Corpus stylistics, norms and comparisons

Mahlberg, Michaela; Wiegand, Viola

Final draft accepted for publication (may contain minor errors and infelicities). Please cite the published version: Mahlberg, M., & Wiegand, V. (2018). Corpus stylistics, norms and comparisons: Studying speech in *Great Expectations*. In R. Page, B. Busse, & N. Nørgaard (Eds.), *Rethinking Language, Text and Context: Interdisciplinary Research in Stylistics in Honour of Michael Toolan* (pp. 123–143). London: Routledge.

Corpus Stylistics, Norms and Comparisons

Studying Speech in *Great Expectations*

Michaela Mahlberg and Viola Wiegand

1. Introduction

Literary stylistics is often described as a discipline that studies literary texts by drawing on a linguistic toolkit. Or, as Simpson (2004, p. 2) describes it, stylistics is 'a method of textual interpretation in which primacy of place is assigned to language'. Because of the centrality of language, stylisticians can be seen to inherit theories 'from the particular kind of linguistics (systemic-functional, corpus, cognitive, etc.) they chiefly employ' (Toolan, 2014, p. 13). While literary stylistics thus brings the study of language and literature together, this does not automatically mean language and literature are seen as fully integrated. Leech and Short (2007, p. 12) refer to Spitzer's philological circle to explain fundamental tenets of literary stylistics. According to Spitzer (1948) a literary text can be seen both as a work of art and as a sample of language, so that an analysis can start from the literary effects of text and study the language to explain these effects or equally it can begin with the study of the language and then seek to explain the literary effects that the language might create. Ideally, a stylistic analysis would then proceed in a cyclical fashion moving between the linguistic and literary view. Any approach, however, that aims to emphasize the integration of the study of language and literature necessarily also stresses that language and literature are fundamentally different and so non-literary and literary texts need different treatment. This difference is sometimes also expressed through value judgements. Stockwell and Whiteley (2014, p. 1) go as far as to say that stylistics is 'the proper study of literature', while Toolan (2014) points out that literary texts operate like any other

Final draft accepted for publication (may contain minor errors and infelicities). Please cite the published version: Mahlberg, M., & Wiegand, V. (2018). Corpus stylistics, norms and comparisons: Studying speech in *Great Expectations*. In R. Page, B. Busse, & N. Nørgaard (Eds.), *Rethinking Language, Text and Context: Interdisciplinary Research in Stylistics in Honour of Michael Toolan* (pp. 123–143). London: Routledge.

type of language, but are more intellectually interesting. Carter (2016) approaches the distinction through the notion of "literariness". He argues that literariness is best seen as a cline, so that it is possible to see literary features to varying degrees in all types of texts. The way in which early dictionaries used examples from literary texts is another more practical illustration that the boundaries are fuzzier than the terminology around literary and non-literary language suggests.

In this chapter, we look at fictional speech in Dickens's *Great Expectations* to lend further support to the argument that the notion of a clear-cut distinction between literary and non-literary language is 'an unhelpful one', as Carter (2016, p. 69) puts it. In particular, we aim to demonstrate that literary stylistics cannot just rely on linguistic models and methods but also needs to push the boundaries of the field by stressing how the analysis of the language of literature also impacts on how we describe language more generally. Toolan (2014, p. 13) lists corpus linguistics among the examples of linguistic approaches from which stylistics has inherited theories. Leech and Short (2007, p. 286) even talk about a "corpus turn" in stylistics and the term "corpus stylistics" is used to position work in this field. In this chapter, we want to argue that, while existing corpus linguistic methods and emerging tools for textual analysis within the digital humanities offer useful potential for stylistic analysis, there is still a need for more approaches that are tailored towards the textual analysis of literature. We will introduce the CLiC web app that has been specifically designed for the study of literary texts and demonstrate how adjustments to conventional corpus tools make it possible to approach literary questions in a more focused way. We will propose an innovative method to identify linguistic patterns that are associated with "spokenness" in fiction and suggest that corpus methods can also help to identify likely candidates for free indirect speech.

Final draft accepted for publication (may contain minor errors and infelicities). Please cite the published version: Mahlberg, M., & Wiegand, V. (2018). Corpus stylistics, norms and comparisons: Studying speech in *Great Expectations*. In R. Page, B. Busse, & N. Nørgaard (Eds.), *Rethinking Language, Text and Context: Interdisciplinary Research in Stylistics in Honour of Michael Toolan* (pp. 123–143). London: Routledge.

2. Corpus Stylistics Methods and Theory

Many corpus stylistic studies have applied standard corpus methods to literary texts. A popular corpus linguistic approach to literary texts is the analysis of keywords. Keywords are words that are statistically significantly more frequent in a specific text compared to a reference corpus. For example, Scott (2006) examines keywords of *Romeo and Juliet* retrieved in comparison with all Shakespeare plays. The study illustrates the keyword procedure for examining the "aboutness" of an individual text by identifying thematic keywords (*love*, *lips*, *light*, *death*, *poison*, etc.). Unsurprisingly, Scott (2006) also finds proper nouns (*Romeo*, *Juliet*, *Mercutio*, *Verona*, etc.) and exclamations (*O*, *Ah*) among the keywords. He points out that unexpected items like exclamations should be investigated further in a concordance analysis; indeed, that is also a popular procedure for corpus stylistics (see e.g. Fischer-Starcke, 2010). Scott's interest is not primarily a literary stylistic one. He mainly uses the Shakespeare examples to illustrate the concept of keyness. Still, the corpus methodological approach he takes is similar to studies in corpus stylistics.

Stylistic concerns especially come into play when decisions are made on the type of comparisons. Culpeper (2009) illustrates keyword (as well as key part-of-speech and semantic categories) comparisons between subcorpora that consists of the speech by individual characters in *Romeo and Juliet* which allows him to draw conclusions on characterisation. Because of the text format of plays, subcorpora for character speech are straightforward to select. Murphy (2015) also runs keywords for Shakespeare, but with the focus on the language of soliloquies in Shakespeare's plays. He creates specific subcorpora based on existing typologies of soliloquy and dialogue, drawing on criteria of direction of address. These corpora, make it possible for Murphy (2015) to compare the language of soliloquies with dialogue.

Final draft accepted for publication (may contain minor errors and infelicities). Please cite the published version: Mahlberg, M., & Wiegand, V. (2018). Corpus stylistics, norms and comparisons: Studying speech in *Great Expectations*. In R. Page, B. Busse, & N. Nørgaard (Eds.), *Rethinking Language, Text and Context: Interdisciplinary Research in Stylistics in Honour of Michael Toolan* (pp. 123–143). London: Routledge.

The approach of "text-internal" keyword comparisons has also been used in studies of narrative text. Toolan (2006) draws on the keywords procedure in his study of narrative progression in short stories. As Toolan's general framework is sentence-based (also see Toolan, 2009; 2016), instead of carrying out a conventional concordance-analysis of the keywords, he pulls out all full sentences of a particular keyword. Aiming at tracing progression and coherence in a short story, Toolan (2006) argues that this set of keyword sentences acts as an "abridgement" of the story.

