Article

Language and Literature
2019, Vol. 28(4) 326–353

DOI: 10.1177/0963947019886754
journals.sagepub.com/home/lal

Speech-bundles in the 19th-century English novel

Michaela Mahlberg
Viola Wiegand
University of Birmingham, UK

Peter Stockwell
University of Nottingham, UK

Anthony Hennessey
University of Birmingham, UK

Abstract

We propose a lexico-grammatical approach to speech in fiction based on the centrality of 'fictional speech-bundles' as the key element of fictional talk. To identify fictional speech-bundles, we use three corpora of 19th-century fiction that are available through the corpus stylistic web application CLiC (Corpus Linguistics in Context). We focus on the 'quotes' subsets of the corpora, i.e. text within quotation marks, which is mostly equivalent to direct speech. These quotes subsets are compared across the fiction corpora and with the spoken component of the British National Corpus 1994. The comparisons illustrate how fictional speech-bundles can be described on a continuum from lexical bundles in real spoken language to repeated sequences of words that are specific to individual fictional characters. Typical functions of fictional speech-bundles are the description of interactions and interpersonal relationships of fictional characters. While our approach crucially depends on an innovative corpus linguistic methodology, it also draws on theoretical insights into spoken grammar and characterisation in fiction in order to question traditional notions of realism and authenticity in fictional speech.

Keywords

CLiC, characterisation, corpus linguistics, lexical bundles, 19th-century fiction, realism, spoken grammar

Corresponding author:
Michaela Mahlberg, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK.
Email: m.a.mahlberg@bham.ac.uk

1. Introduction

In fictional narrative, the centrality of character has always been recognised, and the importance of the speech of characters within narratives has received a great deal of critical attention. Indeed, it is possible in the most general terms to discern a broad shift in narratology from a definition of narrative mainly as the sequencing of events (in which characters are caught up) to a definition that narrative is primarily the representation of minds negotiating the events that they perceive and articulate (most obviously, contrast Labov, 1972, and Genette, 1980, with Herman, 2013, and Zunshine, 2006, for indicative examples). The stylistic representation of forms of dialogue in fictional narrative has been an abiding focus of interest, as part of a literary-linguistic account of characterisation, narratorial voice and authorial composition (from, for example, Cohn, 1978, and Page, 1988, to Leech and Short, 1981/2007, Semino and Short, 2004, and Busse, 2010).

In cognitive stylistics, characterisation has been re-described systematically as a process in the mind of the reader where schematic background knowledge and information triggered by cues in the text interact (Culpeper, 2001, 2009). A significant part of this background knowledge are the norms of language. The continuity between fictional worlds and the real world receives even greater emphasis through the concept of mind-modelling: this describes the way in which readers create a sense of the minds of fictional characters. It refers to the human capacity for imagining and maintaining a working model of the knowledge, beliefs, feelings, motivation and consequent behaviour of others (Stockwell, 2009). This human capacity works in similar ways whether minds of real or fictional people are modelled. The textual information that readers draw on in the process includes descriptions of physical appearance, the speech of characters, social relationships and reactions of other characters. Mind-modelling is a cognitive poetic concept, which can be extended through a corpus linguistic dimension (Mahlberg and Stockwell, 2016; Stockwell and Mahlberg, 2015): corpus linguistic methods are employed to find the linguistic patterns that contribute to the cumulative picture of character information in a text.

Ultimately, the language of fictional texts is part of language in general – so language forms a clear link between fiction and the real world. However, the continuity that mind-modelling proposes between real and fictional people has so far been described in terms of the process of modelling and the information used for it, that is, based on information drawn from the text. The fluid linguistic relationship between text and previous experience has not been studied. While information about the physical appearance of characters is less directly comparable, patterns of fictional speech and real spoken language can indeed be compared.

We propose a novel approach to describing the relationship between real and fictional speech along a continuum. This approach enables a more nuanced account of the role of speech in the creation of fictional characters. At the same time, it also utilises innovative corpus methodology to be able to arrive at these theoretical insights. We focus on speech in the novels of Charles Dickens and the 19th century, partly because of Dickens' influential and canonical status in the 19th-century literary tradition, partly because he is celebrated for his characterisation and verisimilitude and partly because of the practical consideration that, in the 19th-century novel, direct speech is a common form of speech presentation and relatively straightforwardly identifiable.

2. Direct speech itself

In general, the indirect and narratorially mediated forms of speech presentation have received more scholarly attention than straightforward examples of direct speech within narrative fiction. Narrative voice, stream-of-consciousness techniques and free indirect discourse perhaps offer more interesting tensions to explore than what appears to be the simple reportage of a character's words, represented plainly with a framing reporting clause. This might be explained by the prevalence of what Palmer (2011: 211) calls an 'internalist' perspective on the mind: a focus on 'those aspects that are inner, introspective, private, solitary, individual, psychological, mysterious and detached'. As a corrective, he puts forward the notion of the 'social mind' to emphasise the need for an external perspective to minds that recognises the physical and social contexts of the fictional characters' experience. Like Palmer (2011), we are interested in an externalist perspective on the minds of fictional characters. Our approach differs from his, as we focus on direct speech as a central feature of the social context of fictional characters.

Our approach is also distinguished from most previous literary-linguistic treatments in that we aim to explore squarely the *content* of the speech, rather than simply its framing presentation. Indeed, we can characterise most previous discussions of fictional dialogue as being interested in the peripheral framing of the speech or in only those stylistic aspects of the speech itself which pertain to the narrative frame. For example, the presence or absence of the reporting clause, the nature of narratorial description, the stylistic extent to which the character's speech is assimilated syntactically with the narrator and even the deictic style in which the speech is altered to align with the narrator's mind (as typically discussed in work in the tradition of Leech and Short, 1981/2007) are all aspects of speech representation that are mainly interested in narrator-character agency, rather than speech itself.

Perhaps another reason why direct speech has not been so thoroughly explored as indirect narratorial forms is that, as the psychological norm (Leech and Short, 1981/2007), it is interactively the least deviant. That is, the phenomenon by which someone tells you what someone else said is the most direct and natural form of relaying dialogue in everyday discourse. Since we cannot access people's thoughts directly, their actual speech is the primary and most direct form of access that we have. (This is why, contrastively, indirect thought is the psychological norm for internal mental presentation).

Comparisons of the content of direct speech in fiction and in natural conversation have tended not to feature in literary-linguistic discussions, except where the concern is a sociolinguistic one (see, for example, Hodson, 2014). The verisimilitude or not of fictional direct speech is set aside with the caveat of poetic licence. In the opening chapter of his classic *Speech in the English Novel*, Page (1988) describes the nature of fictional speech not as 'a slavish reproduction of actual features of speech' (p. 11); rather, he puts the emphasis on the creation of effects of spoken language. The focus on effects, that is, the impressions that linguistic forms create on the reader, has significantly influenced how literary stylistics has dealt with fictional speech. No systematic and large-scale attempt has yet been made to relate real spoken language and direct speech in fiction. The compelling account that Page (1988) makes for the 'inevitable gap' between real speech and dialogue in fiction seems to have been prevalent in literary stylistics. Page (1988: 7–10) identifies three main reasons for this gap:

1. Characteristics of spoken language, such as silence fillers, incomplete words or phrases, and grammatical inconsistencies, would be "unacceptable in the *written* medium of the novel" where editing and revising is possible, and they would be regarded as thematically meaningful in fiction. [italics in original]
2. Spoken language depends on the context of situation to a degree that can only be partially recreated in the novel.
3. The "phonological component" of spoken dialogue cannot be adequately represented by the written medium.

On the basis of these fundamental differences, the question is not how alike real and fictional speech are (they are evidently radically different) but how fictional speech creates a sense of realism and authenticity (see also McIntyre, 2016, on credibility). As Chapman (1994) points out, in the Victorian period 'the demand for realism in fiction was increasing' (p. 247), which did not mean that readers expected more transcription-like representations of speech, but rather conventions that would create a fiction of realism appropriate for the time. In this sense, the representation of regional or social differences through language is as relevant as the use of real-life settings. For fictional speech, the relation to theatrical forms is also important. A speech by Dickens is often quoted to make this point: 'every writer of fiction, though he may not adopt the dramatic form, writes in effect for the stage' (Fielding, 1960: 262).

The concept of speech realism or authenticity has manifested itself in narratology and literary-linguistics as the notion of 'faithfulness'. The stylistic model developed by Leech and Short (1981/2007) and colleagues at Lancaster University has a strong commitment to the 'principle of faithfulness': the notion that the form of speech presentation can be evaluated in terms of its degree of correspondence with the actual words uttered. So direct speech ('*I am Heathcliff!', he said*) presents what appears to be a verbatim report, whereas indirect speech (*he said that he was Heathcliff*) is an apparently less faithful representation of the original utterance. The distinction between report and representation is important in the Lancaster model and is significant when it comes to critical discourse analyses of newspaper reports and political, legal and commercial texts. Some researchers (Fludernik, 1993; Sternberg, 1982; Tannen, 1989) have pointed out that faithfulness is problematic in fictional literary texts, where there is of course no recoverable original utterance: the representation *is itself* the original. The Lancaster model resolves this by pointing to the pretence of faithfulness:

> [i]n fiction the 'original' speech has no independent existence whatsoever, and is only accessible *via* the report itself [. . .] the pretence of faithful reproduction is one of the consequences of the suspension of disbelief that applies to the reading of fiction. (Short et al., 2001: 499)

Since fictionality itself is essentially predicated on inauthentic pretence, and faithfulness is a matter of authenticity, this resolution is not a simple one. The Lancaster model manages the paradox because it is primarily concerned with the graphological and syntactic formal aspects of speech presentation framing, rather than with the semantic content of the speech itself. Direct speech, for example, is identifiable by its punctuation and reporting clause, and indirect forms can be differentiated by their different clausal

structure and backshifting of tense, aspect and deixis. This formal stylistic description allows for the arrangement of speech types along a cline from most free to most narratorially controlled (Leech and Short, 1981/2007: 276).

