Fiction – one register or two?

Speech and narration in novels

Jesse Egbert & Michaela Mahlberg
Northern Arizona University | University of Birmingham

In this paper our focus is on analyzing register variation within fiction, rather than between fiction and other registers. By working with subcorpora that separate text within and outside of quotation marks, we appromixate fictional speech and narration. This enables us to identify and compare linguistic features with regard to different situational contexts in the fictional world. We focus in particular on the novels of Charles Dickens and a reference corpus of other 19th-century fiction. Our main method for the register analysis is Multi-dimensional Analysis (MDA) for which we draw on altogether four dimensions from two previous MDAs. The linguistic distinctions we identify highlight similarities between fictional speech and involved registers such as face-to-face communication, and between narration and more informational and narrative prose. In addition to the detailed information on register features that characterize speech and narration, the paper raises more general questions about the ability of register studies to deal with situational contexts within fiction.

Keywords: register variation, fictional speech, narration, 19th-century fiction, Charles Dickens

1. Introduction

Fiction has often been included in register studies. So far, however, there has been little large-scale research focused on linguistic variation and associated functions within narrative fiction, and in particular novels. A register is a variety of texts associated with a particular situational context and particular linguistic features. The linguistic features of a register are functional, i.e. their (co-)occurrence in a register tends to reflect the register's purposes and situational context (Biber & Conrad 2009: 6). So registers and the functions for which language is used within them affect the linguistic choices that language users make when producing texts

https://doi.org/10.1075/rs.19006.egb
Register Studies 2:1 (2020), pp. 72–101. issn 2542-9477 | e-issn 2542-9485

(e.g., Biber 1988; Halliday 1991). Examples of registers include casual conversations, text messages, research articles, and business meetings. Speakers in a casual conversation, for instance, frequently use contracted forms and *that* deletion to increase fluency, and they will often use personal pronouns and demonstratives to make reference to people and objects within the shared situational context.

As Biber and Conrad (2009: 132) point out, "[f]rom a situational perspective, fiction is one of the most complicated registers". This is because the real-world context in which a fictional text is produced is "almost irrelevant", what matters is the construction of the fictional world: "the relevant situational context for a fictional text is the fictional world that the author creates in the text itself" (Biber & Conrad 2009: 132). If the real-world context is considered to be of little relevance, much of the linguistic variation in fiction appears as stylistic variation. Especially in quantitative approaches, methods for the linguistic analysis of features of register and features of style can be very similar. Lexical and grammatical features are identified, quantified and compared across sets of texts. In register studies, a crucial difference will be the interpretation of the reasons for any variation. Register variation results from functional language use associated with the communicative situation, whereas "causes of style variation are related to aesthetic preferences and attitudes about language" (Biber & Conrad 2009: 21).

This distinction between register and style is one that is made from a linguistic point of view, where fiction can be analyzed along with other registers, such as newspaper writing or academic prose. From a literary point of view, however, the situational context of fiction may be given different weight. The social and historical context in which novels were produced and received plays a major role in literary scholarship. While Biber and Conrad (2009: 156) only briefly touch upon the demographic change and the growth of the reading public across the 18th and 19th centuries that might have had an impact on linguistic features in the novel, Underwood (2019: xii) points out that "quantitative literary research now starts with social evidence about things that really interest readers of literature – like audience, genre, character, and gender". Some of this social evidence does relate to the communicative situation and purpose of fiction that will affect linguistic choices.

From a literary stylistic point of view, the distinction between aesthetic and functional choices is even more difficult to uphold. Cognitive stylistics has emphasized that meaning is made in the mind of reader. Especially with regard to the creation of fictional characters, the relationship between information in the text and real-world knowledge has received much attention. Culpeper (2001) specifically describes characterization as a process where textual cues and the reader's background knowledge work together to create an impression of fictional people in the mind of the reader. With his approach to mind-modelling, Stockwell

(2009) gives even more emphasis to the relationship between real and fictional people. The application of corpus linguistic methods helps to identify lexico-grammatical patterns that show the continuity between real-world knowledge and elements of fictional worlds (Mahlberg & Stockwell 2016, Mahlberg et al. 2019a). So what might appear to be an aesthetic choice can have a function in the creation of fictional worlds.

To be able to address the complexity of fiction, nothing less than a fully interdisciplinary approach will be required. But even if we stay within the framework of register studies, the question of the situational context deserves more attention than it has received to date. If the relevant situational context is the fictional world, as Biber and Conrad (2009: 132) indicate, what does that this mean for the way in which we interpret linguistic features in fiction functionally, i.e. as features of a register? In his comprehensive study of register variation, Biber (1988) includes fiction and compares it to other varieties such as telephone conversations, interviews, press reportage or official documents. His analysis of linguistic features indicates that fiction is a special case, where a text-internal physical and temporal situation has to be taken into account. While there can be linguistic features that might look like exophoric references to the discourse production (e.g. *last night, the kitchen table, David's school*) the situation they refer to is a fictional one and the reader understands such references "in terms of the internal physical and temporal situation developed in the text rather than any actually existing external context" (Biber 1988: 148).

'The fictional world' is a rather broad concept and within a specific fictional world, a range of imaginary situational contexts can be developed. So it is reasonable to expect register variation *within* fiction. However, register studies so far has not given much attention to this variation. In this paper, we consider variation within fiction by comparing fictional speech and narration. Fictional speech is a crucial component of novels for the representation of spoken interaction between characters. Narration on the other hand presents a different situational context for the telling of the story. As we will explain in the following, the distinction we make between speech and narration is based on formal criteria, i.e. punctuation, and we focus on 19th-century novels, so our approach concentrates on a specific type of fiction. This focus, however, is crucial to enable an initial comparison that can serve as a baseline for future studies.

2. Fiction: Variety and variation in corpus linguistics

Most previous research on registers has been carried out from a text-linguistic perspective. In text-linguistic research, the unit of observation is the text and the goal is to describe language varieties by analyzing the linguistic characteristics of the texts within those varieties (see Biber & Jones 2009). Traditionally, a major assumption of this line of research is that *texts* are nested within *registers*, but the opposite is not true: *registers* are not nested within *texts*. This may explain, at least in part, why register studies to date has largely disregarded potential differences between narration and fictional speech within novels. To subdivide a novel into two parts, i.e. into fictional speech and narration, would be to divide a coherent text into segments that, while situationally different, are contextually interdependent.