In addition to the keywords procedure, other comparative methods from corpus linguistics include "multi-dimensional analysis" (MD). MD analysis was developed by Biber (1988) for the comparison of spoken and written registers. It works with corpora that have been tagged for a selection of mainly grammatical features. These features are then quantified and assigned to five dimensions. Shepherd and Berber Sardinha (2013) illustrate the use of MD analysis to compare works by the writer Julian Barnes to a number of registers such as telephone conversations, face-to-face conversations, professional letters, and adventure fiction.

Generally, for corpus linguistics, any type of analysis will involve comparison. The notion of comparison is also important to capture theoretical links between corpus linguistics and literary stylistics by relating comparison to the notion of "foregrounding". Foregrounding means that linguistic features are made prominent and stand out against the norms of general language or against the background of their textual context. It is the psychological effect brought about in the reader (hearer) when deviant features of a text are made perceptually prominent (cf. Leech, 1985, p. 47; Short, 1996, p. 11). The two main textual patterns that can account for linguistic means to achieve foregrounding are "deviation" and "repetition". Deviations from linguistic norms are, for instance, ungrammatical forms or uncommon semantic combinations.

Final draft accepted for publication (may contain minor errors and infelicities). Please cite the published version: Mahlberg, M., & Wiegand, V. (2018). Corpus stylistics, norms and comparisons: Studying speech in *Great Expectations*. In R. Page, B. Busse, & N. Nørgaard (Eds.), *Rethinking Language, Text and Context: Interdisciplinary Research in Stylistics in Honour of Michael Toolan* (pp. 123–143). London: Routledge.

Repetition is also a form of deviation in that it goes against normal usage patterns by being overfrequent (cf. Wales, 2001, p. 157). Deviant textual patterns can theoretically be described as the results of comparisons against various norms. From a corpus linguistic point of view, deviation can be practically identified through various forms of corpus comparison. Leech (1985) distinguishes three types of deviation: "primary deviation" is deviation from norms of the language as a whole (Leech, 1985, p. 45), "secondary deviation" is deviation from norms of literary composition, including norms of author or genre (Leech, 1985, p. 48), "tertiary deviation" (also called "internal deviation") is deviation from norms internal to a text (Leech, 1985, p. 49).

Louw (1993) is now a classic example of how corpus methods can aid the identification of primary deviation. He refers to his method as 'matching texts against corpora' (Louw, 1993, p. 161). This means individual words or phrases that are identified in a specific text passage are compared to a general reference corpus. With the help of a concordance analysis in the general corpus, Louw (1993) shows how the meaning in the text passage can be described as unusual and thus creating literary effects. An example of secondary deviation is Scott's (2006) comparison of *Romeo and Juliet* to all of Shakespeare's plays. Whereas Culpeper's (2009) comparison of individual characters' speeches can be seen as describing a form of internal deviation. While the framework by Leech (1985) is useful to systematically relate the notion of deviation to corpus comparison, corpus linguistics emphasises that the notion of norms can only ever present a simplified view. Focusing on parameters for comparison and the description of variation might be more productive (cf. Mahlberg, 2013 for a more detailed discussion).

Shifting the emphasis from deviation to comparison, concordances, as a standard corpus tool, also provide crucial methods for corpus stylistic research.

Final draft accepted for publication (may contain minor errors and infelicities). Please cite the published version: Mahlberg, M., & Wiegand, V. (2018). Corpus stylistics, norms and comparisons: Studying speech in *Great Expectations*. In R. Page, B. Busse, & N. Nørgaard (Eds.), *Rethinking Language, Text and Context: Interdisciplinary Research in Stylistics in Honour of Michael Toolan* (pp. 123–143). London: Routledge.

Running concordance searches across an individual text will retrieve all the instances of a word / words or phrase in that text and thus provide the opportunity for identifying the cumulative picture of how specific textual meanings are created. Ruano San Segundo's (2017) study of reporting verbs in *Nicholas Nickleby* is one example of the potential of concordance searches.

Common corpus tools such as WordSmith (Scott, 2017), AntConc (Anthony, 2018), or WMatrix (Rayson, 2008) are popular choices for corpus stylisticians. Reasons for this popularity include their availability and accessibility. A key word comparison with any of these three tools will be easier to run than, for instance, an MD analysis not only because of the statistical complexity but also because the Biber tagger that is used for the initial tagging is less readily available[1]. In addition to standard corpus tools, which typically include concordance, keyword, cluster and collocation functions, a handful of new tools have been developed that are particularly suited to support stylistic analyses. Although in principle built for the analysis of any type of text, the WordWanderer web app (http://wordwanderer.org/) (Dörk and Knight, 2015) currently features exclusively literary texts as preloaded examples (but allows the user to upload their own text). This tool was developed to provide both novices and experts with the opportunity to explore a text via a 'playful' navigation through the lexical links in a text (Dörk and Knight, 2015, p. 84). By contrast, the WorldBuilder tool (http://viv-research.info/TWT/system/index.html; Wang et al., 2016) works with conceptual rather than lexical patterns. Having been developed to facilitate the annotation of a text

[1] Andrea Nini developed The Multidimensional Analysis Tagger (MAT; Nini, 2014) that replicates Biber's (1988) tagger. However, it is also important that the underlying selection of features can be modified to fit the set of texts under analysis (cf. Biber 2006, pp. 181f.).

Final draft accepted for publication (may contain minor errors and infelicities). Please cite the published version: Mahlberg, M., & Wiegand, V. (2018). Corpus stylistics, norms and comparisons: Studying speech in *Great Expectations*. In R. Page, B. Busse, & N. Nørgaard (Eds.), *Rethinking Language, Text and Context: Interdisciplinary Research in Stylistics in Honour of Michael Toolan* (pp. 123–143). London: Routledge.

with the elements of the cognitive stylistic framework of Text World Theory (Gavins, 2007; Werth, 1999), the tool makes it possible to quantify these categories and produce "cognitive diagrams" based on the text worlds.

Work in corpus stylistics also needs to be seen in the context of wider developments in computational linguistics and digital humanities. Secondary deviation as referred to by Leech (1985) is especially relevant to studies in stylometry. Hoover (2007), for instance, studies the distinctive features of Henry James's style and specifically the division into early and late James. Burrows and Craig (2012, p. 293) discuss wider issues in authorship attribution studies to 'show the literary fundamentals of the relations between character styles and authorial styles'. Drawing on examples from Shakespeare and his contemporaries Burrows and Craig (2012) show that while character idiolects are identifiable, authorial differentiation transcends variation in characters' speech styles. Eder (2017, p. 51) argues that methods of stylometry are popular with literary scholars, 'because they offer convincing visualizations'. Eder (2017, p. 51) also refers to the 'immense popularity of beautiful yet relatively simple plots' of research concerned for instance with literary history or "distant reading" and "macro analysis" (Moretti, 2005, Jockers, 2013). Other terms that have come to be used for work of a similar nature are "cultural analytics" or "culturomics". What tends to distinguish such studies from research in corpus stylistics is the amount of data under analysis and the emphasis on larger trends. But the results of such studies provide an important context to assess the norms, in the sense of Leech (1985), which impact on the literary stylistic analysis of individual texts or even text extracts. Underwood et al. (2018), for instance, investigate how the language used to describe fictional men and women has changed since the 18th century, which can provide a useful reference point for the study of gendered language in an individual novel.