In fact, the most that can be said about literary faithfulness is that the form of presentation in a fictional text makes a claim to authenticity that a reader will acquiesce in. Importantly, as stressed in Stockwell (2009), the plausibility and authenticity of characters 'do not depend on realism and believability' (p. 115). For speech in particular, this implies that direct speech does not have to be exactly the same as spoken language to create an impression of authenticity, or as Short et al. (2001) point out, '[w]hen we read novels we pretend that the reports are real' (p. 494). However, we need to address this readerly engagement with fictionality directly in order to recognise that in this case literary fiction is not like other modes and genres of discourse. The main difference from everyday discourse is the necessary fictional switch into another ontological level (or 'world') when we begin to engage with fiction. This crucial, defining difference necessarily alters our understanding of issues such as faithfulness, authenticity and character-autonomy, but it has no automatic bearing on the forms of expression and the meaningful content of speech. So for these latter issues, we can be sure that analytical techniques for exploring everyday discourse, and (as in this article) corpus linguistic methods, can be relied on to be consistent even across a world-boundary. It is the textual functions of linguistic forms that create readerly effects of naturalness. Because of the continuity between real and fictional people, and the similarity in the way in which readers mind-model both, a deeper understanding of the similarities between real and fictional speech will thus enable us to better account for these effects.

The study of speech presentation is not the only approach to fictional speech. Thomas (2012) sets out to bring together some of the theories and models of fictional dialogue that take account of developments and changing perspectives in various fields. Crucially, she is interested in 'the specific ways in which novelists have responded to changing attitudes to and modes of speech' (Thomas, 2012: 4). Particularly relevant to the present approach is her observation that often the individualising function of fictional speech is viewed in too simple terms. Thomas (2012) argues that rather than specific language features, the development of conversations and the gradual emergence of habitual behaviour of characters that only becomes apparent in the course of the text contribute to the creation of fictional characters. Linguistic approaches that can account for such development have been illustrated in, for instance, Short (1996) and Toolan (1985), where pragmatic principles, conversation analysis or discourse analysis are drawn on to describe fictional dialogue. In this article, we add corpus linguistic methods and approaches to the rich study of fictional speech.

3. Real spoken language

Page (1988) makes a point that is crucial to this article: 'it seems probable that the whole concept of realism as applied to fictional speech is often based on an inadequate or inaccurate notion of what spontaneous speech is really like' (pp. 3–4). The implications are significant if we assume, as argued in the previous section, that analytical techniques for

the analysis of linguistic forms of spoken language work across world-boundaries. Since Page (1988) wrote his criticism, corpus linguistic research has made it possible to capture evidence of real spoken language at scale and develop new descriptions of spoken registers. So our 'notion of what spontaneous speech is really like' is rather different today. Leech (2000) observed two trends in corpus linguistics: a 'sameness' approach and a 'differentness' approach. He views the 'Nottingham School' as a particular proponent of the differentness approach where no assumption can be made that speech and writing share the same grammatical framework (Leech, 2000: 689). With their *Cambridge Grammar of English*, Carter and McCarthy (2006) tackle the fact that the written language has long been taken 'as a benchmark for what is proper and standard in the language' (p. 9). Stressing the recentness of advances in audio-recording and technology that enable corpus research, Carter and McCarthy (2006) argue that new insights into spoken language also require new grammatical concepts and terminology.

For Leech (2000: 689), the sameness approach is illustrated by the *Longman Grammar of Spoken and Written English* (Biber et al., 1999) which continues the framework of Quirk et al. (1985). Biber et al. (1999) generally use the same descriptive framework for spoken and written language, which allows them to make comparisons across different registers (Leech, 2000: 690). What Carter and McCarthy (2006) and Biber et al. (1999) crucially share is the recognition of the importance of frequently repeated sequences of words in a corpus-informed grammar of English. Biber et al. (1999) identify 'lexical bundles' as 'the sequences of words that most commonly co-occur in a register' (p. 989). They use lexical bundles to compare conversation and academic prose and find, for instance, that about 30% of the words in conversation occur in lexical bundles; in academic prose it is 21% (Biber et al., 1999: 995). In conversation, most lexical bundles are part of declarative clauses or questions (e.g. *can I have a*), whereas in academic prose, lexical bundles are often part of noun phrases or prepositional phrases (e.g. *the nature of the*). Although Carter and McCarthy (2006) use the term 'cluster', they describe the same phenomenon when they emphasise that 'the most common clusters differ between written and spoken texts' (p. 828). Lexical bundles have become an important means to compare varieties of the language and to hone in on specific discourse functions within a register. These repeated sequences of words are a manifestation of what Sinclair (1991) describes as the 'idiom principle', that is, the selection of semi-pre-constructed phrases that account for an important part of language use, because they facilitate both language production and reception (see also Wray, 2002). Especially because they are frequent and reduce processing effort, language users are not easily aware of lexical bundles. It was only through corpus linguistics that the extent of the lexico-grammatical patterning of the language came into focus (and lexical bundles are just one aspect of this patterning). So it is not surprising that Page (1988), and research in his tradition, approached speech in fiction in a particular way. Now that we have insights from corpus linguistics, it is time to take a fresh view.

Corpus linguistics has even further relevance for this re-assessment. Carter (2004) makes the case for a 'cline of literariness' (see also Carter and Nash, 1990). Investigating linguistic creativity with the help of corpus data, he argues that the division between literary and non-literary language cannot be reasonably upheld. With their corpus approach to the speech and thought representation model, Semino and Short (2004) also

demonstrate similarities between fiction and non-fiction. In this article, we argue that the fuzziness between literary and non-literary language, or the cline of literariness, is relevant to the nature of fictional speech, too, and what is considered a sense of realism needs to be understood in that context.

4. Methodology

We aim to identify and describe crucial similarities between fictional speech and real spoken language that can create a readerly sense of naturalness and authenticity. Our methodology is innovative in several respects. We focus on the actual utterances of direct speech in fiction rather than the framing that integrates these utterances into the text. This focus is made possible through our innovative corpus tool CLiC (Corpus Linguistics in Context) and the corpora associated with it (see section 4.1 below). The way we analyse direct speech is through the study of clusters – which we interpret functionally as 'speech-bundles' (section 4.2). While there are limitations to lexical bundles mainly because of their fixedness, repeated sequences of words are a useful starting point to identify textual functions (Mahlberg, 2013). Finally, we propose a comparative approach to the description of speech-bundles that overcomes the limitations of the use of cut-off points in the study of lexical bundles (section 4.3). Corpus linguistic methods have been applied to narrative fiction in general before; such approaches are often captured under the umbrella of 'corpus stylistics' (for an overview of such work and links to digital humanities see, for example, Mahlberg and Wiegand, in press). The analysis of the content of fictional speech is technically more straightforward for drama or TV dialogue because of the format of the data (e.g. Bednarek, 2018; McIntyre, 2016). Small-scale corpus studies on fictional speech have been conducted, for instance, by De Haan (1996) and Oostdijk (1990), but with a focus on more formal features. Axelsson (2009) specifically observes, '[i]t is remarkable that so little quantitative research on the language of direct speech in fiction has been published so far' (p. 190). She sees a particular reason for this in the lack of suitable corpora – which we directly address with CLiC.

4.1. CLiC and the corpora for this study

We developed the CLiC tool (http://clic.bham.ac.uk) as our main research instrument for the study of textual patterns in 19th-century fiction. CLiC is a combined set of corpora with a web interface that allows for a range of search possibilities including concordances, keywords and clusters. What distinguishes CLiC from standard corpus tools is primarily the facility to separately search quoted material (typically this will be direct speech, rather than direct thought or writing) and the narratorial material surrounding these quotes. In Example (1), the text within quotation marks belongs to the 'quotes' subset of the corpora. The text outside quotation marks (italicised in the example) is included in 'non-quotes'. CLiC distinguishes further subtypes not relevant to this article, and the initial annotation process of subsets is explained in Mahlberg et al. (2016) (as CLiC is an expanding resource, see also the online user guide for updates: Mahlberg et al., 2019).

Table 1. Corpora used in this study (frequencies based on CLiC 1.6.1 and the XML version of the BNC1994).

Corpus	Corpus size (in million words)	Quotes subset/spoken subcorpus size (% of corpus)
Dickens' Novels (DNov)	3,835,807	1,369,029 (36)
19th Century Reference Corpus (19C)	4,513,070	1,611,083 (36)
19th Century Children's Literature (ChiLit)	4,443,542	1,511,497 (34)
BNC1994	97,639,023	9,899,403

BNC: British National Corpus.

(1) *Leaving the carriage at the posting-house and ordering fresh horses to be ready, my companion gave me his arm, and we went towards home*.

"As this is your regular abode, Miss Summerson, you see," *he observed*, "I should like to know whether you've been asked for by any stranger answering the description, or whether Mr. Jarndyce has. I don't much expect it, but it might be." (*Bleak House*)

The CLiC corpora we used for this article are the corpus of Dickens' Novels (DNov), a 19th Century Reference Corpus (19C) by other prominent British 19th-century writers and a corpus of 19th Century Children's Literature (ChiLit). 19C was originally created as a reference corpus to contextualise findings from DNov (Mahlberg, 2013). ChiLit (Cermakova, 2018) widens the perspective by contributing data from children's literature, which shows how genre has an effect on speech in fiction, too.

Table 1 provides the number of words in each corpus, as well as the number of words in the quotes subsets. We use the spoken component of the BNC (British National Corpus) as data of real spoken language. Following Love et al. (2017), we refer to the 'BNC1994' to distinguish the corpus from the 'BNC2014'. The time gap between the CLiC corpora and the BNC deserves comment. As present-day readers, we do not have exposure to real spoken English of the 19th century. When we read 19th-century fiction, we bring our real-life experience to bear. The BNC serves as a proxy for this spoken language experience. Whether we know what real 19th-century speech would have sounded like is not crucial. What matters is the readers' basis for comparison, that is, contemporary speech (as well as the reader's experience of reading fiction). When we started this research, the full text of the BNC2014 was not available yet. In future research, this corpus can add additional dimensions for comparisons that might enable a more nuanced assessment of the experience of 'contemporary readers', because demographic factors of readers, such as age, will affect how well the BNC1994 or BNC2014 serves as a proxy for their background knowledge.

4.2. Speech-bundles

Biber et al. (1999) define a lexical bundle with regard to particular thresholds of frequency (per million words) and range (the number of texts it appears in). These criteria can vary depending on the aims of the research. Hence, the use of the terms 'lexical

bundles', 'clusters' or also 'chunks' and 'n-grams' is not consistent across corpus linguistic research. The *Longman Grammar of Spoken and Written English* (Biber et al., 1999), which gave initial prominence to lexical bundles, is based on a large (40-million-word) general corpus with relatively evenly balanced 'core' register components. Biber et al. (1999) set their frequency and range thresholds in order to avoid 'idiosyncrasies' and to describe the general usage of lexical bundles across a given register. As shorter bundles are more common than longer ones, they suggest a minimum threshold for four-word bundles at ten per million and for five- and six-word bundles at five per million. For studies of large general corpora, the range criterion can be redundant in practice, because frequency thresholds often presuppose range (see Biber et al., 2004: 376). The situation is different for smaller, specialised corpora, where a highly frequent cluster may not necessarily be widely spread across texts or where functions of clusters can still be identified if the frequency threshold criterion is not met (Mahlberg, 2013).