For register studies, the identification and quantification of lexical and grammatical features is crucial so that different registers can be compared. Such comparisons typically rely on corpus linguistic methods, including the design and compilation of adequate corpora. The first version of ARCHER (A Representative Corpus of Historical English Registers) had the potential to set the register study of fiction onto a different path. The early version of the corpus subdivided novels into narration and dialogue. Based on this corpus, Biber and Finegan (1994) used MDA to describe change, or 'drift', over time in the linguistic characteristics of several historical registers. In doing this, they reported separate linguistic results for narration and dialogue from novels, revealing that narration was much more 'Informational' – relying on linguistic features such as nouns, prepositions, and attributive adjectives – while dialogue was more 'Involved' – using linguistic features such as private verbs, contractions, second person pronouns, and emphatics. Unfortunately, the narration-dialogue distinction was eliminated in the second edition of ARCHER[1] and questions about fiction as a homogenous register have not been pursued in subsequent studies using the original version of ARCHER.

While corpus linguistic research generally does not distinguish corpora of fictional speech and narration, several major corpora include novels as one of their register strata, including the British National Corpus (BNC), the Corpus of Contemporary American English (COCA), the Corpus of Historical American English (COHA) and the Longman Corpus of Spoken and Written English. The design of such 'reference' corpora has had a major effect on how linguistic studies have treated fiction and included insights from this variety in fundamental reference

[1] <http://www.helsinki.fi/varieng/CoRD/corpora/ARCHER/updated%20version/background.html> (6 February 2020)

works. The Longman Corpus of Spoken and Written English, for instance, formed the basis for the *Longman Grammar of Spoken and Written English* (Biber et al. 1999) and the Corpus of Contemporary American English led to the compilation of *A Frequency Dictionary of American English* (Davies & Gardner 2010). In Biber et al. (1999), fiction tends to appear as an intermediate benchmark in comparisons of linguistic features across registers, where conversation and academic writing show much more striking contrasts. If we assume that fictional speech will to some extent display similarities with actual spoken language, for instance, in terms of spatial and temporal references, we might conclude that dealing with fictional speech and narration in a single text will simply produce an 'average' of different register features. However, the relative proportions of fictional speech and narration might also have an effect. The results of Biber's (1988) register study suggest that the features of narration carry greater weight in the overall picture (we will return to this point in Section 4.2).

While the separation between speech and narration has not received much attention in the compilation of corpora, questions around the need for full-length novels versus text extracts have been discussed. Copyright restrictions may make it unavoidable to limit the fiction section of a corpus to text extracts. In COCA, for example, novels are represented by only the first chapters.[2] As Biber (1993) has argued, for most studies of relatively frequent linguistic features, 2000-word samples and hence extracts will be sufficient. As the assumption is that register is defined by frequent features, working with parts of novels has generally not been seen as detrimental. However, when we start to drill down into situational variation within fictional worlds, it may become more crucial to consider full texts. Relatively recent work in corpus linguistics has also highlighted that meanings and functions of words and phrases can crucially depend on where in a text they occur (Scott & Tribble 2006; Römer 2010; O'Donnell et al. 2012; Barlow 2016). Such insights relate to work by genre researchers who study the internal structure – or moves and steps – of texts (see, e.g., Swales 1990; Bhatia 2014). This research typically deals with academic writing, as illustrated by Swales (1990), who provides in-depth descriptions of the textual structure of research articles detailing the IMRD (Introduction, Methods, Results, Discussion) organization. Flowerdew's (2003, 2008) research into the 'problem-solution pattern' in technically-oriented report writing adds detail on lexico-grammatical patterning in the types of culturally popular patterns that Hoey (2001) describes across a variety of narrative and non-narrative texts. These studies offer clear evidence that not all texts are linguistically homogeneous, which provides even more reason to consider fictional speech and narration separately.

[2] <https://corpus.byu.edu/coca/compare-anc.asp> (6 February 2020)

Recent studies of register variation online have argued for the existence of 'hybrid' registers composed of texts that combine the situational characteristics of two or more text types (Rosso 2008). For example, research by Egbert et al. (2015) relied on a bottom-up, user-based approach to classifying web documents into register categories. This approach revealed that a sizeable proportion of online texts are hybrid in nature (see also Biber & Egbert 2018). Examples of this include hybrids between Personal Blogs and Advice and between Description of a Thing and FAQs about Information. Unlike fiction, however, for such hybrid registers the relevant situational context is not imagined, but the result of new online situational factors in which language is produced.

In research outside of register studies, there have been a small number of corpus or more generally quantitative studies that argue for the consideration of internal variation in fiction (e.g. Burrows 1987; De Haan 1996; Hubbard 2002; Oostdijk 1990). From a stylistic point of view, Burrows (1987) shows how treating the speech of individual characters in Jane Austen can distinguish different fictional people. These results are complemented by findings from De Haan (1996), who compared the fictional dialogue from seven novels and found that they varied with respect to several variables, including length of sentences and type of reporting verbs used. De Haan (1996: 38) also observes: "fiction takes sort of a middle position between more formal writing on the one hand, and face-to-face conversation on the other" and he calls for future research to investigates stylistic variation in fictional dialogue more comprehensively. Hubbard (2002) takes up this challenge in a study focused entirely on a linguistic description of the dialogue portions of Jane Austen's *Sense and Sensibility.* Oostdijk (1990) works with a small data set of samples from five fictional texts, and looks at a range of textual features that suggest that specifically dialogue in fiction takes up a middle position on a continuous scale of planned and unplanned discourse. Pointing out the limitations of research on the language of fiction, Axelsson (2009) lists the lack of suitably annotated corpora, as well as challenges of working with samples from novels among the reasons for the fact that "so little quantitative research on the language of direct speech in fiction has been published so far" (Axelsson 2009: 190). She argues "[m]ore research on direct speech […] as well as research on the narration parts of fiction would give a fuller picture of 'the language of fiction'" (Axelsson 2009: 191).

For register studies, the comparison of linguistic features across different sets of texts is crucial. In this sense, fictional speech can be contrasted with narration, but it can also be compared to actual speech. Especially in historical linguistics, assumed similarities between fictional and real speech form the basis for the use of samples of fictional texts for the purpose of estimating the characteristics of speech in time periods that pre-date audio recording devices (e.g. Biber &

Finegan 2001; Culpeper & Kytö 2010; Kytö, Rudanko, & Smitterberg 2000). In the context of literary studies, however, the usefulness of comparing fictional and real speech has been viewed more sceptically. In his influential work *Speech in the English novel,* Page (1988) argues that: "for various reasons it seems overwhelmingly likely that no dialogue in novel or play will consist merely, or even mainly, of an accurate transcript of spontaneous speech". Importantly, however, Page (1988: 4) also points out that the study of speech in fiction will benefit from a better understanding of real spoken language, which has for a long time been prevented by "widespread ignorance of the detailed anatomy of actual speech". With the availability of corpora and associated methods of comparison it is now possible to take a fresh view not only on the anatomy of speech in fiction (Mahlberg et al. 2019a), but also on fiction as a register more generally. So in this paper we tackle three research questions:

1. To what extent do fictional speech and narration differ from each other?
2. How do fictional speech and narration compare to other registers?
3. What are the linguistic features that characterize fictional speech and narration?