Final draft accepted for publication (may contain minor errors and infelicities). Please cite the published version: Mahlberg, M., & Wiegand, V. (2018). Corpus stylistics, norms and comparisons: Studying speech in *Great Expectations*. In R. Page, B. Busse, & N. Nørgaard (Eds.), *Rethinking Language, Text and Context: Interdisciplinary Research in Stylistics in Honour of Michael Toolan* (pp. 123–143). London: Routledge.

3. A Tool for Corpus Stylistics: The CLiC Web App

Corpus linguistics has not yet used the range of data visualisation tools that other areas of digital humanities might have. However, concordances are fundamentally a form of linguistic visualisation. Concordances enable the researcher to see patterns and identify meanings associated with these patterns. This approach to meaning is central to the innovative contribution that corpus linguistics has made to modern linguistics. Sinclair (1991, p. 100) already makes the point: '[t]he language looks rather different when you look at a lot of it at once'. To identify meanings in literary texts, and be able to focus on individual texts and even text extracts in the way that literary stylistics does, concordances and related tools also play an important role. In the present chapter, we propose an approach that adjusts standard corpus methods and tools so that they best serve the exploration of literary texts.

The CLiC (Corpus Linguistics in Context) web app (http://clic.bham.ac.uk; Mahlberg et al., 2016) has been specifically designed for this purpose. It offers standard corpus functionalities (concordancing, generating clusters and keywords) but also additional options to aid the analysis of literary features. These provide the user with access to different textual subsets (e.g. character speech and narration) along with the possibility to "KWICGroup" concordance lines based on shared lexical patterns. CLiC also contains a tag menu that allows researchers to "tag" concordance lines as part of their analysis and there an option for different users to merge their results to facilitate the measurement of inter-rater agreement.

The basic principle behind the CLiC interface is designed around "subsets" or intratextual subcorpora. Accordingly, apart from accessing "all text" of a particular book, it is possible to navigate to "quotes" (mostly corresponding to character speech) and "non-quotes" (narration). A specific type of non-quotes are "suspensions". These

Final draft accepted for publication (may contain minor errors and infelicities). Please cite the published version: Mahlberg, M., & Wiegand, V. (2018). Corpus stylistics, norms and comparisons: Studying speech in *Great Expectations*. In R. Page, B. Busse, & N. Nørgaard (Eds.), *Rethinking Language, Text and Context: Interdisciplinary Research in Stylistics in Honour of Michael Toolan* (pp. 123–143). London: Routledge.

are narratorial disruptions of character speech, based on Lambert's (1981) concept of the "suspended quotation", which tend to contain reporting clauses and/or additional character information such as body language. In CLiC, suspensions are categorised according to length into "short suspensions" containing up to four words and "long suspensions" for any stretches longer than that. As Examples (1) and (2) – both from Chapter 2 of *Great Expectations* – illustrate, the longer suspensions provide the narrator with more opportunity to give additional character information such as body language.[2]

(1) "Yes, Pip," said Joe; "and what's worse, she's got Tickler with her." (GE, Chapter 2)

(2) "You know, Pip," said Joe, solemnly, with his last bite in his cheek and speaking in a confidential voice, as if we two were quite alone, "you and me is always friends, and I'd be the last to tell upon you, any time. But such a--" he moved his chair and looked about the floor between us, and then again at me - "such a most oncommon Bolt as that!" (GE, Chapter 2)

While CLiC makes it possible to read the novels with this annotation in the Chapter view, as illustrated for Example (2) in Figure 1, the main advantage of the corpus stylistic features of the CLiC web app is that the subsets are searchable independently of each other.

[2] All examples are taken from the CLiC text, so no page numbers are included.

Final draft accepted for publication (may contain minor errors and infelicities). Please cite the published version: Mahlberg, M., & Wiegand, V. (2018). Corpus stylistics, norms and comparisons: Studying speech in *Great Expectations*. In R. Page, B. Busse, & N. Nørgaard (Eds.), *Rethinking Language, Text and Context: Interdisciplinary Research in Stylistics in Honour of Michael Toolan* (pp. 123–143). London: Routledge.

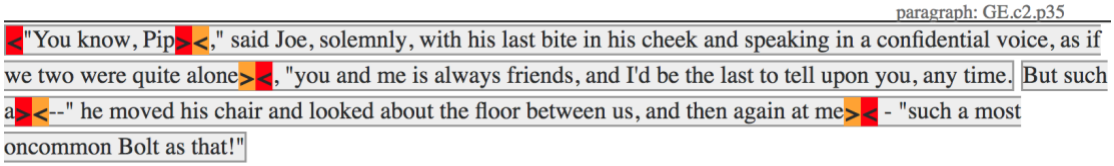

Figure 1: Screenshot of the CLiC chapter view for Example (2) from *Great Expectations* (quotes annotated in red; long suspensions annotated in orange)

The searchable subsets make it possible to focus on the language used in quotes, non-quotes and suspensions. A simple option is, for instance, to check the distribution of a word or phrase that is found when reading a small text extract. Example (2) contains the phrase *you know*. Checking quotes and non-quotes shows that *you know* occurs 106 times in quotes and once in non-quotes of *Great Expectations*. In Example (2), it is used as a discourse marker. Going through the concordance lines, the tagging function allows the researcher to highlight all occurrences where *you know* functions as a discourse marker by adding a tick to the line (see Concordance 1). This makes it easy to count the examples, but also sort the concordance so that examples with a tick are displayed together.

4	pless bite, and looked at me again	."You know,	Pip," said Joe, solemnly, with his la	GE	2	35	84	✓
5	his food, has he?" cried my sister	."You know,	old chap," said Joe, looking at me,	GE	2	37	87	✓
6	oat had revived."Dressed like you,	you know,	only with a hat," I explained, trembl	GE	3	38	85	✓
7	ear about it, before it's done with,	you know.""	I know, but this is another pint, a s	GE	5	62	208	✓
8	moment when you came in.Don't	you know,	Pip?""So," said my convict, turning	GE	5	70	220	
9	g, Pip, or the pot won't bile, don't	you know?"	I saw that, and said so."'Consequen	GE	7	32	73	
10	pointed!Good indeed!Now Joseph,	you know	the case.""No, Joseph," said my sis	GE	7	86	196	
11	prehended in the answer "No.""Do	you know	what I touch here?" she said, laying	GE	8	42	108	
12	. Joe, "I wish you had him always:	you know	so well how to deal with him.""Now	GE	9	18	42	

Concordance 1: Sample of the 106 occurrences of *you know* in quotes – the tick at the end of the concordance lines indicates occurrences of the discourse marker

Final draft accepted for publication (may contain minor errors and infelicities). Please cite the published version: Mahlberg, M., & Wiegand, V. (2018). Corpus stylistics, norms and comparisons: Studying speech in *Great Expectations*. In R. Page, B. Busse, & N. Nørgaard (Eds.), *Rethinking Language, Text and Context: Interdisciplinary Research in Stylistics in Honour of Michael Toolan* (pp. 123–143). London: Routledge.