Kopaczyk (2013) provides an overview of the cut-off points used in several lexical bundle studies, covering both general and specialised corpora (as small as 220,000 words in the case of Culpeper and Kytö, 2010). She finds wide variation, with frequency being reported both in normalised terms (e.g. per million words) and in absolute terms (e.g. ten occurrences in the target corpus), and some studies treating the range of bundles in terms of a minimum number of texts or a percentage of the files in the corpus. Kopaczyk (2013) concludes that '[t]here is no uniform practice in lexical bundle studies [. . .] Every researcher takes an informed but idiosyncratic decision' (p. 152).

An argument against fixed thresholds comes from psycholinguistics, where frequency effects on learners have been shown to be variable rather than divisible into frequency bins (Arnon and Snider, 2010). Based on this argument, Durrant (2017: 170) develops a formula to quantify the overlap of the number of bundles (in terms of tokens rather than types) shared between subcorpora of student writing. The move away from cut-offs and differences between (sub)corpora towards a graded perspective makes Durrant's (2017) approach similar to ours.

Our unit of analysis is the five-word cluster. A length of five has been shown to be a useful starting point for the analysis of fiction (Mahlberg, 2013). Following Scott (2019), we use the term 'cluster' to refer to a unit that is purely defined by surface features. We refer to 'speech-bundles' to emphasise those textual functions of clusters that create a sense of spokenness, that is, the way in which they are able to create readerly effects of naturalness of speech. So every speech-bundle is a cluster, but not every cluster is a speech-bundle. When our analysis aims to focus on the most frequent examples, we do refer to cut-offs. Following Biber et al. (1999: 993), we use a frequency of five per million words and occurrence across a range of at least five texts. Through additional comparisons, we ensure our study is not limited to speech-bundles defined by fixed criteria, but takes a more nuanced approach.

4.3. Comparisons and the continuum of fictional speech

Frequency thresholds are not absolute; they have to be seen along a continuum. To describe the continuum of real spoken language and speech in fictional texts, we take the following steps to compare clusters:

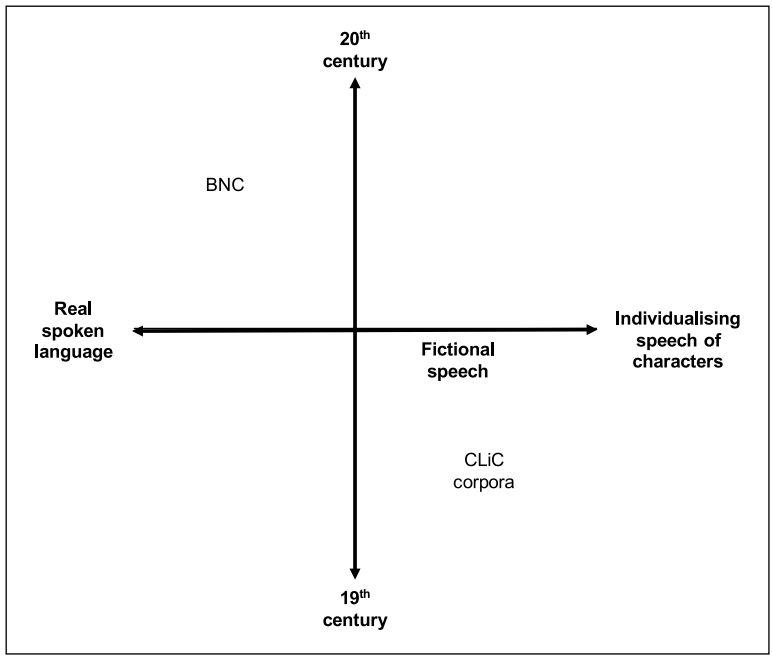


Figure 1. The continuum of real and fictional speech and change over time.

1. We compare the BNC with the CLiC corpora to identify similarities between fiction and real spoken language;
2. We identify similarities in fictional speech across the CLiC corpora;
3. We look at author-specific fictional speech;
4. We compare the quotes and non-quotes subsets of the CLiC corpora.

Steps 1–3 allow us to study clusters along two dimensions – real spoken language and fictional speech on one, and 19th- and 20th-century speech patterns on the other, as shown in Figure 1. Step 4 adds an intratextual comparison, that is, a key cluster comparison that complements the other three steps by showing how fictional speech is characterised through being different from narratorial text. A key cluster comparison uses clusters, instead of words, as units in key comparisons (Mahlberg, 2007). For the four comparisons, we use rank-ordered frequency lists to complement minimal occurrence thresholds, but in both cases the comparison across corpora will be crucial to account for the fuzzy boundaries between different data sets. Detail on the individual comparisons will be provided in the following sections.

Across all sections we used a combination of the CLiC web interface and its 'API' ('application programming interface'; a way of accessing the data with a programming language). For this article, we used a development version of the CLiC API (for documentation of the current API, see https://clic.readthedocs.io/en/latest/advanced/api_usage.html). It is important to note that CLiC is a developing resource. The data for this article

are mainly based on versions 1.6.1 and 1.7. CLiC has since been further expanded with version 2.0 (used for Table 10 of this article), so if readers of this article run searches on live CLiC there might be small quantitative differences, which, however, will not affect the overall results. For comparisons with real spoken language, we used the 'original' BNC1994 (XML version http://www.natcorp.ox.ac.uk/) to generate cluster counts and, for concordance searches, we used the BNCweb (Hoffmann et al., 2008). For each results table below, we indicate the respective data source. The R code (R Core Team, 2018) for retrieving the data sets and the large-scale comparisons between the CLiC corpora and the XML version of the BNC is available from our GitHub repository (https://github.com/birmingham-ccr/sm/tree/master/Mahlberg_et_al_supplementary_material).

5. A speech-bundle approach to fictional talk

The separate 'quotes' and 'non-quotes' subsets of CLiC are crucial data sets to gain insights into the nature of fictional speech. Table 2 shows the top of the frequency list of five-word clusters across all three fiction corpora combined, based on cluster counts in the whole texts. The table shows 17 clusters, which cover the top 16 frequency ranks (clusters 16 and 17 have the same frequency so appear on the same rank). The clusters in Table 2 have a frequency of at least ten per million words for the whole texts ('all text'). Most of them are relatively more frequent in the non-quotes subset, as indicated by the final two columns, providing relative frequencies for quotes and non-quotes (of the 12.8 million words in the three corpora combined, 8.3 million appear in non-quotes and 4.5 million in quotes). Only *what do you think of* and *I should like to know* are relatively more frequent in quotes. So features of fictional speech might appear less noticeable when texts are treated as whole units.

A cluster list for 'all text' also illustrates the effects of thresholds to define lexical bundles. With the commonly used cut-off of five per million (see section 4.2), *I think it would be* is an example that would not be captured as a lexical bundle – it occurs 3.05 per million in 'all text'. In quotes, however, its relative frequency of 8.68 makes it above this cut-off. We will look at this cluster in more detail in section 5.1.

5.1. Fictional and natural speech-bundles

Once quotes are separated out, features of spokenness become more clearly visible. In this section, we look at the quotes subsets of the three fiction corpora individually to identify patterns in a more nuanced way for the comparison with the Spoken BNC1994. Table 3 provides an overview of the top five ranks for each corpus (as in Table 2, bundles with the same frequency are presented on the same rank), which immediately shows there is overlap between the most frequent bundles across the corpora. The bundles in bold appear in the top five frequency ranks of at least two corpora. There is overlap between the fiction corpora but not with the Spoken BNC1994. The relative frequencies in the BNC are the highest – the top bundle has a relative frequency of over 70 per million. In DNov, the top bundle occurs almost 50 times per million, whereas for 19C and ChiLit the top frequency is just over 30 times per million. In the Spoken BNC1994, *at the end of the* is the most frequent bundle. Its frequency is largely due to the six-word

Table 2. Clusters with a relative frequency of at least ten per million, in 'all text' of the DNov, 19C and ChiLit corpora combined (raw frequencies based on the CLiC 1.7 Cluster tab).

Rank	Cluster	Raw frequency all text	Freq pm all text	Freq pm in quotes	Freq pm in non-quotes
1	the other side of the	294	22.98	14.92	27.37
2	in the middle of the	288	22.51	11.80	28.58
3	at the end of the	263	20.56	9.13	26.77
4	on the other side of	221	17.28	11.35	20.47
5	a quarter of an hour	220	17.20	12.47	19.74
6	at the bottom of the	199	15.56	6.68	20.47
7	in the course of the	197	15.40	7.79	19.62
8	as if he had been	186	14.54	2.89	21.07
9	in the direction of the	161	12.59	2.45	18.17
10	as if it had been	160	12.51	4.68	16.96
11	at the top of the	158	12.35	6.46	15.74
12	at the foot of the	146	11.41	4.23	15.38
13	at the back of the	142	11.10	4.23	14.78
14	what do you think of	141	11.02	30.50	0.48
15	in the midst of the	129	10.08	2.00	14.53
16	the other end of the	128	10.01	5.12	12.60
	I should like to know	128	10.01	26.49	0.97

DNov: Dickens' Novels; 19C: 19th Century Reference Corpus; ChiLit: 19th Century Children's Literature; freq pm: frequency per million.

bundle *at the end of the day*, which is often used to mean 'ultimately', as illustrated in Example (2). The six-word bundle accounts for almost half of the occurrences of *at the end of the* – 332 out of 705 – and relates to the bundle *the end of the day* on rank 4 of the frequency list.

(2) She's had her chance and I'm doing all I can to help her but **at the end of the day** it's on her shoulders. (Spoken BNC1994, file H02)

In Table 2, *at the end of the* also occurs near the top. However, in the fiction corpora, it is more frequent in non-quotes, where it is used with a time or place meaning, as in Example (3).