By answering these questions we are able to reconsider how fiction has been dealt with within register studies more widely, while at the same time providing a more detailed account of fictional speech and narration. Our study is based on 19th-century novels, with a particular focus on the works of Charles Dickens. The main reasons for this are that fiction from this era is easy to access and free of copyright restrictions, novels of this time period have large amounts of direct speech that lends itself to automatic annotation (Mahlberg et al. 2016), and Dickens in particular is a well-studied author.

3. Methododolgy

For our methodology it is crucial to use corpora that separate direct fictional speech from the rest of the narration. By using texts from the CLiC corpora, we address "the main problem" that Axelsson (2009: 191) highlights concerning the "lack of specially adapted corpora". We explain the corpora in Section 3.1. Our main method of analysis to study fictional speech and narration is a Multi-Dimensional Analysis (MDA), as explained in Section 3.2.

3.1 Corpus

The data in this study come from two of the CLiC corpora: the 19th Century Fiction (19C) corpus and the Dickens Novel (DNov) corpus. DNov contains all 15 novels by Charles Dickens. 19C contains 29 novels by 19th-century novelists other than Dickens. The CLiC corpora are annotated for 'quotes', i.e. text within quotation marks, and 'non-quotes', i.e. text outside quotation marks. Generally, quotes are mainly direct speech rather than direct thought or direct writing presentation (in the sense of Semino & Short 2004), and we treat non-quotes as approximation of narration (Mahlberg et al. 2016; Mahlberg et al. 2019b), so more indirect forms of speech presentation will be included in non-quotes. The CLiC corpora are accessible through a web application (clic.bham.ac.uk), but based on the annotation, we can also use the 'quotes' and 'non-quotes' subsets as separate corpora.[3] In this way, we can split the 44 novels into 88 texts. Taken together, these corpora contain a total of more than 8.3 million words of running text, with about 5.3 million words (64%) of narration and nearly 3 million words (36%) of fictional speech. Table 1 contains quote, non-quote, and total word counts for the DNov and 19C corpora.

Table 1. Word counts (quotes, non-quotes, and total) for the texts in DNov and 19C

	Quotes		Non-quotes		TOTAL	
Novel	Texts	Words	Texts	Words	Texts	Words
DNov	15	1,369,029	15	2,456,994	30	3,835,807
19C	29	1,611,083	29	2,882,635	58	4,513,070
Total	44	**2,980,112**	44	**5,339,629**	88	**8,348,877**

The proportions of non-quotes and quotes in each of the texts included in this study is displayed in Figure 1. On average, fictional texts in these corpora contain more narration than fictional speech. However, the relative proportions of text in quotes and non-quotes varies widely across the novels, ranging from 49% non-quotes and 51% quotes (*The Hound of the Baskervilles*) to 83% non-quotes and 17% quotes (*Vanity Fair*). The mean proportions are nearly identical in DNov and 19C.

[3] The texts are available from the CLiC webpage, details here: <https://www.birmingham.ac.uk/schools/edacs/departments/englishlanguage/research/projects/clic/publications.aspx> (6 February 2020)