If a word or phrase occurs predominantly in quotes, the occurrences in non-quotes can point to specific textual functions. The only example of *you know* in non-quotes is shown in context in Example (3). Here the narrator Pip uses the discourse markers to pick up on Herbert's wording in the preceding sentence.

(3) "But the thing is," said Herbert Pocket, "that you look about you. That's the grand thing. You are in a counting-house, you know, and you look about you."

> It struck me as a singular implication that you couldn't be out of a counting-house, **you know**, and look about you; but I silently deferred to his experience. (GE, Chapter 22)

4. Common and Idiosyncratic Speech Patterns in *Great Expectations*

The corpus interpretation of the three types of deviation outlined by Leech (1985) has already pointed towards how corpus comparison can contribute to literary stylistic concerns. In this section, we are particularly interested in how corpus methods can support the analysis of fictional speech in a single novel. Page (1988) has argued that for the description of fictional speech the main question is not how similar it is to real spoken language. Because of the challenges that lie in representing spoken language in written form a detailed comparison would be an unproductive exercise. He also raises the point that the conceptualisation of fictional speech will depend on the current model of real spoken language that he still sees as inadequate. However, especially because of corpus advances, today's current model of real spoken language is significantly different from the one that Page (1988) was referring to and our corpus approach takes account of these developments.

Final draft accepted for publication (may contain minor errors and infelicities). Please cite the published version: Mahlberg, M., & Wiegand, V. (2018). Corpus stylistics, norms and comparisons: Studying speech in *Great Expectations*. In R. Page, B. Busse, & N. Nørgaard (Eds.), *Rethinking Language, Text and Context: Interdisciplinary Research in Stylistics in Honour of Michael Toolan* (pp. 123–143). London: Routledge.

In literary stylistics, the analysis of fictional speech has drawn on various linguistic models and approaches, including pragmatic principles, conversation or discourse analysis. The speech and thought presentation model by Leech and Short (1981) has been an extremely influential approach in literary stylistics not least because it directly accounts for the specific forms in which speech is presented in literature. An updated model also includes writing presentation (Leech and Short 2007), and the large-scale corpus study by Semino and Short (2004) has provided an empirical basis that has shown the relevance of the model beyond literary texts. Busse (2010) has specifically developed the speech, writing and thought presentation categories based on a corpus of 19th century fiction and suggested repetitive linguistic features for their automatic annotation.

From a theoretical point of view, direct speech generally seems to be the most straightforward category as it is formally indicated by quotation marks. In CLiC the identification of quotes is also entirely based on punctuation. Direct thought and direct writing appear to be less frequent in the CLiC corpora so, as a short-hand, we refer to all quotes as fictional speech to start with. As part of the speech, writing and thought presentation model, direct speech tends to be seen in relation to the other categories so that the actual patterns of the content of what is said receive less interest than the structures that define what category the speech presentation falls into. In corpus stylistics, patterns of what is said are mainly studied for drama (as in the above example of Shakespeare) or television dialogue (e.g. Bednarek 2011), as in both cases direct speech comes in a more straightforward format than in narrative fiction. CLiC, however, makes it easy to focus on patterns in speech.

Beginning from the example of *Great Expectation* (GE), in this section we will illustrate how CLiC can help identify patterns in direct speech that appear particularly

Final draft accepted for publication (may contain minor errors and infelicities). Please cite the published version: Mahlberg, M., & Wiegand, V. (2018). Corpus stylistics, norms and comparisons: Studying speech in *Great Expectations*. In R. Page, B. Busse, & N. Nørgaard (Eds.), *Rethinking Language, Text and Context: Interdisciplinary Research in Stylistics in Honour of Michael Toolan* (pp. 123–143). London: Routledge.

"speech-like". We use "clusters", i.e. repeated sequences of words such as *I don't want to* to look at three types of examples:

1) General speech clusters in GE quotes;
2) Idiosyncratic character clusters in GE;
3) Clusters for which the tendency to occur in quotes affects their function when occurring in non-quotes.

GE forms a relatively small corpus of around 185,000 words and we mostly focus on the even smaller "quotes" subset of the corpus of approximately 53,000 words. We compare GE to the CLiC corpus of Dickens's novels (DNov) and to the spoken part of the British National Corpus (BNC). There is no general reference corpus of transcribed spoken 19th century English that we could have used. More importantly, however, we use the BNC spoken as a proxy for linguistic background knowledge of a the 20th century reader. This knowledge is relevant to the extent to which examples from GE can create an effect of spokenness in the reader. Table 1 provides an overview of the corpora used in this chapter.

Table 1: Corpora used in this study

Corpus	**Corpus size**	**Speech subset used**
BNC	97,639,023	9,899,403
Dickens's Novels (DNov)	3,835,807	1,369,029
Great Expectations (GE)	185,199	53,221

1) General speech clusters in GE quotes

This section examines which "general speech clusters" – i.e. repeated sequences of words that are frequent in authentic 20th century spoken language and Dickens's overall fictional speech – can also be found in GE. Our argument is that frequent clusters both in corpora of authentic and fictional speech can act as a list of candidates that we use

Final draft accepted for publication (may contain minor errors and infelicities). Please cite the published version: Mahlberg, M., & Wiegand, V. (2018). Corpus stylistics, norms and comparisons: Studying speech in *Great Expectations*. In R. Page, B. Busse, & N. Nørgaard (Eds.), *Rethinking Language, Text and Context: Interdisciplinary Research in Stylistics in Honour of Michael Toolan* (pp. 123–143). London: Routledge.

to look through an individual text. Even if these speech clusters only occur once in the individual text, this provides evidence for their speech-like quality.

Table 2 shows the most frequent clusters in GE quotes that also reach high frequency thresholds in the spoken component of the BNC and the entire quotes subset of all of DNov. The three main columns following the individual clusters provide the frequency information for each of the three corpora. The clusters are displayed in the order of their frequency in GE, which is shown in the first column following the clusters ('freq.'). Accordingly, *I don't want to* is with 8 occurrences the novel's most frequent quote cluster that is also highly frequent in the two larger corpora DNov and BNC. Note that purely based on cluster frequency in GE quotes, the top cluster would be *I am going to* (occurring 10 times), but this cluster does not reach the thresholds in the other corpora and hence is not included in Table 2. As all three corpora are of different sizes (cf. Table 1), Table 2 displays both the raw and relative frequencies (normalised per million words). For very small corpora, relative frequencies generally lead to inflated numbers. So even a single occurrence like *I don't know what* reaches a relative frequency of almost 19 per million words in GE. However, this is not a problem in the present chapter as we are mainly interested in the frequency of the clusters in the larger corpora to make our point. Importantly, as our method enables us to find clusters that are defined by comparison, we can pick up the examples at the bottom of Table 2 that only occur once in GE. If clusters are run for an individual text, they need to be found at least twice to become noticeable. And often, minimum thresholds, e.g. at least five occurrences in the text under investigation, as in one of our earlier studies on GE (Mahlberg 2007), introduce a particular selection that prevents to find the kind of examples we identify in this chapter.