(3) "What can it matter?" cried Dorian Gray, laughing, as he sat down on the seat **at the end of the** garden. (*The Picture of Dorian Gray*)

The top most frequent bundles provide a good starting point to identify crucial textual functions in a corpus. Extending Table 3, Table 4 covers all bundles that appear in the top 15 cluster frequency ranks in at least two of the corpora. The table is ordered alphabetically and indicates the frequency rank for each shared bundle. For instance, *I should like to know*

Table 3. Top five most frequent cluster ranks per corpus (frequencies based on the API cluster endpoint for CLiC 1.6.1 and the XML version of the BNC; extended tables available in the supplemental material, Appendix 1).

Rank	Spoken BNC1994	Freq pm	DNov	Freq pm
1	at the end of the	71.32	**what do you mean by**	48.94
2	mm mm mm mm mm	60.10	**what do you think of**	41.64
3	you know what i mean	49.40	I beg your pardon sir	35.79
4	the end of the day	35.05	very much obliged to you	27.03
5	one two three four five	33.84	to tell you the truth	24.10
Rank	**19C**	**Freq pm**	**ChiLit**	**Freq pm**
1	**I should like to know**	30.41	the other side of the	31.10
2	**what do you think of**	29.79	**I should like to see**	28.45
3	**I should like to see;** I want to speak to; it seems to me that	18.00	**I should like to know**	26.46
4	I am not going to	17.38	**what do you mean by**	25.80
5	a quarter of an hour	16.76	on the other side of	23.82

API: application programming interface; BNC: British National Corpus; DNov: Dickens' Novels; 19C: 19th Century Reference Corpus; ChiLit: 19th Century Children's Literature; freq pm: frequency per million.

ranks sixth in DNov, first in 19C and third in ChiLit. For the 'top 15' criterion, the overlap is greater between the fiction corpora than with the BNC. Still, speech-bundles shared across the fiction corpora do reflect features of spoken language. In Table 4, most refer to interactions between people. First- and second-person pronouns, as in *it seems to **me** that* and *what are **you** going to*, or question fragments (*what do you mean by*, *what do you say to*) are features of what Carter and McCarthy (2004) call the 'speaker-listener world'. In addition, bundles that contain verbs like *know*, *think* or *mean* show how knowledge is monitored and assertions hedged, which belongs to the central functions of lexical bundles in spoken language (Carter and McCarthy, 2006: 835).

While fewer of the top frequency bundles of real spoken language are found in Table 4, Table 5 focuses on the overlap between real language and fictional speech by highlighting the speech-bundles that occur at least five times per million and appear in at least five different texts in all four corpora. From Table 5 we can see that even if speech-bundles are shared across all corpora, there are author-specific tendencies. Referring to the continuum of real and fictional speech in Figure 1, *what do you think of* and *what do you mean by* tend more towards the fictional end of the cline. Both are particularly frequent in DNov, and especially *what do you mean by* is an example of a bundle that helps to individualise fictional characters. As shown in Mahlberg (2013), Dickens uses *what do you mean by* to depict confrontation between fictional characters. It is a device of formulaic impoliteness (section 5.3 goes into more detail of speech-bundles to individualise characters). Table 5 further shows how speech-bundles may reflect features of genre. In Example (4), *the other side of the* illustrates how references to physical space and activity are commonly found in children's stories.

Table 4. Shared speech-bundles among the top 15 cluster frequency ranks per corpus, ordered alphabetically. All examples in Table 4 occur at least five times per million and in at least five texts per corpus (based on the API cluster endpoint for CLiC 1.6.1 and the XML version of the BNC; extended tables available in the supplemental material, Appendix 1).

Cluster	Spoken BNC1994	DNov	19C	ChiLit
as well as I do		14	9	
at the end of the	1		15	11
I am not at all		12	15	
I am not going to		7	4	8
I am sure you will		15	8	
I don't know that I		8		15
I don't know what I		6	13	
I don't know what you			13	11
I don't want to be			14	8
I should like to know		6	1	3
I should like to see			3	2
I want to speak to		7	3	
in the middle of the	12		15	7
it seems to me that	14		3	12
the other side of the	15			1
to tell you the truth		5	13	
very much obliged to you		4	12	
what am I to do			9	15
what do you mean by		1		4
what do you say to		13	11	
what do you think of		2	2	7
what is the matter with			14	15
what is to be done			9	6
you don't mean to say		8	12	

API: application programming interface; BNC: British National Corpus; DNov: Dickens' Novels; 19C: 19th Century Reference Corpus; ChiLit: 19th Century Children's Literature.

(4) "Never mind, we can hurl them like javelins," said Cyril, "or drop them on people's heads. I say--there are lots of stones on **the other side of the** courtyard. If we took some of those up? Just to drop on their heads if they were to try swimming the moat." (*Five Children and It*)

Because of the relatively low frequencies in Table 5, we do not apply statistical tests, but look at some examples in more detail. In a similar way to Table 4, the examples in Table 5 reflect the interpersonal meanings of speech-clusters. Like devices for hedging the assertiveness of the speaker, vagueness expressions are common in everyday conversation. The speech-bundle *and all that sort of* in Table 5 is such an example. In Example (5), it illustrates how vague language indicates 'an assumed shared knowledge' (Carter and McCarthy, 2006: 202).

(5) "My belief is, that she's been an ill-used woman," said Cradell. "If she had a husband that she could respect and have loved, **and all that sort of** thing, she would have been a charming woman." (*The Small House at Allington*)

Table 5. All five-word speech-bundles (occurring at least five times per million, in at least five texts per corpus) that are shared across all four corpora, based on the API cluster endpoint for CLiC 1.6.1 and the XML version of the BNC.

Rank[a]	Speech-bundle	Spoken BNC1994	DNov		19C		ChiLit	
			Freq pm	Range	Freq pm	Range	Freq pm	Range
1	it seems to me that	16.77	7.30	8	18.00	10	14.56	12
2	the other side of the	15.86	5.84	7	7.45	10	31.10	25
3	I think it would be	12.83	6.57	6	7.45	10	11.91	13
4	what do you think of	10.40	41.64	14	29.79	17	21.17	16
5	are you going to do	10.00	5.84	7	8.07	12	11.25	15
6	what are you going to	9.80	7.30	8	8.07	12	9.92	12
7	and all that sort of	7.37	10.96	5	6.83	5	10.59	10
8	I don't know what to	6.16	18.99	11	8.69	8	11.91	13
9	what do you mean by	5.76	48.94	15	8.69	11	25.80	21

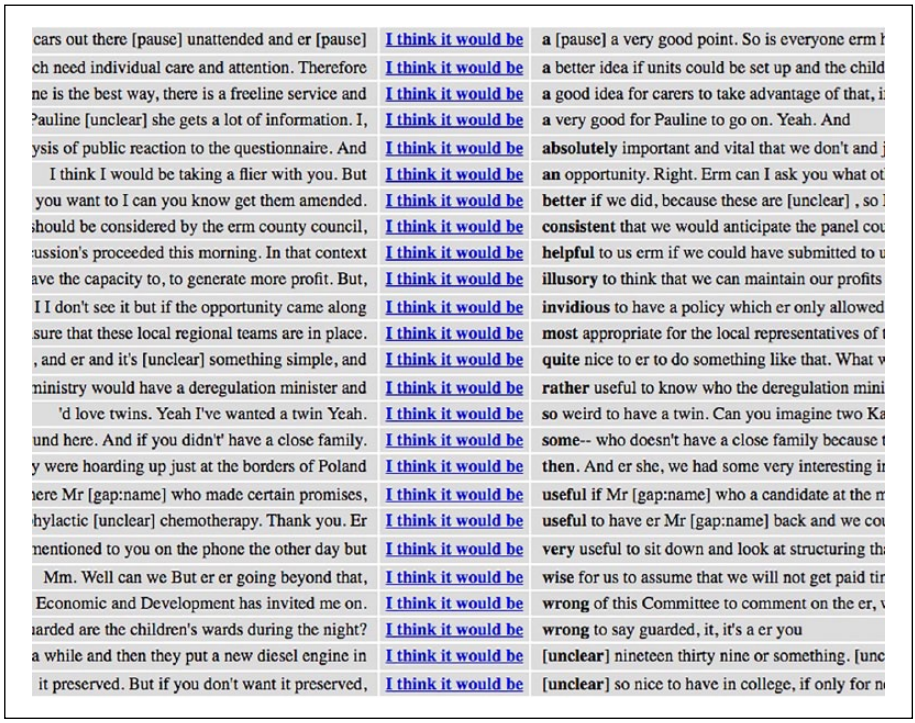
API: application programming interface; BNC: British National Corpus; DNov: Dickens' Novels; 19C: 19th Century Reference Corpus; ChiLit: 19th Century Children's Literature; freq pm: frequency per million.
[a]This is not the rank of an entire frequency list, but a rank of the bundles meeting the set criteria. The same applies to Tables 7 and 9.

Only *I think it would be* is more frequent in the BNC than in the other corpora. Note this is an example that does not make the five-per-million threshold for lexical bundles when the fiction corpora are treated together as one corpus with whole texts, as in Table 2. It strikingly shows how focusing on quotes enables a more nuanced analysis. In all four corpora, *I think it would be* is mostly used to hedge suggestions that speakers make about what should be done or would be better or advisable to do. Figure 2 shows 25 random examples from the BNC. Adjectives from Figure 2 that are also among the repeated adjectives in the full concordance of the Spoken BNC1994 include *better*, *helpful*, *appropriate*, *nice*, *useful*, *wise*, *wrong.*

Figure 3 shows 15 of the 39 examples in the fiction corpora, where we find a similar picture. Lines 12–15 are four of the eight examples that are followed by *better*; in two of these cases, the phrase is *it would be much better.* As an example of context, the wider text from line 12 in Figure 3 is given in (6). Figure 4 illustrates a case that functions in a similar way in the Spoken BNC1994, where people are speaking about what would be better for another person.