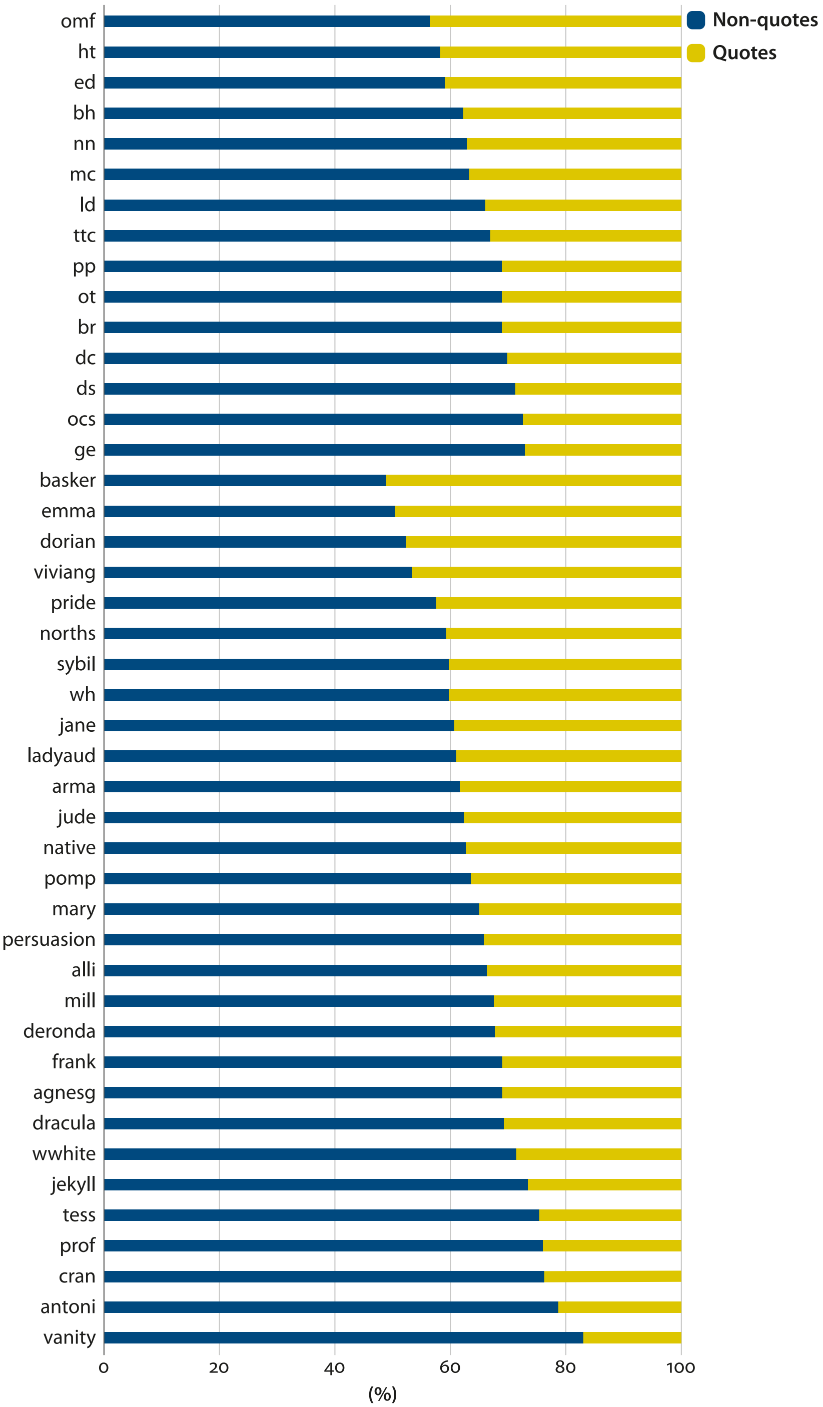

Figure 1. Proportion of text in quotes and non-quotes in 19C and DNov

3.2 Data analysis

The main methodology to answer our three research questions is an MDA approach. MDA is a common method in register studies. Whereas many linguistic studies are based on analyses of one or more individual linguistic features, MDA takes a more comprehensive approach by reducing a large set of relevant linguistic variables down to a smaller set of underlying dimensions of linguistic variation, which are then interpreted functionally or stylistically (see Biber 1988, also e.g. Clarke & Grieve 2017). This is done using factor analysis, a statistical technique that accounts for co-occurrence patterns among variables to produce a parsimonious set of latent factors. There are two types of MDA (see Conrad & Biber 2001). The first type entails carrying out a full MDA, which includes selecting linguistic variables, performing a factor analysis, interpreting the resulting dimensions, calculating dimension scores, and analyzing variation in those dimension scores across the texts or text categories in the corpus. In the second type, researchers rely on the dimension structure of a past MDA study that is based on a similar corpus, or a much more general corpus. The researcher calculates dimension scores for each text in the new corpus based on the dimensions that emerged in the previous study, and then uses those scores to analyze variation across the texts and text categories in the corpus.

In this study, we apply the second type of MDA – and we use two previous studies as our reference points. To answer our research questions 1 and 2, we draw on Biber (1988) and especially Dimension 1 of this study: 'Involved versus Informational Production'. This dimension was based on a corpus of texts from the LOB and London-Lund corpora. It was the most robust dimension that emerged from an analysis of a very general corpus that included narrative fiction along with a wide array of other written and spoken registers. Biber (2014) showed that the 'involved/oral' versus 'informational/literate' opposition captured so clearly in this dimension has emerged, in some form, in nearly every MDA that has been performed during the past 30 years. Using this dimension offers a point of reference for analyzing the oral – literate dimension. It also provides a familiar dimension that has proven to be extremely stable and robust across many studies and in many domains. We will refer to this dimension as Biber_D1. By comparing our quotes and non-quotes corpora against Biber_D1, we will be able to see how these corpora are different from one another, and at the same time, we will be able to compare them to other registers that Biber (1988) had plotted along dimension 1.

In addition to Biber_D1, we use three dimensions from Egbert (2012). They come from an MDA based on the FLAG (FABLE, Longman, ARCHER, and Gutenberg) corpus, a large corpus of fiction from various genres and time periods. The wide array of fiction included in FLAG makes the dimensions ideal for studying linguistic patterns specific to novels. So we use them to answer our third research question that is aimed at describing the actual features that characterize fictional speech and narration. The three dimensions from Egbert (2012) are Egbert_D1 (Thought Presentation versus Description), Egbert_D2 (Abstract Exposition versus Concrete Action), and Egbert_D3 (Dialogue versus Narration). To our knowledge, no previous study has analyzed corpus data based on existing dimension structures from more than one previous MDA. The combination of the two studies enables us to both identify differences between fictional speech and narration through comparison, but also to add detail that can help us explain why we find these differences. Biber_D1 is ideal for comparing fictional speech and narration based on a dimension that has been useful for describing general patterns of language variation across registers. And the three Egbert dimensions are useful for comparing fictional speech and narration based on dimensions that are specific to the domain of fictional prose. It should be noted, however, that there is a certain amount of overlap between the features of Biber_D1 and the three Egbert dimensions.

Complete lists of the positive and negative loading features on each of the four dimensions are contained in Tables 1 to 5. To be able to calculate dimension scores for these dimension, each of the texts in the quotes and non-quotes corpora was automatically tagged using the Biber Tagger (see Biber 1988, 2006). This tagger achieves accuracy levels comparable to other existing taggers (see, e.g., Gray 2011), but it analyzes a larger set of linguistic characteristics than most other taggers. Each tagged text was processed to calculate normalized rates of occurrence (per 1,000 words) for 150+ linguistic features using Biber's TagCount program (see Biber 2006). These tag counts were the basis for the quantitative linguistic analyses performed in this study. Using the tag counts, we calculated dimension scores for each text in the 19C and DNov corpora on the four dimensions from the two previous MD analyses. Dimension scores for each text were calculated by standardizing the linguistic counts for all relevant features using the *z*-score formula based on means and standard deviations from the corpora used in the Biber and Egbert studies. Group differences for research questions 1 and 3 were measured using 2 x 2 factorial ANOVAs. A Bonferroni adjusted alpha was set at .0125 (.05 / 4) to adjust for an increased Type I error rate due to multiple comparisons. In cases where a statistically significant interaction effect was detected, we analyzed and interpreted simple effects. Main effects (i.e. differences among levels of each variable separately) were used for cases where a significant interaction was not found (see Egbert 2015 for a discussion of main effects versus simple effects in corpus linguistic data). For research question 2, we compared the mean scores for the speech and narration corpora to a subset of the registers from Biber's (1988) study. For all comparisons, we treated DNov and 19C separately, so that we were able to see authorial differences, too. This was particularly important because literary scholars have noted differences between authors' effectiveness in representing authentic patterns of spoken language, especially with regard to Dickens (see Page 1988; Sorenson 1989).

Table 2. Positive and negative loading linguistic features on Biber_D1

Positive features – Involved Production
Verbs: private, present tense, do as pro-verb, BE as main verb,
Pronouns: 1st person, 2nd person, indefinite, IT, demonstrative
Stance: emphatics, amplifiers, hedges, possibility modals
Clauses: WH clauses, Causative subordination, nonphrasal coordination, relative clauses
Other: WH questions, THAT deletion, contractions, discourse particles, negation, final prepositions
Negative features – Informational
Noun phrase: nouns, attributive adjectives
Other: word length, type/token, prepositions

Table 3. Positive and negative loading linguistic features on Egbert_D1

Positive features – Thought Presentation
Verbs: perfect aspect, mental, existence
Adverbials: factive, adverbs, likelihood
THAT-clauses: controlled by attitudinal verbs, controlled by factive verbs, controlled by likelihood verbs, controlled by non-factive verbs
Pronouns: indefinite, IT
Other: THAT deletion, TO-clauses controlled by desire verbs, possibility modals, emphatics, WH clauses
Negative features – Description
Noun phrase: nouns, attributive adjectives, prepositions

Table 4. Positive and negative loading linguistic features on Egbert_D2

Positive features – Abstract Exposition
WH relatives: subject position, object position, pied pipes
Nouns: nominalizations, cognition, abstract, process
Passives: BY, agentless
Verbs: split auxiliaries, BE as main verb
Other: Amplifiers, Infinitives
Negative features – Concrete Action
Adverbials: place, time
Verbs: activity, present progressive aspect
Other: Concrete nouns, Contractions, Color adjectives

Table 5. Positive and negative loading linguistic features on Egbert_D3

Positive features – Dialogue
Verbs: present tense, HAVE as main verb, communication, DO as pro-verb **Pronouns**: second person, first person, demonstrative **Modals**: prediction, necessity **Other**: WH questions
Negative features – Narration
Verbs: past tense, occurrence **Other**: Third person pronouns

4. Results and discussion

To answer research question 1, Section 4.1 shows how we measure group differences between fictional speech and narration between Dickens and the other 19th-century authors along Biber_D1. Section 4.2 addresses the second research question, comparing the mean Biber_D1 scores for fictional speech and narration to a subset of the registers from Biber's (1988) study. Section 4.3 compares the fictional speech and narration in DNov and Dickens across Egbert_D1, Egbert_D2, and Egbert_D3 to answer our third research question.

