction has comparatively fewer different lexical bundles than any of the four registers in Biber's (2006) study. One of the factors that can have an effect on the number of bundles is the corpus design and sample size. Referring to work by Cortes (2002), Biber (2006: 175 note 8) points out that analyses of smaller corpora can result in more lexical bundles, because of inflated frequencies.

Table 3.6 Four-word lexical bundles in 37Ns30

N	Lexical bundle	Freq.	Texts	% of Texts
1	*for the first time*	84	31	83.78
2	*at the same time*	69	27	72.97
3	the rest of the	55	21	56.76
4	*i should like to*	52	23	62.16
5	*at the end of*	49	23	62.16
6	in the course of	49	19	51.35
7	*the end of the*	47	18	48.65
8	i do not know	45	14	37.84
9	*in the midst of*	43	21	56.76
10	for the sake of	41	23	62.16

The texts in 37N are complete texts and therefore relatively long, i.e., between 16,000 and 350,000 words. Table 3.6 presents the lexical bundles in a subcorpus of 37N, referred to as 37Ns30, that only contains samples of the texts up to about 30,000 words.[8] The size of this sample corpus amounts to nearly 1.06 million words. To work with the same cut-off of 40 per million, a lexical bundle has to occur at least 42 times to be included in the analysis. As Table 3.6 shows, the smaller corpus provides two more clusters than the larger one. There is also greater variation in the distribution across texts. In both corpora, the most frequent cluster is *for the first time*. Whereas this cluster occurs in all texts of 37N, it is only found in 31 of the text samples in the subcorpus. None of the lexical bundles in the smaller corpus occur in all text samples, and one of the bundles, *I do not know,* is found in only about 38 per cent of the text samples. The lexical bundles that are italicised in Tables 3.5 and 3.6 are shared by both lists, i.e., the two corpora have six lexical bundles in common. With about 1.06 million words the subcorpus on which Table 3.6 is based is more similar in size to Biber's (2006) corpora of conversation and textbooks, than the full 37N is (cf. Biber 2006: 24, 175, note 8). Still, the figures suggest that there are fewer lexical bundles in fiction than in any of the registers discussed in Biber (2006).

Biber (2006: 134) stresses that the cut-off of 40 per million is somewhat arbitrary and a conservative choice. If we go further down the list of lexical bundles in 37Ns30 and include the next one (*for the sake of*) in the discussion, all 7 of the lexical bundles from 37N are among the top 10 lexical bundles in 37Ns30. Equally, extending the list of 37N by 4 lexical bundles (the lexical bundles in Table 3.5 appearing below the line), yields 10 lexical bundles that are shared by both corpora. The lexical bundle in brackets (*it would have been*) is the only one not shared. From Tables 3.5 and 3.6 we can thus draw the following conclusions. Fiction seems to be a register that has fewer lexical bundles than some other registers, here particularly the university registers investigated by Biber (2006). The number of different lexical bundles per million words decreases slightly when instead of text samples full texts are used and the size of the corpus is thus increased (in the present example by about 6 times). However, in the larger corpus the lexical bundles are distributed across more different texts. These figures suggest that for fictional texts (more specifically novels), there is a set of lexical bundles that is likely to occur at some point throughout the text.

For his taxonomy of discourse functions, Biber (2006: 146) stresses that the categories aim to include functions "that can potentially be realised in any register". Discourse functions that are most prominent from Tables 3.5 and 3.6 belong to the category that Biber (2006: 139) calls 'referential bundles': "Referential bundles make direct reference to physical or abstract entities, or to the textual context itself, either to identify the entity or to single out some particular attribute of the entity as especially important". For referential bundles further subcategories can be distinguished. The following are examples of Biber's category of 'time/place/text-deixis' bundles:

(1) One day near *the end of the* long vacation, when he had been making a tour in [. . .]

(George Eliot,
Daniel Deronda)

(2) The light appeared *at the end of* the street leading from the more central [. . .]

(Wilkie Collins,
Antonina or, the Fall of Rome)

(3) They landed some thirty or forty yards lower, *in the midst of* our kitchen-garden, [. . .]