(6) "Poor Alicia is rather jealous of any attention Mr. Audley pays me, and--and--**I think it would be** better for her happiness if your nephew were to bring his visit to a close." (*Lady Audley's Secret*)

For most speech-bundles in Table 5, their highest frequency is found in one of the fiction corpora rather than in the BNC. This might relate to the larger range of bundles in the Spoken BNC1994 overall. Table 6 shows the total number of clusters in each corpus. These totals refer to cluster types (the different clusters found in the corpus). Table 6 provides frequency brackets, so [5,10) means DNov contains 296 clusters that occur (in



cars out there [pause] unattended and er [pause]	I think it would be	**a** [pause] a very good point. So is everyone erm h
ch need individual care and attention. Therefore	I think it would be	**a** better idea if units could be set up and the child
ne is the best way, there is a freeline service and	I think it would be	**a** good idea for carers to take advantage of that, i
Pauline [unclear] she gets a lot of information. I,	I think it would be	**a** very good for Pauline to go on. Yeah. And
ysis of public reaction to the questionnaire. And	I think it would be	**absolutely** important and vital that we don't and j
I think I would be taking a flier with you. But	I think it would be	**an** opportunity. Right. Erm can I ask you what ot
you want to I can you know get them amended.	I think it would be	**better** if we did, because these are [unclear] , so I
should be considered by the erm county council,	I think it would be	**consistent** that we would anticipate the panel cou
ussion's proceeded this morning. In that context	I think it would be	**helpful** to us erm if we could have submitted to u
ave the capacity to, to generate more profit. But,	I think it would be	**illusory** to think that we can maintain our profits
I I don't see it but if the opportunity came along	I think it would be	**invidious** to have a policy which er only allowed
sure that these local regional teams are in place.	I think it would be	**most** appropriate for the local representatives of t
, and er and it's [unclear] something simple, and	I think it would be	**quite** nice to er to do something like that. What w
ninistry would have a deregulation minister and	I think it would be	**rather** useful to know who the deregulation mini
'd love twins. Yeah I've wanted a twin Yeah.	I think it would be	**so** weird to have a twin. Can you imagine two Ka
und here. And if you didn't' have a close family.	I think it would be	**some**-- who doesn't have a close family because t
y were hoarding up just at the borders of Poland	I think it would be	**then**. And er she, we had some very interesting in
here Mr [gap:name] who made certain promises,	I think it would be	**useful** if Mr [gap:name] who a candidate at the n
hylactic [unclear] chemotherapy. Thank you. Er	I think it would be	**useful** to have er Mr [gap:name] back and we cou
nentioned to you on the phone the other day but	I think it would be	**very** useful to sit down and look at structuring th
Mm. Well can we But er er going beyond that,	I think it would be	**wise** for us to assume that we will not get paid tir
Economic and Development has invited me on.	I think it would be	**wrong** of this Committee to comment on the er, v
uarded are the children's wards during the night?	I think it would be	**wrong** to say guarded, it, it's a er you
a while and then they put a new diesel engine in	I think it would be	**[unclear]** nineteen thirty nine or something. [unc
it preserved. But if you don't want it preserved,	I think it would be	**[unclear]** so nice to have in college, if only for n

Figure 2. Twenty-five randomly selected examples of the 127 occurrences of *I think it would be* in the Spoken BNC1994, sorted on the right (retrieved with the BNCweb and its random selection function).

1	rees, some of which are probably very good to eat,	I think it would be	a capital plan to make bows and arrows, with which	coral
2	s up properly.Before they started, William observed,	"I think it would be	a good thing, if Ready and I were to take	masterman
3	can't, after all I have gone through.""Well," I said,	"I think it would be	a real shame if you were to bite or kick	beauty
4	here.""I'll tell you why I proposed it," said Margaret	."I think it would be	a relief for Ethel to escape from Miss Winter's belo	daisy
5	our planning it, your saying to yourself one idle day,	`I think it would be	a very good thing for Miss Taylor if Mr. Weston	emma
6	new place.Let me see, where can I put it	?""I think it would be	a very good plan if you did put it up	woodmagic
7	have the life of a fellow-creature on my conscience	."I think it would be	advisable, Edward, that I should set off early to-mo	forest
8	nd have worked there, and have slept soundly there	.I think it would be	almost cowardly and cruel not to have some little a	LD
9	ll we move farther back upon higher ground?" "Yes,	I think it would be	as well."So as the fog-bank flowed onward we fell	basker
10	from a great many strange places, sir," said I; "but	I think it would be	as well to tell you where and how in a	kidnap
11	ttle.'This proceeding is more absurd than the other	."I think it would be	best to take it in,' replied Mr. Ben Allen; 'it	PP
12	lous of any attention Mr. Audley pays me, and--and	--I think it would be	better for her happiness if your nephew were to bri	LadyAud
13	them now that it was he who sent the necklace	."I think it would be	better for men not to gamble.It is a besotting	Deronda
14	ward to the happiness of having Ada with me again,	I think it would be	better for us."I hope it was not a poor	BH
15	ter.""I will talk to him," said Margaret, "and, indeed,	I think it would be	better than worrying papa.""Well," said Ethel, "of co	daisy

Figure 3. Fifteen of the 39 concordance lines of *I think it would be* in quotes in DNov, 19C and ChiLit, sorted on the right (one additional concordance line occurs in non-quotes in the first-person narration of *Bleak House*), retrieved with CLiC 1.7.

Note: In the current version of CLiC 2.0.2, two instances are found in non-quotes (and only 38 in quotes), because one instance is mistagged in ChiLit due to a different tagging algorithm.

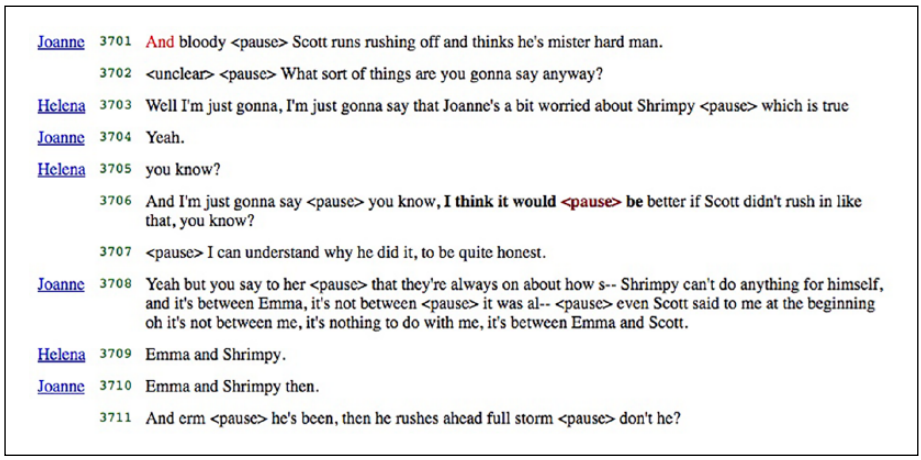

Joanne 3701 And bloody <pause> Scott runs rushing off and thinks he's mister hard man.
3702 <unclear> <pause> What sort of things are you gonna say anyway?
Helena 3703 Well I'm just gonna, I'm just gonna say that Joanne's a bit worried about Shrimpy <pause> which is true
Joanne 3704 Yeah.
Helena 3705 you know?
3706 And I'm just gonna say <pause> you know, **I think it would <pause> be** better if Scott didn't rush in like that, you know?
3707 <pause> I can understand why he did it, to be quite honest.
Joanne 3708 Yeah but you say to her <pause> that they're always on about how s-- Shrimpy can't do anything for himself, and it's between Emma, it's not between <pause> it was al-- <pause> even Scott said to me at the beginning oh it's not between me, it's nothing to do with me, it's between Emma and Scott.
Helena 3709 Emma and Shrimpy.
Joanne 3710 Emma and Shrimpy then.
3711 And erm <pause> he's been, then he rushes ahead full storm <pause> don't he?

Figure 4. Dialogue in the Spoken BNC1994 (file KCE) containing the bundle *I think it would be*.

Table 6. Number of five-word clusters per corpus per million words (all with a range ⩾five texts per corpus, that is, a minimum raw frequency of five), based on the API cluster endpoint for CLiC 1.6.1 and the XML version of the BNC.

Frequency band per million	Spoken BNC1994	DNov	ChiLit	19C
[0,5)	13,165	186	356	324
[5,10)	97	296	238	175
[10,20)	28	85	39	28
[20,40)	5	13	8	2
[40,Inf)	3	2	NA	NA

API: application programming interface; BNC: British National Corpus; DNov: Dickens' Novels; 19C: 19th Century Reference Corpus; ChiLit: 19th Century Children's Literature.

at least five texts) a minimum of five times per million and a maximum of $9.\overline{9}$ times per million ($5 \leqslant \text{frequency} < 10$). For example, *it seems to me that*, which has a relative frequency of 16.77 in the BNC, is covered by the bracket [10,20) in Table 6. Compared to the other corpora, the BNC contains a lot more cluster types occurring under five times per million (in at least five texts). In addition, the difference between the first and second frequency band is more drastic than for the other corpora. Generally, the number of types decreases as the frequency increases, except for DNov. It seems Dickens especially uses five-word clusters in the frequency band of [5,10).

To describe the furthest end of real spoken language in Figure 1, Table 7 shows the most frequent speech-bundles in the Spoken BNC1994 that are not found in the fiction corpora. The table points to a wider range of 'relational language' (O'Keeffe et al., 2007: 159), which 'serves to create and maintain good relations between the speaker and the hearer', by contrast with 'transactional language' for which the exchange of information is the main function. Discourse markers (*I think*, *I mean*) as well as bundles of vagueness

Table 7. Speech-bundles (freq pm ≥ 5) in the Spoken BNC1994 that do not occur at all in any other corpus, based on the API cluster endpoint for CLiC 1.6.1 and the XML version of the BNC.

Rank	Speech-bundle	Frequency	Range	Freq pm
1	mm mm mm mm mm	595	21	60.10
2	da da da da da	185	32	18.69
3	one two three four one	95	5	9.60
4	I think a lot of	91	64	9.19
5	doo doo doo doo doo	90	15	9.09
6	four one two three four	89	7	8.99
7	I mean i don't know	78	51	7.88
8	a hell of a lot	75	55	7.58
	as a result of the	75	49	7.58
	so on and so forth	75	37	7.58
9	going to be able to	67	53	6.77
10	it's a bit of a	62	49	6.26
11	and button two for no	57	20	5.76
12	those in favour of the	56	24	5.66
	call upon councillor to move	56	10	5.66
13	button one for yes and	55	20	5.56
	for yes and button two	55	20	5.56
	one for yes and button	55	20	5.56
	yes and button two for	55	20	5.56
14	want a cup of tea	54	38	5.45
15	on the left hand side	52	38	5.25
	and all of a sudden	52	37	5.25
	how do you feel about	52	37	5.25
16	what are you gonna do	51	34	5.15
17	the end of the month	50	34	5.05
	in such a way that	50	30	5.05

API: application programming interface; BNC: British National Corpus; freq pm: frequency per million.

and approximation (*it's a bit of a*, *a hell of a lot*) come under this heading. In spoken conversation, frequent discourse markers tend to be shorter than five words.