4.1 Linguistic differences between quotes and non-quotes

In this section we use Biber_D1 scores to measure general linguistic differences between (1) the discourse levels of fictional speech and narration, and (2) the sub-corpora of DNov and 19C. As mentioned above, a similar study was performed previously by Biber & Finegan (1994). They also used Dimension 1 from Biber (1988) to reveal extreme differences between the two discourse levels, with fictional speech being much more 'Involved' and fictional narration being much more 'Informational'. In this section, we attempt to replicate that study and add a comparison between two corpora – DNov and 19C – as well as a qualitative analysis of linguistic differences between fictional speech and narration.

We now turn to our statistical analysis of group differences along Biber_D1 scores. A factorial ANOVA revealed no significant interaction between discourse level (fictional speech versus narration) and sub-corpus (DNov versus 19C) in Biber_D1 scores. The main effects revealed significant differences between the texts in the DNov and 19C corpora ($p = .009$) and between the discourse levels of narration and fictional speech ($p < .001$). The overall model was significant, $F(3, 84) = 157.7$, $p < .001$, $R^2_{adjusted} = .84$. These results show that a novel's discourse level and author can explain more than 84% of the variance in Biber_D1 scores in the corpus. These patterns can be clearly seen in Figure 2.

These statistical results show that fictional speech uses significantly more language features associated with involvement and interaction, whereas the narration uses more features associated with informational production. Additionally, the novels of Charles Dickens were significantly more involved and interactive than the other nineteenth-century novels. However, the effect of this difference is extremely small ($\eta^2 = .01$) compared with the large differences between the two discourse levels ($\eta^2 = .83$) in the texts.

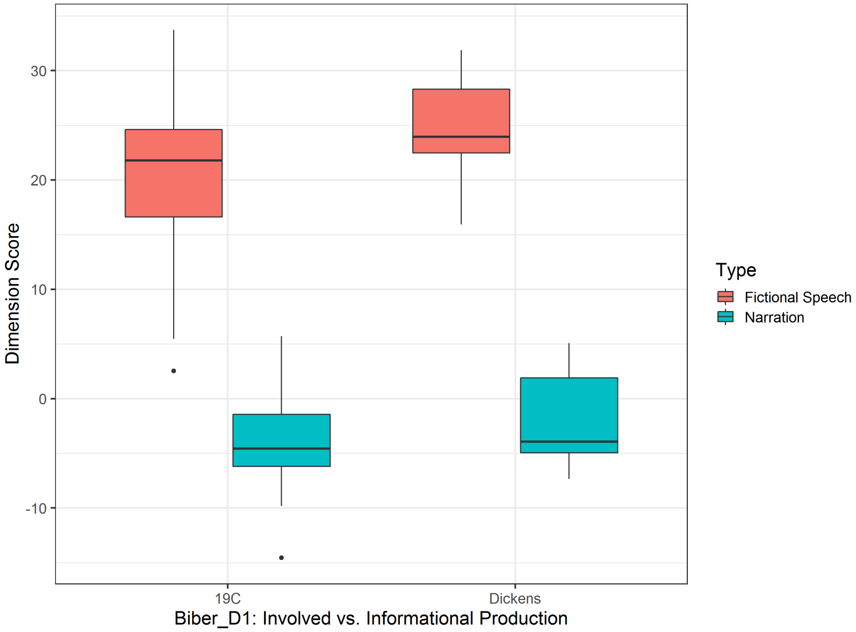

Figure 2. Boxplots of Biber_D1 scores for narration and fictional speech in DNov and 19C. Positive scores are associated with Involved Production; negative scores are associated with Informational Production

On average, Dickens's fictional speech used more features associated with involved production than the fictional speech of other authors. Two of the features associated with involved production are 1st and 2nd person pronouns (including possessive forms). Examples of the use of these pronouns in interactions of characters can be seen in Figure 3, which contains 16 concordance lines of 1st or 2nd person pronouns randomly selected from the quotes subcorpus of DNov. As the concordance lines show, there are also occurrences of names and other forms of address (*my dear, Tom, Mrs Snagsby*) that indicate how characters interact with one another.

Figure 3. A sample of 16 randomly selected concordance lines for 1st and 2nd person pronouns in DNov

bit,' 'My dear, they don't care for **you**, those fellows, if you're NOT hard upon'	OMF_quotes
news do you bring?' 'Oh, Tom!' 'To hear **you** feign that interest in anything that happens to	MC_quotes
these days. Upon my soul, I shouldn't.' '**I** knew it, I was sure of it!' 'My	BR_quotes
's words!" "Now, Mrs. Snagsby, the only amends **you** can make," "is to let me speak a	BH_quotes
rely on me. I have been faithful to **my** post since the days of his Royal Highness	BH_quotes
time, and have forgotten the humble claims upon **my** own, of my hotel, the Crozier.' 'Not at	ED_quotes
got to say to you. You shall give **me** another ten down, and I'll run my	LD_quotes
to have thought that." "Do you, Mr. Pip?" " **I** suppose it will be difficult for you to	GE_quotes
the corner. Not so much. Ha, ha!' 'Yah!' '**you**'re too deep for us, you dog, or	MC_quotes
generally. There are some low minds (not many, **I** am happy to believe, but there are some)	DC_quotes
know what would have become of you, if **I** had not bestowed my hand upon R.W.,	OMF_quotes
of that?' 'Venus tells me so,' 'I tell **you** so,' 'Now, here's my hat, Boffin, and	OMF_quotes
time, my dear, I can assure you (and **you**'ll find this out, Nicholas, for yourself one	NN_quotes
there's any stuff in the world that **I** hate and detest, it's the stuff he	BH_quotes
elligent and respect solicitor is of opinion that **your** affairs are in a bad way, Eugene.' 'Though	OMF_quotes
don't often intrude upon you now, when **I** look round, because I know you are not	LD_quotes

Excerpt 1 is from Dickens's *Oliver Twist*. The fictional speech in this novel had a score of 26.93 and the narration had a score of –6.74. This excerpt illustrates the extreme differences between the linguistic characteristics of fictional speech and narration. We have highlighted in gray three features associated with 'Involved Production' (*1st person pronouns, 2nd person pronouns*, and *contractions*) and in bold three features associated with "Informational Production" (*nouns, attributive adjectives*, and *prepositions*). In the short span of roughly 80 words, the fictional speech text uses a total of 20 occurrences of the three Involved features and only 9 of the Informational features. In contrast, in nearly the same number of words, the narration text uses no occurrences of the 'Involved' features and 28 'Informational' features. We use the format of the table to illustrate the separation into fictional speech and narration that our MDA is based on.