Final draft accepted for publication (may contain minor errors and infelicities). Please cite the published version: Mahlberg, M., & Wiegand, V. (2018). Corpus stylistics, norms and comparisons: Studying speech in *Great Expectations*. In R. Page, B. Busse, & N. Nørgaard (Eds.), *Rethinking Language, Text and Context: Interdisciplinary Research in Stylistics in Honour of Michael Toolan* (pp. 123–143). London: Routledge.

Still, in this study we operate with thresholds, but they take a broader view first before narrowing down the selection in GE. Table 2 only contains clusters that reach a frequency of at least 20 words per million both in DNov and the Spoken BNC and also appear in a minimum of five texts in both of these corpora. By meeting these thresholds, the clusters qualify as what has been termed "lexical bundles" in the corpus linguistic literature (Biber et al., 1999, p. 989): 'the sequences of words that most commonly co-occur in a register'. Clusters that meet these criteria in DNov and the BNC will bring a speech-like quality to GE, even if they only appear there once. The clusters are common in the speech that readers are habitually exposed to so in the fictional text create an effect of spokenness.

Cluster lists for GE, DNov and all other CLiC corpora can be retrieved with the CLiC web app. Currently the app interface only displays the most frequent clusters (occurring at least five times), but the complete cluster information can be retrieved via the CLiC API (the API documentation is available from https://github.com/birmingham-ccr/clic/blob/1.6/doc/api_usage.rst). Accordingly, the GE and DNov frequencies in Table 2 have been collected via the CLiC API (more detail on the data collection procedure is described in Mahlberg et al., in preparation). The BNC clusters have been generated from the spoken component of the XML edition of the BNC from 1994 (available from http://ota.ox.ac.uk/desc/2554).

Table 2: Clusters in GE that are highly frequent in DNov and the Spoken BNC

	cluster	GE		DNov			Spoken BNC		
		freq.	relative freq.	freq.	# of texts	relative freq.	freq.	# of texts	relative freq.
1	i don't want to	8	150.32	51	11	37.25	529	261	53.44
2	what do you mean	7	131.53	196	15	143.17	337	165	34.04
3	what do you think	5	93.95	117	15	85.46	508	242	51.32
4	what do you want	5	93.95	74	14	54.05	359	123	36.26
5	i don't know how	4	75.16	66	14	48.21	413	221	41.72
6	you know what i	3	56.37	30	12	21.91	541	176	54.65

Final draft accepted for publication (may contain minor errors and infelicities). Please cite the published version: Mahlberg, M., & Wiegand, V. (2018). Corpus stylistics, norms and comparisons: Studying speech in *Great Expectations*. In R. Page, B. Busse, & N. Nørgaard (Eds.), *Rethinking Language, Text and Context: Interdisciplinary Research in Stylistics in Honour of Michael Toolan* (pp. 123–143). London: Routledge.

7	are you going to	3	56.37	41	13	29.95	474	190	47.88
8	i tell you what	3	56.37	84	13	61.36	293	119	29.60
9	i don't think i	3	56.37	39	14	28.49	284	162	28.69
10	i was going to	2	37.58	70	13	51.13	416	223	42.02
11	nothing to do with	2	37.58	34	11	24.84	246	149	24.85
12	i don't know what	1	18.79	162	15	118.33	854	283	86.27
13	at the same time	1	18.79	74	14	54.05	459	260	46.37
14	i don't know why	1	18.79	28	12	20.45	276	142	27.88
15	but i don't know	1	18.79	37	12	27.03	236	140	23.84
16	is one of the	1	18.79	30	10	21.91	224	157	22.63
17	for a long time	1	18.79	36	13	26.30	215	151	21.72

Concordance 2 shows all instances of the most frequent quote cluster in GE – *I don't want to* – that also qualifies as a lexical bundle in the BNC and DNov quotes. *I don't want to* is followed by *know* in half of the cases, *be betrayed*, *get into trouble*, *go* and once as part of an answer ('Why don't you cry?' 'Because I don't want to.').[3] This phrase is not tied to a particular character.

	Left	Node	Right	In bk.
1	nd."Why don't you cry?""Because	I don't want to.""	You do," said she."You have been	
2	mpatient movement of her fingers	."I don't want to	know.Are you ready to play?"I was	
3	shilling," observed the coachman	."I don't want to	get into trouble.I know him!"He da	
4	e, when you borrowed that money	.""I don't want to	know what passed between Herber	
5	wled the convict I had recognized	."I don't want to	go.I am quite ready to stay behind	
6	d - any one.Don't tell me anything	:I don't want to	know anything; I am not curious."(	
7	s ago, without his knowledge, and	I don't want to	be betrayed.Why I fail in my ability	
8	d - whichever it may be - you and	I don't want to	know - quite successfully.At the ol	

Concordance 2: All 8 instances of *I don't want to* in *Great Expectations*

Table 2 includes three examples of the trigram *I don't know* followed by *how*, *what*, *why* respectively which reflects that *I don't know* is among the top frequent clusters in spoken English. Importantly, however, not all clusters in Table 2 should automatically

[3] *I don't want to* is an example for which the automatic speech annotation shows some mistake. There are also two instances found in non-quotes, because quotation marks were missed. This annotation mistake, however, does not affect our argument.

Final draft accepted for publication (may contain minor errors and infelicities). Please cite the published version: Mahlberg, M., & Wiegand, V. (2018). Corpus stylistics, norms and comparisons: Studying speech in *Great Expectations*. In R. Page, B. Busse, & N. Nørgaard (Eds.), *Rethinking Language, Text and Context: Interdisciplinary Research in Stylistics in Honour of Michael Toolan* (pp. 123–143). London: Routledge.

be regarded as "speech bundles". The cluster *at the same time* makes the lexical bundles threshold of 20 that we have set here, but it is also a lexical bundle in non-quotes – which is unsurprising as *time* is a general noun, i.e. it is among the top most frequent nouns in the language based on general reference corpora (Mahlberg, 2005). In fact, in our non-quotes set the relative frequency of *at the same time* is even higher (73.26 per million) than in quotes. A comparison as in Table 2 can provide a reference point for clusters that occur infrequently or even just once if a single text is the only data set. By interpreting their functions based on their common patterns valuable information is provided on what makes fictional speech speech-like. The identification of such speech bundles in fiction is particularly important, because idiosyncratic speech clusters, such as those discussed in the next section, tend to be more noticeable to the reader (and the critic) and so receive more attention.