(7) But I, **I mean** when it comes to times tables **that sort of thing** I've never been sorry I learnt those. **I think a lot of** schools do still teach the t-- the children their tables which I think is a good thing. (Spoken BNC1994, file H4C)

Examples like *I mean*, *that sort of thing* and *I think a lot of* are lexical bundles that are frequent in spoken language (e.g. Example (7)), but belong to the features that Page (1988) observes as not being reproduced to the same extent in fiction. Figure 5(a) makes this even clearer. The shorter bundles are significantly more frequent in the Spoken BNC1994 (for the significance test, we used Rayson's (2016) log-likelihood calculator).[1] The bundle *I don't know what I* (Table 4), in contrast, is less frequent in the Spoken BNC1994. To check that what we are seeing in Figure 5(a) are not simply the results of language change over

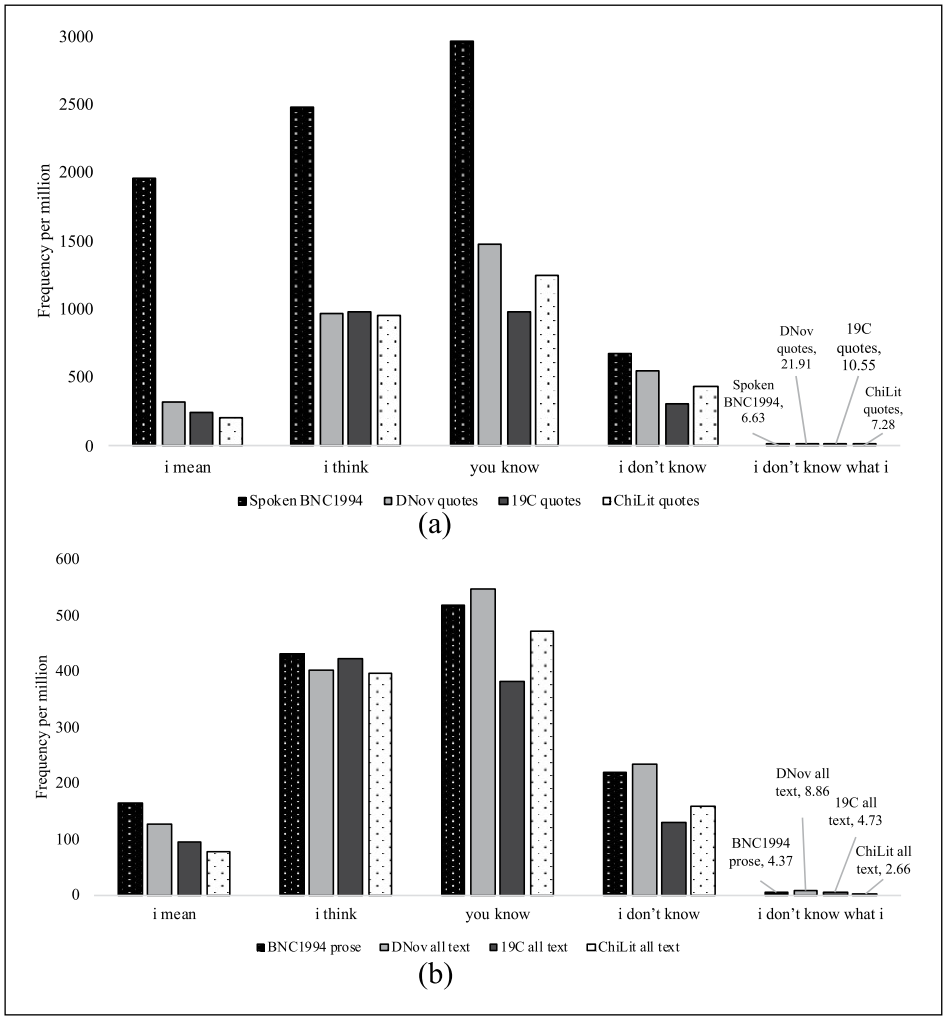


Figure 5. (a) Frequency per million for shorter bundles in the Spoken BNC1994 and the quotes sections of the fiction corpora (frequencies based on the CLiC 1.7 Concordance tab and BNCweb). (b) Frequency per million for shorter bundles in the BNC1994 fictional prose section ('Genre:W:fict:prose') compared to the full text of the CLiC corpora (frequencies based on the CLiC 1.7 Concordance tab and BNCweb).

time, Figure 5(b) compares the same bundles in the full fiction corpora (including non-quotes) with the BNC fictional prose section, finding more similar results.

5.2. Speech-bundles articulating 19th-century politeness

In this section and the next, we focus on speech-bundles that appear to be more specific to fiction. One such group are speech-bundles that function to protect the speaker's face. In

Table 8. Examples of 19th-century fiction politeness bundles across the corpora (based on the BNCweb and the CLiC 1.7 Concordance tab).

Cluster	BNC spoken		DNov		19C		ChiLit	
	Freq pm	Range	Freq pm	Range	Freq pm	Range	Freq pm	Range
I beg your pardon sir	0.1	1	35.79	12	6.21	9	4.63	5
very much obliged to you	0	0	27.03	11	11.17	9	7.94	8
do me the favour to	0	0	21.91	11	0.00	0	0.66	1
I am glad to see	0.1	1	20.45	11	2.48	4	5.95	6
how do you do Mr	0	0	18.26	9	1.24	2	1.32	2
be so good as to	0	0	18.26	7	8.07	8	4.63	5
am glad to see you	0	0	17.53	10	1.24	2	4.63	4
how do you find yourself	0	0	16.07	10	0.00	0	1.32	1
I am obliged to you	0	0	16.07	10	5.59	7	0.66	1
I am very much obliged	0	0	16.07	7	6.21	6	3.97	3
will you allow me to	0.1	1	15.34	8	0.62	1	1.98	3
am very much obliged to	0	0	15.34	7	5.59	6	3.97	3

BNC: British National Corpus; DNov: Dickens' Novels; 19C: 19th Century Reference Corpus; ChiLit: 19th Century Children's Literature; freq pm: frequency per million.

identifying these as politeness markers, we are drawing simply on the traditional Brown and Levinson (1987) model. We started with the top 15 ranks of DNov (cf. Table 4) to select examples because in DNov these ranks have more politeness bundles than any of the other fiction corpora. Table 8 compares the frequencies across all four corpora.

The most frequent politeness bundles refer to common speech acts like apologies (Example (8)), greetings (Examples (9) and (10)), giving thanks (Examples (11) and (12)); such forms are also referred to as 'conversational routines' (Aijmer, 1996; O'Keeffe et al., 2007).

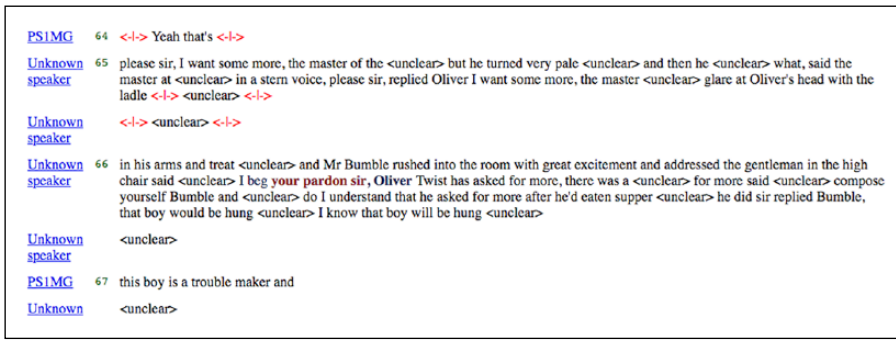
(8) "I'm sure **I beg your pardon, sir**," exclaimed the little woman, seeing that she had made some awful mistake. (*Lady Audley's Secret*)

(9) "**How do you do, Mr**. Vholes? Will you *be so good as to* take a chair here by me and look over this paper?" (*Bleak House*)

(10) "How d'ye do, Crawley? **I am glad to see you**," said Mr. Wenham with a bland smile, and grasping Crawley's hand with great cordiality. (*Vanity Fair*)

(11) "Thank you," said Robert, writing the address in his pocketbook. "**I am very much obliged to you**, and you may rely upon it, Mrs. Vincent shall not suffer any inconvenience through me." (*Lady Audley's Secret*)

(12) "Yes, Uncle," Charlie said cheerfully; "and we are **very much obliged to you**, Mamma and I, for your kindness." (*With Clive in India*)


PS1MG 64 <-I-> Yeah that's <-I->

Unknown speaker 65 please sir, I want some more, the master of the <unclear> but he turned very pale <unclear> and then he <unclear> what, said the master at <unclear> in a stern voice, please sir, replied Oliver I want some more, the master <unclear> glare at Oliver's head with the ladle <-I-> <unclear> <-I->

Unknown speaker <-I-> <unclear> <-I->

Unknown speaker 66 in his arms and treat <unclear> and Mr Bumble rushed into the room with great excitement and addressed the gentleman in the high chair said <unclear> I beg **your pardon sir, Oliver** Twist has asked for more, there was a <unclear> for more said <unclear> compose yourself Bumble and <unclear> do I understand that he asked for more after he'd eaten supper <unclear> he did sir replied Bumble, that boy would be hung <unclear> I know that boy will be hung <unclear>

Unknown speaker <unclear>

PS1MG 67 this boy is a trouble maker and

Unknown <unclear>

Figure 6. Example showing the only use of *I beg your pardon sir* in the Spoken BNC1994 (file F7M).

These bundles are shared across all three fiction corpora. But there seem to be genre differences again. The four bundles around *obliged* are all the least frequent in ChiLit. What appears to be the most formal one, *I am obliged to you*, just occurs once (in F. Anstey's *Vice Versa*). Example (12) referring to Charlie and his mother suggests, for instance, formality is perceived more as part of the adult world. Most politeness bundles in Table 8 do not occur in the Spoken BNC1994. An example with just one occurrence in real spoken language is *I beg your pardon sir*, which occurs in a classroom setting, where the students discuss *Oliver Twist* (see Figure 6), so there is an explicit link to 19th-century speech. Generally, the politeness bundles in Table 8 would mostly sound old-fashioned if used in everyday casual conversation today. Along the dimensions of the continuum in Figure 1, they illustrate change over time indicated by the vertical arrow.