Excerpt 1.
Oliver Twist, by Charles Dickens

Fictional Speech	**Narration**
79 words	**77 words**
'No, no, my **dear**. Not to stop there,'	
	replied the **Jew**.
'We shouldn't like to lose you. Don't be afraid, **Oliver**, you shall come back **to** us again. Ha! ha! ha! We won't be so cruel as to send you away, my **dear**. Oh no, no!'	
	The **old man**, who was stooping **over** the **fire** toasting a **piece of bread**, looked **round** as he bantered **Oliver** thus; and chuckled as if to show that he knew he would still be very glad to get away if he could.
'I suppose,'	
	said the **Jew**, fixing his **eyes on Oliver**,
'you want to know what you're going **to Bill**'s **for** – -eh, my **dear**?'	
'Why, do you think?'	
	inquired **Fagin**, parrying the **question**.
'Indeed I don't know, sir,'	
	replied **Oliver**.
'Bah!'	
	said the **Jew**, turning **away with** a **disappointed countenance from** a **close perusal of** the **boy**'s **face**.
'Wait till **Bill** tells you, then.'	

4.2 Similarities with other registers

As mentioned above, the dimension structure for Biber_D1 was based on corpora that contained many different registers of speech and writing. This makes it possible to compare the dimension scores for the fictional speech and narration in DNov and 19C with other spoken and written registers. This type of comparison can, for instance, provide insights into the degree to which the fictional speech and narration reflect features of naturally occurring spoken language and writing produced in other contexts. Figure 4 displays the mean dimension scores for the four sub-corpora used in this study, as well as for three spoken registers (face-to-face conversation, spoken interviews, and prepared speeches) and three written registers (general fiction, biographies, and academic prose) from Biber (1988). In both DNov and 19C has the fictional speech strong positive Biber_D1 scores (24.5 and 20.1 respectively) and falls in between the scores for face-to-face conversations (35.3) and spoken interviews (17.1). On the other hand, the narration scores for these two corpora are negative for Biber_D1 (−2.0 and −4.0 respectively) and fall between general fiction (−0.8) and biographies (−12.4).

At first glance, it may seem strange that the narration in these two corpora is much closer to 'general fiction' than the fictional speech. One reason for this might be the time difference between the publication of the texts. Biber_D1 is based on findings from the LOB corpus and the fiction component in this corpus was published in the 1960s, so it is possible that texts from this time period have less clearly marked direct speech. Additionally, the LOB texts also contain short stories which might have had an impact. But also the fact that the proportion of non-quotes is higher than quotes (cf. Figure 1), might be relevant here.

Biber and Finegan (1994) also report results comparing Biber_D1 between fictional speech, narration, and other registers of English. Our findings are quite similar to their results, despite differences in the corpora of fiction used in the two studies. The dimension scores for fictional speech in their study ranged between −8 and −24 (note in their study negative values are used for 'Involved Production', whereas in this paper 'Involved Production' has positive values), compared with dimension scores of −20 for spontaneous speeches and −35 for conversation revealing that fictional speech in novels patterns more like these spoken registers than any of the written registers.

4.3 The detail of fictional features

In order to answer the third research question – "What are the linguistic features that characterize fictional speech and narration?" – we compare dimension scores for the three Egbert dimensions between (1) fictional speech and narration and (2) DNov and 19C using factorial ANOVAs. The factorial ANOVA for Egbert_D1 revealed no significant interaction between discourse level and sub-corpus and no significant differences between the DNov and 19C texts. The main effect of discourse level was significant ($p < .001$), resulting in a significant overall model, $F(3, 84) = 66.95$, $p < .001$, $R^2_{adjusted} = .70$. The results reveal that nearly 70% of the variance in Egbert_D1 scores in the corpus can be accounted for by discourse level alone ($\eta^2 = .70$). Fictional speech uses more language associated with 'Thought Presentation', whereas the narration is more descriptive in nature (see Figure 5). It is important to note that 'Thought Presentation' here is not meant to refer to thought presentation in the sense of Leech & Short (2007), but refers to the