2) Idiosyncratic character clusters in GE

In Mahlberg's (2007; 2013) terms, a cluster that is specific to a particular character is referred to as a "label". Table 2 does not include any labels but these can be found by examining a list of all GE clusters. We can distinguish between idiosyncratic clusters that occur only locally at one particular point of the text and those that appear across various chapters. One well-known example of the first type is an extended cluster of Joe's affective address for Pip, based on the core *old chap*, a sample of which is shown in Concordance 2. We have used the KWICGrouper to identify the displayed concordance lines based on matches of *dear*, *old* and *Pip* within three words to the left of *old chap*. The top 6 lines, highlighted in purple in the interface, are the six examples of the five-word cluster *dear old Pip old chap*, which contains two 4-word clusters that are relevant to the data in this chapter, the two fragments: *old Pip old*

Final draft accepted for publication (may contain minor errors and infelicities). Please cite the published version: Mahlberg, M., & Wiegand, V. (2018). Corpus stylistics, norms and comparisons: Studying speech in *Great Expectations*. In R. Page, B. Busse, & N. Nørgaard (Eds.), *Rethinking Language, Text and Context: Interdisciplinary Research in Stylistics in Honour of Michael Toolan* (pp. 123–143). London: Routledge.

chap and *old Pip old chap*. The co-text shows that line 1 is different from the others (see Example (4)). Although the sequence summarises Joe's speech habit of addressing Pip, Pip describes the phrase as one of Joe's "old names" for him, commenting generally on the use of this phrase rather than narrating a particular speech instance.

(4) As I became stronger and better, Joe became a little less easy with me. In my weakness and entire dependence on him, the dear fellow had fallen into the old tone, and called me by the old names, the dear "old Pip, old chap," that now were music in my ears. I too had fallen into the old ways, only happy and thankful that he let me. But, imperceptibly, though I held by them fast, Joe's hold upon them began to slacken; and whereas I wondered at this, at first, I soon began to understand that the cause of it was in me, and that the fault of it was all mine. (GE, Chapter 57)

As illustrated in Concordance 3, the CLiC tags menu makes it possible to annotate an example like this with a customised tag; in this concordance, we have opted to tag this particular instance with 'pip' and all others with 'joe' in order to mark the speaker.

Final draft accepted for publication (may contain minor errors and infelicities). Please cite the published version: Mahlberg, M., & Wiegand, V. (2018). Corpus stylistics, norms and comparisons: Studying speech in *Great Expectations*. In R. Page, B. Busse, & N. Nørgaard (Eds.), *Rethinking Language, Text and Context: Interdisciplinary Research in Stylistics in Honour of Michael Toolan* (pp. 123–143). London: Routledge.

	Left	Node	Right	Book	Ch.	Par.	Sent.	In bk.	joe	pip
1	lled me by the old names, the dear "old Pip,	old chap,"	that now were music in my ears.I too had	GE	57	80	174			✓
2	at last.And so GOD bless you, dear old Pip,	old chap,	GOD bless you!"I had not been mistaken in r	GE	27	65	143		✓	
3	his joy that I knew him."Which dear old Pip,	old chap,"	said Joe, "you and me was ever friends.And	GE	57	19	42		✓	
4	I have been ill, Joe," I said."Dear old Pip,	old chap,	you're a'most come round, sir.""It has been a	GE	57	90	198		✓	
5	Joe, how smart you are!""Yes, dear old Pip,	old chap."	I looked at both of them, from one to the	GE	58	46	112		✓	
6	me, in the time to come!""O dear old Pip,	old chap,"	said Joe."God knows as I forgive you, if I	GE	58	60	138		✓	
7	nly heart as he gave me his hand."Pip, dear	old chap,	life is made of ever so many partings welded	GE	27	65	133		✓	
8	ad at me in very serious remonstrance."Pip,	old chap!	You'll do yourself a mischief.It'll stick somew	GE	2	29	74		✓	
9	eived it as a miracle of erudition."I say, Pip,	old chap!"	cried Joe, opening his blue eyes wide, "what	GE	7	9	30		✓	
10	r the best of friends; an't us, Pip?Don't cry,	old chap!"	When this little interruption was over, Joe re	GE	7	47	109		✓	
11	hand!""Good-bye, Joe!""God bless you, Pip,	old chap!"	I had never parted from him before, and wha	GE	7	91	207		✓	
12	n Joe, Joe contemplated me in dismay."Pip,	old chap!	This won't do, old fellow!I say!Where do you	GE	9	55	127		✓	
13	r.Joe greeted me as usual with "Halloa, Pip,	old chap!"	and the moment he said that, the stranger tu	GE	10	6	24		✓	
14	him, "Dear Joe, how are you?" he said, "Pip,	old chap,	you knowed her when she were a fine figure	GE	35	7	21		✓	
15	Pip, how long have your illness lasted, dear	old chap?""	Yes, Joe.""It's the end of May, Pip.To-morrow	GE	57	23	49		✓	
16	giv' him the name of Pip for your sake, dear	old chap,"	said Joe, delighted when I took another stoo	GE	59	2	4		✓	

Concordance 3: 16 of 43 instances of *old chap* in GE

Another example of an idiosyncratic cluster that is not restricted to one place in the text is *all right John all* (and its variation *right John all right*). The Aged P.'s notorious response to his son John Wemmick occurs across four different chapters. By contrast, the two four-word clusters *put the case that* and *Tom, Jack, or Richard* are each just used in one chapter, by Jaggers and Wemmick, respectively. Both cases are noteworthy in that they occur frequently within the space of only a few paragraphs.

3) Clusters for which the tendency to occur in quotes affects their function when occurring in non-quotes

The eleventh cluster in Table 2, *nothing to do with*, is the first cluster that occurs in GE and is a lexical bundle in both spoken subcorpora of the BNC and DNov, but *does not* contain a personal pronoun. If clusters contain a first or second person personal pronoun

Final draft accepted for publication (may contain minor errors and infelicities). Please cite the published version: Mahlberg, M., & Wiegand, V. (2018). Corpus stylistics, norms and comparisons: Studying speech in *Great Expectations*. In R. Page, B. Busse, & N. Nørgaard (Eds.), *Rethinking Language, Text and Context: Interdisciplinary Research in Stylistics in Honour of Michael Toolan* (pp. 123–143). London: Routledge.

their association with speech is more direct – the pronouns reflect interpersonal relationships between speaker and hearer. In Mahlberg (2013), which was not based on the CLiC subsets yet, the occurrence of the pronouns is used as defining feature for "speech clusters". The speech bundle *nothing to do with* does not make the lexical bundle threshold in non-quotes: it occurs 23 times, i.e. 9.36 per million, in DNov. In GE, the cluster occurs three times altogether, twice in quotes and once in non-quotes. Without the comparison with the reference corpora, these would not be reliable figures to claim a tendency to occur in speech. Based on the comparison in Table 2, however, Example (5), where *nothing to do with* appears in non-quotes, can be seen to illustrate the speech-like quality of the first-person narrator's interpretation of Jaggers's speech.