(R. D. Blackmore,
Lorna Doone, a Romance of Exmoor)

(4) [. . .] and there might be other things said *in the course of* the conversation which history has not condescended to record.

(John Henry Newman,
Loss and Gain: The Story of a Convert)

Example (1) illustrates time reference, examples (2) and (3) reference to places, and example (4) is a form of text reference, taken from an extract where the narrator addresses the reader and comments on a conversation in the preceding text. Examples (1) and (2) also show how shorter lexical bundles can combine into longer bundles. The four-word bundle *the end of the* is part of the five-word bundle *at the end of the*. Altogether, 110 of the

245 cases of *the end of the* occur as part of the five-word bundle *at the end of the*. The lexical bundles *for the first time, at the same time,* and *the rest of the* are further examples of bundles with referential functions. The lexical bundle *the rest of the* illustrates functions of a subcategory that Biber (2006: 145) calls 'bundles specifying attributes'. Examples of this group are bundles that specify quantities or amounts as in example (5), as well as bundles that specify more abstract characteristics as in example (6):

(5) Sebastian Dixon's debts were to be paid off; £1000 was left to Marianne Dixon, and *the rest of the* personal property was to be Amabel's.

(Charlotte M. Yonge,
The Heir of Redclyffe)

(6) But such a decision was *for the sake of* the offspring, and of doubtful justice.

(Charles Reade,
The Cloister and the Hearth)

As Biber (2006: 139) points out, a single lexical bundle can have different functions depending on the context. Examples (1) and (2) show how the same bundle can refer to a time period or a place. Similarly, the bundle *the rest of the* cannot only specify quantities as in example (5) but also refer to time, as in *the rest of the morning,* or to place, as in *the rest of the gardens*. The contextual dependency of the bundles is visible to some extent in their structural incompleteness. Lexical bundles are 'discourse building blocks' (Biber 2006: 156) or structural 'frames' (Biber 2006: 172) to which new information can be attached.

Another category of functions found among the lexical bundles of the fiction corpora used for this study are 'stance expressions'. According to Biber (2006: 139) stance bundles "express attitudes or assessments of certainty that frame some other proposition". The lexical bundles *I should like to* and *I do not know* are examples of this category:

(7) "*I should like to* come to the theatre with you, Lord Henry," said the lad.

(Oscar Wilde,
The Picture of Dorian Gray)

(8) *I do not know* how far matters may have gone between them; but [. . .]

(Harriet Martineau, *Deerbrook*)

Examples (7) and (8) are personal stance expressions that are attributed to the speaker/writer through the pronoun *I*. Both examples are taken from direct speech of characters in the story, with example (8) coming from a longer extract so that the quotation marks are not visible in the example. These examples underline that in fiction we encounter several textual levels with the interaction between characters being an important one among them. Stance bundles link in with the pragmatic categories outlined by O'Keeffe et al. (2007: 71) for the creation of speaker meaning, i.e., the 'speaker-listener world' as opposed to the 'content-world.'

Overall, the lexical bundles of the present section highlight two aspects of textual worlds in fiction. There is a content-world, which is reflected by the broad group of referential bundles, and there is a form of a speaker-listener or interactive world associated with stance bundles or chunks with pragmatic functions, as O'Keeffe et al. (2007) refer to them. Lexical bundles as frequent sequences of words have important discourse functions as textual building blocks in fiction. However, the frequency criteria in the definition of lexical bundles emphasise features that are widely shared between texts. As a consequence, the number of lexical bundles is relatively low. A potential explanation for this may be the variety of topics and themes that can be explored in fiction, but also the fact that fictional prose operates on several discourse levels (see, e.g., Short 1996: 257). The lexical bundles are counted irrespective of whether they occur in the speech of characters in the narration. It is also important to stress that the observations in the present chapter are based on 19th century fiction. So this is the register to which the shared features apply. Twentieth century short stories, for instance, may turn out to show different sets of lexical bundles.