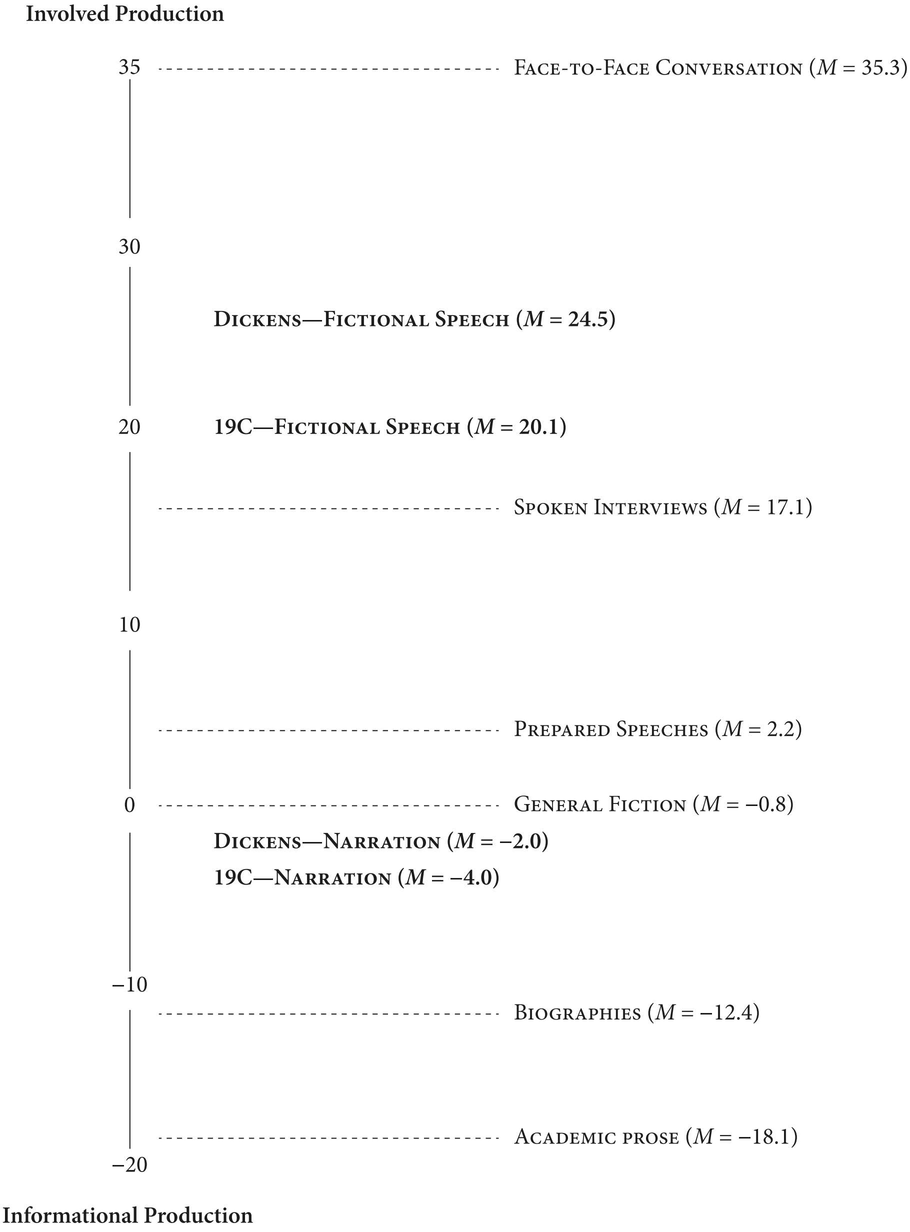

Figure 4. Distribution of the DNov and 19C corpora along Biber_D1 compared to selected registers from Biber (1988), M = mean dimension scores

presence of linguistic features that can refer, for instance, to mental processes or expression of attitudes, likelihood or possibility. Although there was no significant difference between Dickens and his contemporaries in the use of these linguistic features, it is notable how consistent Dickens's writing style was within his fictional speech and narration on this dimension, as evidenced by the small range of variation around the median in the boxplots.

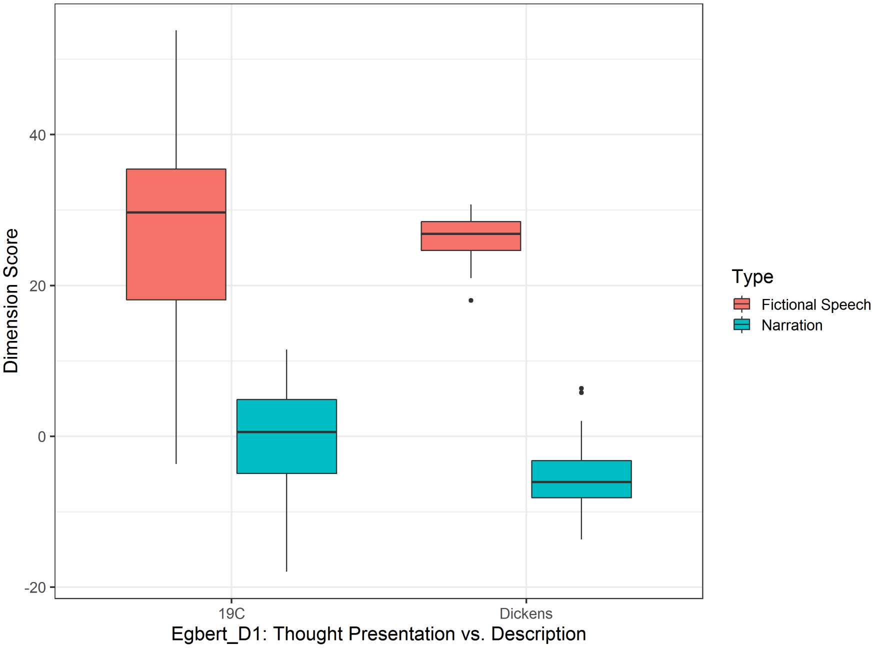

Figure 5. Boxplots of Egbert_D1 scores for narration and fictional speech in DNov and 19C. Positive scores are associated with Thought Presentation; negative scores are associated with Description

Thomas Hardy's *The Return of the Native* has a particularly large difference in Egbert_D1 scores between narration (−4.9) and fictional speech (40.8). Excerpt 2 is from that novel. For reasons of space, in the following we use italics to contrast narration with fictional speech, instead of the table format from Excerpt 1. Excerpt 2 illustrates the extreme differences between the 'Thought Presentation' in fictional speech and the 'Description' in narration. In this excerpt, we have highlighted in gray three features associated with 'Thought Presentation' (*mental verbs, possibility modals*, and *adverbs*). We highlighted in bold three features associated with "Description" (*nouns, attributive adjectives*, and *prepositions*), these are also among the features of Biber_D1 for 'Informational' production. In a span of about 120 words, the fictional speech text uses 19 of the features associated with ‘Thought Presentation’ and only 6 of the ‘Description’ features. On the other hand, in roughly the same number of words, the narration passage uses only 4 of the ‘Thought Presentation’ features, and 40 of the ‘Description’ features.

Excerpt 2.
The Return of the Native, by Thomas Hardy, fictional speech: 119 words, narration: 120 words. The extract was abridged to show roughly the same number of words in speech and narration.

“I care a little, but not enough to break my rest,” *replied the **young man** languidly.* “No, all that’s past. I find there are two flowers where I thought there was only one. Perhaps there are three, or four, or any number as good as the first.... Mine is a **curious fate**. Who would have thought that all this could happen to me?”
*She interrupted with a **suppressed fire of** which either **love** or **anger** seemed an equally **possible issue**,* “Do you love me now?”
“Who can say?”
“Tell me; I will know it!”
[...]
“You know you can’t do otherwise, **for** all your **moods** and **changes**!” she answered defiantly. “Say what you will; try as you may; keep away **from** me all that you can – you will never forget me. You will love me all your **life** long. You would jump to marry me!”
[...]
*She did not answer. Its **tone** was indeed **solemn** and **pervasive**. **Compound utterances** addressed themselves **to** their **senses**, and it was possible to view **by ear** the **features of** the **neighbourhood**. **Acoustic pictures** were returned **from** the **darkened scenery**; they could hear where the **tracts of heather** began and ended; where the **furze** was growing stalky and tall; where it had been recently cut; in what **direction** the **fir-clump** lay, and how near was the **pit in** which the **hollies** grew; for these **differing features** had their **voices** no less than their **shapes** and **colours**.*