(5) "When that person discloses," said Mr. Jaggers, straightening himself, "you and that person will settle your own affairs. When that person discloses, my part in this business will cease and determine. When that person discloses, it will not be necessary for me to know anything about it. And that's all I have got to say."

We looked at one another until I withdrew my eyes, and looked thoughtfully at the floor. From this last speech I derived the notion that Miss Havisham, for some reason or no reason, had not taken him into her confidence as to her designing me for Estella; that he resented this, and felt a jealousy about it; or that he really did object to that scheme, and would have **nothing to do with it**. When I raised my eyes again, I found that he had been shrewdly looking at me all the time, and was doing so still.

Final draft accepted for publication (may contain minor errors and infelicities). Please cite the published version: Mahlberg, M., & Wiegand, V. (2018). Corpus stylistics, norms and comparisons: Studying speech in *Great Expectations*. In R. Page, B. Busse, & N. Nørgaard (Eds.), *Rethinking Language, Text and Context: Interdisciplinary Research in Stylistics in Honour of Michael Toolan* (pp. 123–143). London: Routledge.

> "If that is all you have to say, sir," I remarked, "there can be nothing left for me to say." (GE, Chapter 36)

If we broaden the focus from one novel to the entire DNov, another way to account for differences between quotes and non-quotes is a key comparison. As we argued earlier, key comparison is a popular method in corpus stylistics. A comparison across subsets illustrates another dimension of this approach for the study of fiction. Table 3 shows the top ten 4-word "key clusters", i.e. those clusters that are significantly more frequent in DNov quotes compared to non-quotes, which also include *I don't know what* from Table 2.

	Freq. in quotes	**Freq. in non-quotes**	**Keyness**
what do you mean	196	2	413.03
i beg your pardon	197	3	407.25
i should like to	137	7	252.84
how do you do	120	3	239.78
what do you think	117	3	233.29
i am going to	128	9	224.11
i don't know what	162	37	197.1
what is the matter	92	2	185.86
i am sure you	83	0	183.69
i tell you what	84	2	168.51

Table 3: Top ten 4-word clusters for key comparison quotes vs non-quotes in DNov, for all key clusters $p<0.0001$

If we now move from the broader perspective back to individual textual examples, the key comparison can help find cases where speech clusters take on specific functions when they appear in non-quotes. One of the key clusters is *of such a thing* – occurring 28 times in quotes and 10 times in non-quotes. It is also part of the cluster *I never heard of such a thing.* Figure 2 shows one of the non-quote examples. In Figure 2, *of such a*

Final draft accepted for publication (may contain minor errors and infelicities). Please cite the published version: Mahlberg, M., & Wiegand, V. (2018). Corpus stylistics, norms and comparisons: Studying speech in *Great Expectations*. In R. Page, B. Busse, & N. Nørgaard (Eds.), *Rethinking Language, Text and Context: Interdisciplinary Research in Stylistics in Honour of Michael Toolan* (pp. 123–143). London: Routledge.

thing or even the longer *heard of such a thing* uses spoken features in narration, which is part of the way in which Sir Leicester's and the reaction of the debilitated cousin are presented. Such examples of spokenness can make an important contribution to the discussion of criteria for Free Indirect Speech. As both Toolan (2009) and Busse (2010) have argued, the definition of Free Indirect Speech is difficult to formalise so that examples could be automatically retrieved. The type of comparisons we suggest in this chapter will not be able to resolve the issue of automatization, but taking the spoken-like qualities of speech clusters into account that we have described, does add a new method for finding candidate examples for Free Indirect Speech.

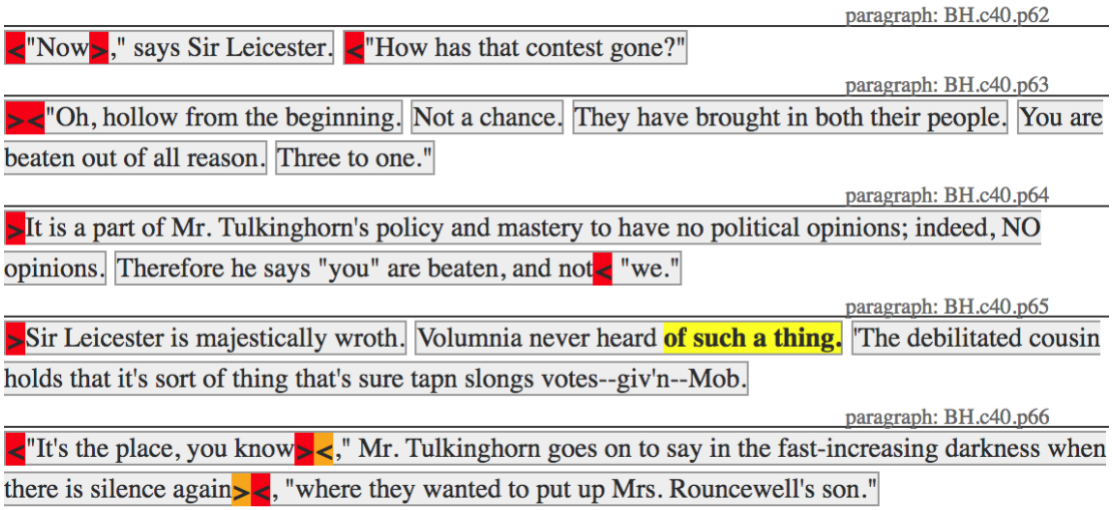

Figure 2: CLiC view of Chapter 40 in *Bleak House* showing *of such a thing* being used in non-quotes

Conclusions

In this chapter, we moved away from a clear-cut distinction into literary and non-literary language by linking the notions of deviation and norms that are drawn on in literary stylistics to corpus linguistic comparisons of different corpora. In this way, we have also situated corpus stylistics within wider trends in digital humanities. Corpus

stylistics strikes a balance between the interest of the stylistician in a detailed textual analysis and the concerns of data science approaches that look for trends across large amounts of data. With functionalities of the CLiC web app we have highlighted how standard corpus linguistic tools can be adjusted to literary stylistic concerns. Our discussion of spokenness in particular stresses that the theoretical concept of the norm hides that categories found in a text are much fuzzier. The thresholds we used in Table 2 as well as the statistics that underlie key comparison only seemingly suggest a clear cut-off and the also depend on the corpora we use. By using the BNC for a comparison with GE and DNov, we have shown how fictional speech shares lexico-grammatical elements with real spoken language. At the same time, our approach underlines that the notion of comparison is crucial and tendencies are more important than exact thresholds. Especially the examples of clusters that tend to occur in quotes but are also found, although less frequently so, in non-quotes show how frequency tendencies affect textual functions in local contexts. These insights have implications for fundamental stylistic models and categories. In particular the concept of Free Indirect Speech – a thorny category in literary stylistics – will need to be reconsidered with regard to quantitative measures to account for features of spokenness.

Reference List

Anthony, L. (2018) AntConc (Version 3.5.6). Tokyo: Waseda University. Available from http://www.laurenceanthony.net/software (Last accessed April 2018).