3.4.2 Increasing the length of clusters

As we have seen in the example of *DC* in Section 3.2, there are usually fewer clusters above a specific frequency cut-off when the length of clusters increases (cf. Biber et al. 1999: 993; Carter and McCarthy 2006: 831). The cut-off for lexical bundles thus has to be adjusted when sequences of more than four words are considered. Biber (2006) concentrates on four-word lexical bundles, but Biber et al. (1999) also investigate longer lexical bundles. Biber et al. (1999: 993) use a cut-off of at least 5 per million words for five- and six-word bundles, but they also use a lower cut-off of 10 per million words for four-word bundles (Biber et al. 1999: 992). Thus, to stay with the more conservative cut-off of 40 of the previous section, for five-word lexical bundles a cut-off of 20 per million words is set here.[9] For the corpora 37N and 37Ns30 this results in a cut-off of 122 and 21, respectively. With these settings, no lexical bundles are found for 37N, and for 37Ns30 there are the following two: *in the middle of the* and *a quarter of an hour.*

Instead of beginning with a frequency cut-off, Table 3.7 presents the top 10 most frequent clusters in 37N and its subcorpus 37Ns30. The two lexical bundles are highlighted in bold and the clusters that are shared between the two lists are italicised: 8 of the 10 clusters are among the top 10 in both lists.

Table 3.7 Top 10 five-word clusters in 37N and 37Ns30

37N					37Ns30			
N	**Cluster**	**Freq.**	**Per million**	**Texts**	**Cluster**	**Freq.**	**Per million**	**Texts**
1	*in the middle of the*	116	19.06	28	***in the middle of the***	26	24.58	17
2	*at the end of the*	110	18.08	30	***a quarter of an hour***	22	20.8	15
3	*a quarter of an hour*	101	16.6	29	at the top of the	19	17.96	13
4	*the other side of the*	98	16.11	24	*at the end of the*	17	16.07	9
5	*in the course of the*	91	14.96	27	*in the direction of the*	17	16.07	7
6	*on the other side of*	80	13.15	26	*in the course of the*	16	15.13	11
7	*i should like to know*	69	11.34	22	for the first time in	14	13.24	9
8	as if it had been	65	10.68	24	*i should like to know*	14	13.24	10
9	*in the direction of the*	64	10.52	18	*on the other side of*	14	13.24	9
10	as if she had been	62	10.19	16	*the other side of the*	14	13.24	9

Similar to the four-word lexical bundles from the previous section, we mainly find clusters that can be grouped into the functional category of referential bundles, as well as the pragmatic expression *I should like to know.* In the list for 37N there is an additional type of cluster that was not found among the four-word lexical bundles. There are two clusters that begin with *as if* and contain only function words: *as if it had been* and *as if she had been.* These clusters are examples of a cluster category with textual functions specific to fiction, and we will return to these clusters in more detail in the following chapter and specifically in Chapter 7. In relation to Tables 3.5 and 3.6, Table 3.7 also shows that the most frequent five-word clusters are distributed across fewer different texts than the most frequent four-word clusters. The range of texts for the five-word clusters in 37N is 16 to 30, out of the 37 texts in the corpus. In contrast, for four-word clusters in 37N, the number of texts ranges from 32 to 37 for the 10 most frequent clusters. So, with increasing cluster length, both the number of clusters and the number of texts in which they occur decreases.

Figure 3.1 compares the number of clusters in DCorp with the number of clusters in 19C. DCorp and 19C are similar in size with both containing about 4.5 million words. In the following, I do not use normalised but absolute numbers, because of the similar sizes of the corpora. All clusters have to occur at least 5 times to be included in the comparison and clusters now do not stop at punctuation. Figure 3.1 shows that there are fewer clusters, as the cluster length increases. The numbers of eight-word clusters are so low that they are difficult to recognise in the figure: in DCorp we find 51 and in 19C 30 eight-word clusters. For each set of clusters, the number in DCorp is higher than the number in 19C (and 19C has about 25,000 more words than DCorp). To some extent, the higher numbers for DCorp have to be seen in relation to the fact that DCorp contains texts by just one author.