Table 8 includes some bundles which function to soften requests, for example, *be so good as to* italicised in Example (9). For pragmatic functions, the classification of speech-bundles again is best dealt with on a cline rather than in terms of hard-and-fast categories. Both the categorisation of speech acts and the range of politeness strategies have been the focus of much debate. Similarly, numerous definitions of specific face-saving devices like hedges and vague language exist. It is neither within the scope nor the focus of this article to enter these debates. The key point we make is that speech-bundles are associated with pragmatic functions that contribute to the depiction of character relations and interactions. Politeness bundles are frequent examples of this relational language, including conversational routines associated with frequent speech acts, but also a range of face-saving devices that contribute to the creation of common ground between the fictional characters.

5.3. Authorial speech-bundle style

On a cline from real spoken language on one end to idiosyncratic phrases of individual characters on the other, an author's individual style takes an intermediate position (cf. Figure 1). The comparison of clusters across corpora can help to identify features of style (cf. the use of clusters/n-grams in authorship attribution research, Grieve et al., 2019). Table 9 shows all speech-bundles that occur in DNov but are not found in any of the other corpora so can be taken as a pointer to Dickens' style. Neither 19C nor ChiLit contain any speech-bundles with zero occurrences in the remaining corpora. The 19C and ChiLit

Table 9. All speech-bundles in DNov that do not occur at all in any other corpus (based on the API cluster endpoint for CLiC 1.6.1 and the XML version of the BNC).

Rank	Speech-bundle	Frequency	Range	Freq pm
1	hope i see you well	13	7	9.50
2	you do me the favour	12	8	8.77
3	would you be so good	10	5	7.30
4	upon my soul I am	9	7	6.57
	oh dear me dear me	9	6	6.57
5	allow me to ask you	8	5	5.84
	did me the honour to	8	5	5.84
	not the least doubt of	8	5	5.84
	the state of the case	8	5	5.84
	will you hold your tongue	8	5	5.84
	you allow me to ask	8	5	5.84
6	I see you well sir	7	6	5.11
	heaven forbid that I should	7	5	5.11

API: application programming interface; BNC: British National Corpus; DNov: Dickens' Novels; freq pm: frequency per million.

speech-bundles also need more detailed scrutiny. In the cluster list of 19C (cf. supplemental material, Appendix 1c), examples like *I do not know what*, *I do not know that* and *I do not wish to* point to Jane Austen's stylistic preference of using non-contracted forms. One obvious reason why stylistic features of Dickens figure so prominently in our data is that he is the author for which most data are included in the study. The 19C reference corpus contains only three novels by Jane Austen, for instance. Table 9 does not imply that *not the least doubt of* or *heaven forbid that I should* is only used by Dickens. Instead, it shows that these examples seem to be relatively rare and so do not occur anywhere in our data set other than in DNov. While the relative frequencies of examples in Table 9 are low, there are still similarities with the speech-bundles we discussed in the previous sections. Generally, speech-bundles are about relational language. While politeness bundles specifically reflect the sociable interaction between characters, speech-bundles have the potential to indicate confrontation, too, as the most frequent speech-bundle in DNov, *what do you mean by*, shows (cf. section 5.1). Examples (13) and (14) similarly illustrate how *will you hold your tongue* functions as a type of phrasal impoliteness.

(13) 'You are a little chafed, but I can make allowance for that, and am, fortunately, myself in the very best of tempers. Now, let us see how circumstances stand. A day or two ago, I mentioned to you, my dear fellow, that I thought I had discovered--'

'**Will you hold your tongue**?' said Jonas, looking fiercely round, and glancing at the door. (*Martin Chuzzlewit*)

(14) '**Will you hold your tongue**--female?' said Mr Mortimer Knag, plunging violently into this dialogue. (*Nicholas Nickleby*)

Speech-bundles that mark authorial style are still not entirely the same as 'labels', that is, phrases that are used to individualise specific fictional characters (Mahlberg, 2007, 2013) such as Mr Snagsby's favourite phrase *not to put too fine a point upon it* in *Bleak House*. Phrases with a strikingly characterising function tend to receive more attention from readers and literary critics. Although they are less relevant to a description of fictional speech in general, they are crucial stylistic techniques authors use to individualise fictional characters. Hence, these bundles are located at the far end of the continuum of real and fictional speech.

5.4. Speech-bundles in comparison with narrative bundles

In the previous three sections, we compared different types of speech-bundles and described them in relation to real spoken language. This comparison was enabled through separating the corpora into 'quotes' and 'non-quotes' subsets. Table 2 above showed that frequency counts for clusters vary across these subsets. A fundamental observation in corpus linguistics is that frequencies relate to meanings and functions (see Mahlberg, 2005), which becomes particularly apparent through comparison. This section contextualises and further validates the earlier findings through a text-internal key comparison. Table 10 contains the top 15 key clusters for a text-internal comparison of all fiction corpora together. The comparison of quotes versus non-quotes shows the relational language we discussed previously: bundles with first- and second-person pronouns,

Table 10. Key comparison of DNov + 19C + ChiLit quotes versus non-quotes and non-quotes versus quotes based on CLiC 2.0.1 (further ranks, frequencies and LL values are available in the supplemental material, Appendix 2).

Rank	Quotes versus non-quotes	Non-quotes versus quotes
1	what do you think of	up and down the room
2	what do you mean by	as if he had been
3	I should like to know	his hands in his pockets
4	I should like to see	in the direction of the
5	I want to speak to	at the door of the
6	what is to be done	said in a low voice
7	I am not going to	in the midst of the
8	I beg your pardon sir	in the centre of the
9	you don't mean to say	was in a state of
10	I don't know what you	with his hands in his
11	very much obliged to you	as if she had been
12	I am very glad to	as soon as he had
13	I am sure you will	by the side of the
14	to tell you the truth	in a corner of the
15	what do you say to	with the air of a

DNov: Dickens' Novels; 19C: 19th Century Reference Corpus; ChiLit: 19th Century Children's Literature; LL: log-likelihood.

question fragments and politeness expressions. On the other hand, the comparison of non-quotes versus quotes shows narrative bundles that refer to fictional characters in the third person (*as if he had been*), descriptions of body language (*his hands in his pockets*), manner of speaking rather than actual speech (*said in a low voice*), more general descriptions (*with the air of a*) and common references to time and place (*up and down the room, in a corner of the*, *at the door of the*). By focusing on bundles in speech alone, we gain an improved picture of fictional speech that specifically highlights pragmatic functions in the speaker-listener world of the characters in the story. With the comparison of quotes and non-quotes, we additionally demonstrate that meanings in fictional speech are markedly different from narrative meanings. The comparison provides further formal evidence for the functions we described in the previous sections, adding to our ability to identify matters of apparent narratorial control and character-autonomy.

6. Conclusion

We have proposed a novel approach for the description of speech in fiction that explores the relationship between fictional speech and real spoken language. This was motivated by a desire to move beyond the narratorial framing of talk in order to focus on the content of direct speech itself. Through a number of comparisons and contrasts, we found fictional speech-bundles that are shared between fictional speech and spoken language, common across fiction corpora or specific to individual authors or even texts. We thus illustrated a continuum of speech-bundles and the boundaries between fiction and real spoken language. Hence, speech-bundles serve to reaffirm the continuity between real and fictional people that the concept of mind-modelling assumes. Linguistic features that fictional and real people share contribute a sense of naturalness to speech in fiction. So speech-bundles can create a strong impression of authenticity.

Speech-bundles are examples of relational language that contribute to the depiction of social and interpersonal relationships that are also relevant to mind-modelling. The continuity between fictional and non-fictional usage explains the efficacy of readers' schematic knowledge from the general language system as applied to literary reading. Typical features such as question fragments, politeness and vagueness draw on norms of spoken language and indicate literary conventions. In our data, the range of speech-bundles in fiction is more limited than the range in spoken language, so it is possible that literature confers an iconic or heightened meaningful effect onto these forms. In particular, shorter lexical bundles (e.g. *I mean*, *you know* and *I don't know*) are less frequent in fiction. In addition, the comparison of 19th- with 20th-century data has suggested some historical change in the norms of naturalness which are worth exploring further. With regard to Dickens specifically, the occurrence of relational language is so far underexplored evidence for the often claimed effect that Dickens' writing had on his contemporary readers. The relational language in the novels can be seen as complementing the relationship between author and readers that Dickens also furthered through his public readings.

Neither authors nor readers are necessarily aware of lexical bundles – which might explain the lack of attention they have received – and corpus methods have enabled us to shed fresh light on this aspect of speech in fiction. In fact, our corpus linguistic

methodology was vital for the identification of fictional speech-bundles. Precisely because of the naturalness they contribute to fiction, speech-bundles are difficult to spot by reading alone. It was crucial to treat the quotes subsets separately as a corpus in its own right and in contrast to the non-quotes. When fictional texts are not separated in this way, features of speech are levelled out to a certain extent, and their value in exploring characterisation is lost. With the help of our freely available CLiC tool and corpora, we have opened up this area of research more widely. For corpus linguistics, the availability of tools and methods has a significant impact on the direction that the field takes. At the same time, the study of fictional speech at scale will allow for connections to be made with work in the digital humanities that uses large sets of novels to study literary history (see Underwood, 2019).

Beyond the contribution to research into fictional language, the way we approached the continuum of speech emphasises the strengths of corpus linguistics in dealing with fuzzy categories. Unlike most studies of lexical bundles, we approached frequency cut-offs in a less selective way. More important than selecting specifically characteristic bundles was the range of comparisons to show fluid boundaries between data sets.

In this article, we have focused on the comparison of corpora. But our work is relevant to the analysis of individual texts too. A sequence of words that is a lexical bundle in a spoken language corpus might only occur once in a fictional text, and so in this text it would not be recognised as a speech-bundle. However, it might still contribute to the naturalness of the text precisely by virtue of its occurrence in real spoken language (for an initial small-scale example, see Mahlberg and Wiegand, 2018). A list of lexical bundles based on a corpus of spoken language can be used to annotate a fictional text for all the sequences that are lexical bundles and in this way contribute to the analysis of an individual text. Such annotation for features of spokenness also has significant potential for the automatic identification of free indirect discourse, which so far is limited (see also the discussion in Toolan, 2009). We hope that the proposal made here, the methodology and CLiC resources, will be used to further the conceptualisation of fictional speech and to develop a new range of corpus approaches to fiction.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Arts and Humanities Research Council (grant number AH/P504634/1).