Bednarek, M. (2011) 'The language of fictional television: A case study of the 'dramedy' Gilmore Girls', *English Text Construction*, 4 (1), pp. 54–84.

Biber, D. (1988) *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, B. (2006) University language: a corpus-based study of spoken and written registers. Amsterdam: John Benjamins.

Biber D., Johansson S., Leech G., Conrad, S., and Finegan, E. (1999) *Longman grammar of spoken and written English*. Harlow: Longman.

Burrows, J., and Craig, H. (2012) 'Authors and Characters', *English Studies*, 93 (3), pp. 292–309.

Busse, B. (2010) *Speech, writing and thought presentation in a corpus of 19th-century narrative fiction*. Bern: University of Bern.

Carter, R. (2016) *Language and creativity: the art of common talk*. 2nd ed. London: Routledge.

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), pp. 29–59.

Dörk, M., and Knight, D. (2015) 'WordWanderer: a navigational approach to text visualisation', *Corpora*, 10 (1), pp. 83–94.

Eder, M. (2017) 'Visualization in stylometry: cluster analysis using networks', *Digital Scholarship in the Humanities*, 32 (1), pp. 50–64.

Fischer-Starcke, B. (2010) *Corpus linguistics in literary analysis: Jane Austen and her contemporaries*. London/New York: Continuum.

Gavins, J. (2007) *Text world theory: an introduction*. 1 ed. Edinburgh: Edinburgh University Press.

Hoover, D. L. (2007) 'Corpus Stylistics, stylometry, and the styles of Henry James', *Style*, 41 (2), pp. 174–203.

Jockers, M. L. (2013) *Macroanalysis: digital methods and literary history*. Champaign, IL: University of Illinois Press.

Lambert, M. (1981) *Dickens and the suspended quotation*. New Haven: Yale University Press.

Leech, G. (1985) 'Stylistics', in van Dijk, T. A. (ed.) *Discourse and literature*. Amsterdam: John Benjamins, pp. 39–57.

Leech, G., and Short, M. (2007) [1981] *Style in fiction: a linguistic introduction to English Fictional prose*. 2nd ed. Harlow: Pearson.

Love R, Dembry C, Hardie A, Brezina, V., and McEnery, T. (2017) 'The Spoken BNC2014: designing and building a spoken corpus of everyday conversations', *International Journal of Corpus Linguistics*, 22 (3), pp. 319–344.

Louw, W. E. (1993) 'Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies', in Baker, M., Francis, G., and Tognini-Bonelli, E. (eds.) *Text and technology: in honour of John Sinclair*. Amsterdam: John Benjamins, pp. 157–74.

Mahlberg, M. (2005) *English general nouns: a corpus theoretical approach*. Amsterdam: John Benjamins.

Mahlberg, M. (2007) 'A corpus stylistic perspective on Dickens' *Great Expectations*', in Lambrou, M. and Stockwell, P. (eds.) *Contemporary stylistics*. London: Continuum, pp. 19–31.

Mahlberg, M. (2013) *Corpus stylistics and Dickens's fiction*. London: Routledge.

Mahlberg, M. and Smith, C. (2012) 'Dickens, the suspended quotation and the corpus', *Language and Literature*, 21 (1), pp. 51–65.

Mahlberg, M., Stockwell, P., Joode, J. de, Smith, J., C., and O'Donnell, M. B. (2016) 'CLiC Dickens: novel uses of concordances for the integration of corpus stylistics and cognitive poetics', *Corpora*, 11 (3), pp. 433–463.

Mahlberg, M., Stockwell, P., Wiegand, V., and Hennessey, A. (in preparation) 'Speech in the 19th century English novel'.

Moretti, F. (2005). Graphs, Maps, Trees: Abstract Models for a Literary History. London/New York: Verso.

Murphy, S. (2015) 'I will proclaim myself what I am: corpus stylistics and the language of Shakespeare's soliloquies', *Language and Literature*, 24 (4), pp. 338–354.

Nini, A. (2014) Multidimensional Analysis Tagger 1.2 - Manual. Retrieved from: http://sites.google.com/site/multidimensionaltagger (Last accessed April 2018).

Page, N. (1988) *Speech in the English novel*. 2nd ed. Houndmills: Macmillan.

Rayson, P. (2008) 'From key words to key semantic domains', *International Journal of Corpus Linguistics*, 13 (4), pp. 519–549.

Ruano San Segundo, P. (2017) 'Reporting verbs as a stylistic device in the creation of fictional personalities in literary texts', *Atlantis*, 39 (2), pp. 105–124.

Scott, M. (2006) 'Keywords of individual texts: aboutness and style', in, Scott, M. and Tribble, C. (eds) *Textual patterns: key words and corpus analysis in language education*. Amsterdam: John Benjamins, pp. 55–72.

Scott, M. (2017) WordSmith Tools version 7. Stroud: Lexical Analysis Software. Available from http://lexically.net (Last accessed April 2018).

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

Shepherd, T.M.G., and Berber Sardinha, T. (2013) 'A rough guide to doing corpus stylistics', *Matraga, Rio de Janeiro*, 20 (32), pp. 66-89.

Short, M. (1996) *Exploring the language of poems, plays and prose*. London: Longman.

Simpson, P. (2004) *Stylistics: a resource book for students*. London/New York: Routledge.

Sinclair, J. (1991) *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Spitzer, L. (1948) *Linguistics and literary history: essays in stylistics*. Princeton: Princeton University Press.

Stockwell, P. and Whiteley, S. (2014) 'Introduction', in Stockwell, P. and Whiteley, S. (eds.) *The Cambridge handbook of stylistics*. Cambridge: Cambridge University Press, pp. 1–9.

Toolan, M. (2006) 'Top keyword abridgements of short stories: a corpus linguistic resource?', *Journal of Literary Semantics,* 35 (2), pp.

Toolan, M. (2009) *Narrative progression in the short story: a corpus stylistic approach*. Amsterdam: John Benjamins.

Toolan, M. (2014) 'The theory and philosophy of stylistics', in Stockwell, P. and Whiteley, S. (eds.) *The Cambridge handbook of stylistics*. Cambridge: Cambridge University Press, pp. 13–31.

Toolan, M. (2016) *Making sense of narrative text: situation, repetition, and picturing in the reading of short stories*. New York: Routledge.

Underwood T., Bamman, D., and Lee, S. (2018) 'The Transformation of Gender in English-Language Fiction', *Journal of Cultural Analytics*.

Wales, K. (2001) *A dictionary of stylistics*. 2nd ed. Harlow: Longman.

Wang, J., Ho, Y., Xu, Z., McIntyre, D. and Lugea, J. (2016) 'The visualisation of cognitive structures in forensic statements', in *20th International Conference Information Visualisation*, Universidade NOVA de Lisboa, Lisbon, Portugal,

20 July 2016. Available at: http://eprints.hud.ac.uk/id/eprint/29086/ (Last accessed April 2018)

Werth, P. (1999) *Text Worlds: Representing Conceptual Space in Discourse*. Harlow: Longman.