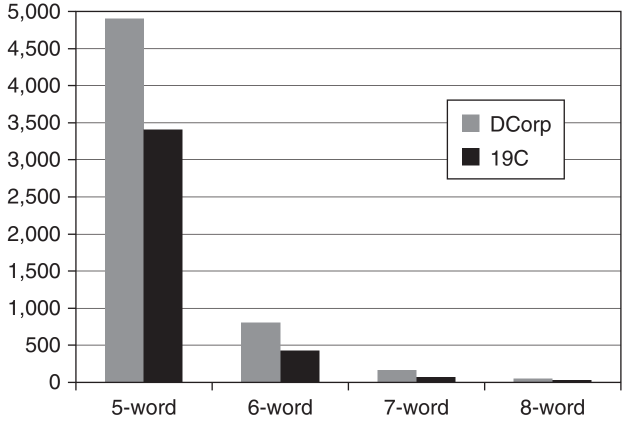

Figure 3.1 Five- to eight-word clusters in DCorp and 19C

With the version of *WordSmith Tools* (Scott 2008) that was used for the present study the maximum length of clusters that could be retrieved was 12. Table 3.8 presents the results for clusters from length 9 to 12. In contrast to Figure 3.1, the picture is now reversed and we find a higher number of clusters in 19C. An important aspect of the development of the cluster figures is that at some point all clusters form patterns that are specific to particular texts. This point is reached with the nine-word clusters in 19C and with the ten-word clusters in DCorp. In 19C there are 25 nine-word clusters, each cluster occurs in only one text, and altogether there are only two different texts from which the set of clusters comes. One of the nine-word clusters in 19C is *a fair day's wage for a fair day's work* occurring 5 times in Disraeli's *Sybil*. The other 24 nine-word clusters are from *Vivian Grey*. They are all clusters that are parts of larger clusters, building up to two sequences. One is an 18-word cluster occurring 5 times (here given with capitalisation and punctuation to be easier to read): *I feel the greatest pleasure in seeing you. My greatest pleasure is to be surrounded by my friends*. The other is a 22-word cluster occurring 6 times: *Again, Sir Stranger, briefly, but heartily, welcome! Welcome from us, and welcome from all; and first from us, and now from the*. All of the ten-, eleven-, and twelve-word clusters in Table 3.8 are then also made up by parts of these very long clusters. These very long clusters illustrate how including or excluding punctuation affects the kind of clusters that are obtained. If

Table 3.8 Nine- to twelve-word clusters in DCorp and 19C

	9-word	10-word	11-word	12-word
DCorp	24	13	7	3
19C	25	22	20	18

Table 3.9 Ten-word clusters in DCorp

10-word cluster	Freq.	Text
improved hot muffin and crumpet baking and punctual delivery company	6	NN
metropolitan improved hot muffin and crumpet baking and punctual delivery	6	
united metropolitan improved hot muffin and crumpet baking and punctual	6	
the united metropolitan improved hot muffin and crumpet baking and	5	
at the delightful village of dotheboys near greta bridge in	5	NN
the delightful village of dotheboys near greta bridge in yorkshire	5	
but i never own to it before her discipline must	5	BH
i never own to it before her discipline must be	5	
never own to it before her discipline must be maintained	5	
monotony of bells and wheels and horses' feet and no	5	DS
of bells and wheels and horses' feet and no rest	5	
let him remember it in that room years to come	7	DS
no disparity in marriage like unsuitability of mind and purpose	5	DC

In this table I added the apostrophe for *horses*' as this is not included in the generation of clusters.

clusters were to stop at punctuation, the sequence of 22 words would only result in clusters of a maximum length of four words.

In DCorp, the 13 ten-word clusters are the first set where all clusters are part of text-specific patterns. Similar to 19C, we find clusters that form even longer sequences. However, the 13 clusters still originate from four different texts. Table 3.9 shows the 13 ten-word clusters in DCorp grouped according to related clusters and to the texts they come from. From Table 3.9 we can already see how the eleven- and twelve-word clusters will develop. For the eleven-word clusters we still have three different texts (*NN, BH,* and *DS*) and two for the twelve-word clusters (*NN* and *BH*). In fact, the longest cluster that we find in DCorp occurs in *NN* and is the following 13-word cluster: *the united metropolitan improved hot muffin and crumpet baking and punctual delivery company.* So, whereas we find one text in 19C that has longer clusters than any of the texts in DCorp, there are still several texts in DCorp where Dickens makes use of relatively long clusters for the creation of special effects. This strategy can be seen as one of the reasons why we find more clusters in DCorp than in 19C. The longer clusters also add to the figures for shorter clusters, because they are built up by shorter ones.