Supplemental material

Supplemental material for this article is available online.

Note

1. Log-likelihood (LL) values for the comparison of the Spoken British National Corpus 1994 versus the fictional corpora (+ indicates overuse and – indicates underuse relative to the respective fiction corpus):

   I mean: versus DNov + LL = 2634.97; versus 19C + LL = 3555.62; versus ChiLit + LL = 3565.87
   I think: versus DNov + LL = 1481.86; versus 19C + LL = 1701.06; versus ChiLit + LL = 1668.91
   You know: versus DNov + LL = 1117.88; versus 19C + LL = 2587.02; versus ChiLit + LL = 1712.89
   I don't know: versus DNov + LL = 28.15; versus 19C + LL = 346.88; versus ChiLit + LL = 127.26
   I don't know what I: versus DNov – LL = 24.73; versus 19C – LL = 2.68; versus ChiLit – LL = 0.08

References

Aijmer K (1996) *Conversational Routines in English: Convention and Creativity*. London: Longman.

Arnon I and Snider N (2010) More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language* 62(1): 67–82.

Axelsson K (2009) Research on fiction dialogue: Problems and possible solutions. In: Jucker AH, Schreier D and Hundt M (eds) *Corpora: Pragmatics and Discourse*. Amsterdam: Rodopi, pp. 189–201.

Bednarek M (2018) *Language and Television Series: A Linguistic Approach to TV Dialogue*. Cambridge: Cambridge University Press.

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

Biber D, Johansson S, Leech G, et al. (1999) *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Brown P and Levinson SC (1987) *Politeness: Some Universals in Language Usage*. Cambridge: Cambridge University Press.

Busse B (2010) *Speech, Writing and Thought Presentation in a Corpus of 19th-century Narrative Fiction*. Bern: University of Bern.

Carter R (2004) *Language and Creativity: The Art of Common Talk*. London: Routledge.

Carter R and McCarthy M (2004) This that and the other: Multi-word clusters in spoken English as visible patterns of interaction. *Teagna* 21: 30–52.

Carter R and McCarthy M (2006) *Cambridge Grammar of English: A Comprehensive Guide to Spoken and Written English Usage*. Cambridge: Cambridge University Press.

Carter R and Nash W (1990) *Seeing through Language: A Guide to Styles of English Writing*. London: Blackwell.

Cermakova A (2018, 14 February) ChiLit: The GLARE 19th Century Children's Literature Corpus in CLiC [Blog post]. Available at: https://blog.bham.ac.uk/glareproject/2018/02/14/chilit-the-glare-19th-century-childrens-literature-corpus-in-clic/ (accessed 18 August 2019).

Chapman R (1994) *Forms of Speech in Victorian Fiction*. London: Longman.

Cohn D (1978) *Transparent Minds: Narrative Modes for Presenting Consciousness in Fiction*. Princeton, NJ: Princeton University Press.

Culpeper J (2001) *Language and Characterisation: People in Plays and Other Texts*. London: Longman.

Culpeper J (2009) Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's *Romeo and Juliet*. *International Journal of Corpus Linguistics* 14(1): 29–59.

Culpeper J and Kytö M (2010) *Early Modern English Dialogues: Spoken Interaction as Writing*. Cambridge: Cambridge University Press.
De Haan P (1996) More on the language of dialogue in fiction. *ICAME Journal* 20: 23–40.
Durrant P (2017) Lexical bundles and disciplinary variation in university students' writing: Mapping the territories. *Applied Linguistics* 38(2): 165–193.
Fielding KJ (1960) *The Speeches of Charles Dickens (1841–1870)*. Oxford: Clarendon.
Fludernik M (1993) *The Fictions of Language and the Languages of Fiction*. London: Routledge.
Genette G (1980) *Narrative Discourse: An Essay in Method*. Ithaca, NY: Cornell University Press.
Grieve J, Clarke I, Chiang E, et al. (2019) Attributing the *Bixby Letter* using n-gram tracing. *Digital Scholarship in the Humanities* 34(3): 493–512.
Herman D (2013) *Storytelling and the Sciences of Mind*. Cambridge, MA: MIT Press.
Hodson J (2014) *Dialect in Film and Literature*. Basingstoke: Palgrave MacMillan.
Hoffmann S, Evert S, Smith N, et al. (2008) *Corpus Linguistics with BNCweb: A Practical Guide*. Frankfurt: Peter Lang.
Kopaczyk J (2013) *The Legal Language of Scottish Burghs: Standardization and Lexical Bundles (1380–1560)*. Oxford: Oxford University Press.
Labov W (1972) *Language in the Inner City*. Philadelphia, PA: University of Pennsylvania Press.
Leech G (2000) Grammars of spoken English: New outcomes of corpus-oriented research. *Language Learning* 50(4): 675–724.
Leech G and Short M (1981/2007) *Style in Fiction: A Linguistic Introduction to English Fictional Prose*. 2nd edition. Harlow: Pearson.
Love R, Dembry C, Hardie A, et al. (2017) The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics* 22(3): 319–344.
McIntyre D (2016) Dialogue: Credibility versus realism in fictional speech. In: Sotirova V (ed.) *The Bloomsbury Companion to Stylistics*. London: Bloomsbury, pp. 430–443.
Mahlberg M (2005) *English General Nouns: A Corpus Theoretical Approach*. Amsterdam: John Benjamins.
Mahlberg M (2007) Clusters, key clusters and local textual functions in Dickens. *Corpora* 2(1): 1–31.
Mahlberg M (2013) *Corpus Stylistics and Dickens's Fiction*. New York, NY: Routledge.
Mahlberg M and Stockwell P (2016) Point and CLiC: Teaching literature with corpus stylistic tools. In: Burke M, Fialho O and Zyngier S (eds) *Scientific Approaches to Literature in Learning Environments*. Amsterdam: John Benjamins, pp. 251–267.
Mahlberg M and Wiegand V (2018) Corpus stylistics, norms and comparisons: Studying speech in *Great Expectations*. In: Page R, Busse B and Nørgaard N (eds) *Rethinking Language, Text and Context: Interdisciplinary Research in Stylistics in Honour of Michael Toolan*. London: Routledge, pp. 123–143.
Mahlberg M and Wiegand V (in press) Literary stylistics. In: Adolphs S and Knight D (eds) *Routledge Handbook of English Language and Digital Humanities*. London: Routledge.
Mahlberg M, Stockwell P, de Joode J, et al. (2016) CLiC Dickens: Novel uses of concordances for the integration of corpus stylistics and cognitive poetics. *Corpora* 11(3): 433–463.
Mahlberg M, Wiegand V, Lentin J, et al. (2019) CLiC User Guide v2.0.1 documentation. Available at: http://clic.bham.ac.uk/docs/ (accessed 18 August 2019).
O'Keeffe A, McCarthy M and Carter R (2007) *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press.
Oostdijk N (1990) The language of dialogue in fiction. *Literary and Linguistic Computing* 5(3): 235–241.
Page N (1988) *Speech in the English Novel*. 2nd edition. Houndmills: MacMillan.
Palmer A (2011) Social minds in fiction and criticism. *Style* 45(2): 196–240.

Quirk R, Greenbaum S, Leech G, et al. (1985) *A Comprehensive Grammar of the English Language*. London: Longman.
Rayson P (2016) Log-likelihood and effect size calculator spreadsheet. Available at: http://ucrel.lancs.ac.uk/llwizard.html (accessed 18 August 2019).
R Core Team (2018) *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Available at: http://www.R-project.org/ (accessed 18 August 2019).
Scott M (2019) Single words v. clusters. Available at: https://lexically.net/downloads/version7/HTML/single_words.html (accessed 18 August 2019).
Semino E and Short M (2004) *Corpus Stylistics: Speech, Writing and Thought Presentation in a Corpus of English Writing*. London: Routledge.
Short M (1996) *Exploring the Language of Poems, Plays, and Prose*. London: Longman.
Short M, Semino E and Wynne M (2001) Revisiting the notion of faithfulness in discourse report/(re)presentation using a corpus approach. In: Biermann I and Combrink (eds) *A Poetics Linguistics and History*. Potchefstroom: Potchefstroom University, pp. 484–509.
Sinclair J (1991) *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
Sternberg M (1982) Proteus in quotation-land: Mimesis and the forms of reported discourse. *Poetics Today* 3(2): 107–156.
Stockwell P (2009) *Texture: A Cognitive Aesthetics of Reading*. Edinburgh: Edinburgh University Press.
Stockwell P and Mahlberg M (2015) Mind-modelling with corpus stylistics in *David Copperfield*. *Language and Literature* 24(2): 129–147.
Tannen D (1989) *Talking Voices: Repetition, Dialogue and Imagery in Conversational Discourse*. Cambridge: Cambridge University Press.
Thomas B (2012) *Fictional Dialogue: Speech and Conversation in the Modern and Postmodern Novel*. Lincoln, NE: University of Nebraska Press.
Toolan M (1985) Analysing fictional dialogue. *Language & Communication* 5(3): 193–206.
Toolan M (2009) *Narrative Progression in the Short Story: A Corpus Stylistic Approach*. Amsterdam: John Benjamins.
Underwood T (2019) *Distant Horizons: Digital Evidence and Literary Change*. Chicago, IL: University of Chicago Press.
Wray A (2002) *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.
Zunshine L (2006) *Why We Read Fiction: Theory of Mind and the Novel*. Columbus, OH: Ohio State University Press.

Author biographies

Michaela Mahlberg is Professor of Corpus Linguistics at the University of Birmingham, where she is also the Director of the Centre for Corpus Research. She is the Editor of the *International Journal of Corpus Linguistics* (John Benjamins) and the Principal Investigator on the CLiC Dickens project.

Viola Wiegand is a Research Fellow on the CLiC Dickens project at the University of Birmingham. Her main research interest is the study of textual patterns with tools and frameworks from corpus linguistics, stylistics and discourse analysis. She is Assistant Editor of the *International Journal of Corpus Linguistics* (John Benjamins).

Peter Stockwell is Professor of Literary Linguistics at the University of Nottingham, UK, and a Fellow of the English Association. He has published 12 books and over 80 articles in stylistics, sociolinguistics and applied linguistics.

Anthony Hennessey is a visiting academic at the University of Birmingham, with extensive programming experience and a PhD in statistics from the University of Nottingham. He is currently a Senior Fellow at CERN.