In addition to the strategy of using long clusters, the higher number of clusters in DCorp as compared to 19C can also result from clusters that are more specific to Dickens and used in several texts in DCorp. The eight-word cluster *with the air of a man who had,* for instance, occurs 15 times and in nine texts in DCorp. However, there is no text in DCorp in which the cluster occurs more than 3 times. In 19C the cluster is not found at all (it also does not occur in 37N). Such points have to be taken into consideration to avoid overinterpretations.

Overall, the comparison of figures for DCorp and 19C suggests the following picture. In DCorp we find clusters whose repetitions are spread across a number of different texts, including clusters that reach the minimum occurrence of five, because DCorp contains five different texts in which they occur. Such clusters point to stylistic characteristics of Dickens. Their frequencies and functions would require more detailed comparison with 19C and other reference data to assess their significance. At the same time, we find clusters that are not repeated across texts, but are specific to individual texts. Dickens did not necessarily repeat the same words, but the same strategy of repeating longer sequences in the same text. So the figures point to at least two aspects of authorial habit: the repetition of the same clusters in different texts and the use of repetition within individual texts.

As outlined previously, frequently recurring sequences tend to have common discourse functions and are assumed to relieve processing pressure. Longer, text-specific repetitions, as illustrated in this section, form local and more creative patterns. From a psycholinguistic point of view it seems worth investigating what the creation of such patterns means in terms of processing effort. To what extent is there a relationship between the number of repetitions required for a sequence to become a pattern and the number of words in the repeated sequence? Wray (2008: 69) points out that for memorising formulaic sequences rhythm and rhyme are more important than the meaning of the unit. Thus, the kind of words in the sequence might also affect the formation of the pattern. While psycholinguistic issues are beyond the scope of the present study, they underpin the rationale for setting the cluster length at five in the following chapters. It is generally assumed that working memory can hold seven plus-or-minus two items. When the items are words, the number is more likely to be five than nine (see also Reisberg 2001: 140f., Matlin 2003: 80). So, to capture local patterns that can be remembered, five seems to be a good starting point. For the more general patterns, five also is a reasonable length based on the findings for lexical bundles.

CONCLUSIONS

This chapter has highlighted differences between a corpus linguistic and a corpus stylistic focus. Corpus linguistics aims to find generalisations that hold across a range of different texts, while corpus stylistics is also interested in features of individual texts. The discussion of lexical bundles has shown that frequency cut-offs applicable to large corpora can be too restrictive if applied to specific literary text collections. As a consequence, descriptive tools such as lexical bundles are limited in their applicability to specific aspects of the study of literature. On the other hand, literary texts can contain features that are not typically of interest in corpus linguistic studies. The 'long clusters' that we find in several of Dickens's texts are a case in point. The way in which clusters are used in the present study is thus specific to the aim of identifying local textual functions in Dickens's fiction. When clusters increase in length they become more text-specific; at the same time, the number of different clusters decreases. In the following, the focus will be on clusters of length five. They appear to be sufficiently long to find text-specific features and at the same time provide sufficient data to have a critical mass for the identification of functions shared across texts. Still, the focus on five-word clusters is to some extent arbitrary and exploratory. What this chapter has also shown is that settings for the retrieval of clusters can vary with the aims of the study. While lexical bundles do not cross punctuation, the clusters in the present study do not stop at punctuation. In this way clusters are included that are made up of parts from both a character's speech and the narration. Overall, the figures that are dealt with in the present study are small by the standards of large-scale corpus studies. A corpus stylistic study has to find a useful trade-off between general quantitative information and finding ways of selecting examples that can serve as a basis for more detailed textual analyses. The next chapter will look at this trade-off.