The overall factorial model was significant for Egbert_D2, $F(3,84) = 8.43$, $p < .001$, $R^2_{adjusted} = .20$. On this dimension there was no significant interaction effect ($p = .17$, $\eta^2 = .02$) or main effect of discourse level ($p = .14$, $\eta^2 = .02$). One of the reasons why this dimension is less distinguishing than others is a certain amount of overlapping features, e.g. both the positive and the negative features contain nouns, although different types of nouns. There was, however, a significant difference between the DNov and 19C corpora ($p < .001$, $\eta^2 = .19$). And especially the narration in Dickens is significantly different from narration in the 19C corpus, as shown in Figure 6. Dickens generally uses more features of concrete action and fewer features of abstract exposition than the 19C writers. This is true for both the fictional speech and narration written by Dickens. Excerpts 3 and 4 illustrate Dickens's use of *concrete nouns*, especially body part nouns. It seems that the greater difference between Dickens's narration and that of other 19th-century authors might also be partly due to specific patterns in which concrete nouns can occur. In Excerpt 4, the stretch of text *he said, smiting the table with his* ***fist***, is a special type of text outside of quotations. It interrupts the speech of a fictional character with contextual information. Such 'suspensions' tend to be useful places for body language information, and Dickens seems to have been able to make particularly good use of them (Mahlberg 2013). While for Dickens Figure 6 shows a slightly greater weight of narration along the 'Concrete Action' dimension, for the other authors the situation is exactly the other way round with narration scoring more highly on abstract exposition.

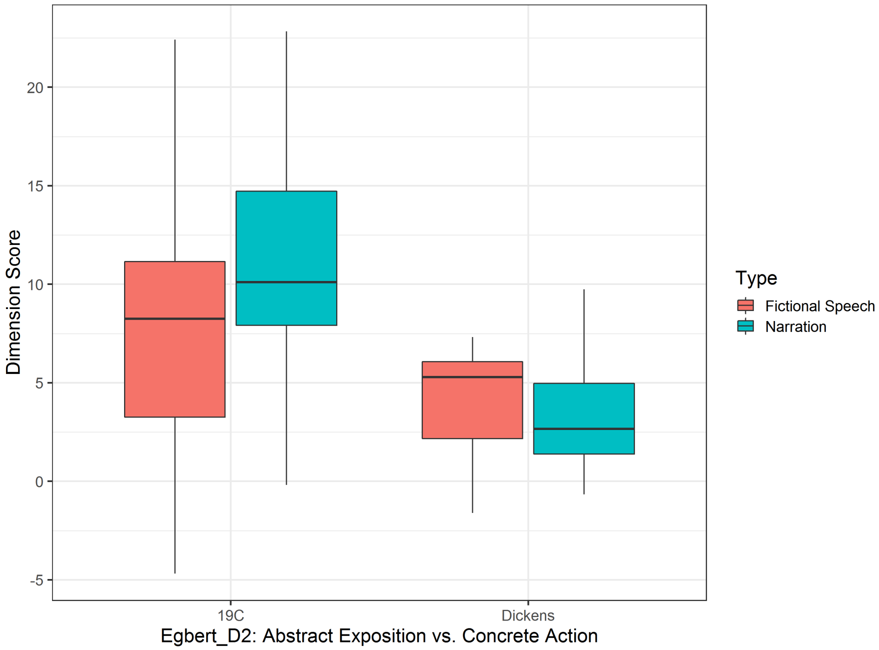

Figure 6. Boxplots of Egbert_D2 scores for narration and fictional speech in DNov and 19C. Positive scores are associated with Abstract Exposition; negative scores are associated with Concrete Action

Excerpt 3.
Our Mutual Friend, by Charles Dickens, fictional speech: 12 words, narration: 17 words

> ‘My **head**’s a bit light, and my **feet** are a bit heavy,’ *said old Betty, leaning her **face** drowsily on the **breast** of the woman who had spoken before.*

Excerpt 4.
Oliver Twist, by Charles Dickens, fictional speech: 54 words, narration: 31

> *The spirit of contradiction was strong in Mr. Grimwig's **breast**, at the moment; and it was rendered stronger by his friend's confident smile.*
> ‘No,’ *he said, smiting the table with his **fist***, ‘I do not. The boy has a new suit of clothes on his **back**, a set of valuable books under his **arm**, and a five-pound note in his pocket. He'll join his old friends the thieves, and laugh at you. If ever that boy returns to this house, sir, I'll eat my **head**.’

The differences between ‘Dialogue’ versus ‘Narration’ as described in Egbert_D3, are clearest. The features that account for these differences are similar to Biber_D1, in that they include first and second person pronouns, as well as present tense verbs for ‘Dialogue’ and ‘Involved’ production. Egbert_D3 emerged on the basis of a corpus containing only fiction. Thus, the functional interpretation of this dimension was based on fictional dialogue versus narration, rather than on speech and writing that is more or less involved or informational. This becomes particularly visible in the narration features from Egbert_D3, which contain third person pronouns and past tense verbs – features that are not accounted for in Biber_D1.

The overall model for this dimension was significant, $F(3,84) = 380.7$, $p < .001$, $R^2_{adjusted} = .93$. These results show that more than 93% of the variance in Egbert_D3 scores can be explained by the independent variables, primarily by the variable of discourse level ($p < .001$, $\eta^2 = .925$). There was some evidence for the presence of an interaction effect, but it was not strong enough to be significant ($p = .018$, $\eta^2 = .01$). There was no significant effect of author ($p = .18$, $\eta^2 = .002$). Still, Figure 7 suggests that Dickens uses more features of dialogue in his fictional speech than his contemporaries. A *t*-test confirmed that this difference is significant with a medium to large effect size ($p = .04$, $d = .75$).

Excerpt 5, taken from *Barnaby Rudge*, illustrates the use of highly colloquial and involved style of Dickens's fictional speech. In this excerpt, we have highlighted in gray three features associated with Dialogue (*1st person pronouns, 2nd person pronouns*, and *WH questions*) and bolded two features associated with Narration (*3rd person pronouns* and *past tense verbs*). In a span of about 120 words, the fictional speech text uses 20 of the features associated with ‘Dialogue’ and only 5 of the ‘Narration’ features. On the other hand, the associated narration passage uses none of the 3 ‘Dialogue’ features, and 27 of the ‘Narration’ features.

Excerpt 5.
Barnaby Rudge, by Charles Dickens, fictional speech: 122 words, narration: 128 words

‘Oh indeed!’ ***said*** *Mr Chester gaily.* ‘What else **did** you take from **her**?’

‘What else?’

‘Yes,’ ***said*** *the other, in a drawling manner, for* ***he was*** *fixing a very small patch of sticking plaster on a very small pimple near the corner of his mouth.* ‘What else?’

‘Well a kiss,’ ***replied*** *Hugh, after some hesitation.*

‘And what else?’

‘Nothing.’

‘I think,’ ***said*** *Mr Chester, in the same easy tone, and smiling twice or thrice to try if the patch* ***adhered*** – ‘I think there **was** something else. I have **heard** a trifle of jewellery spoken of – a mere trifle – a thing of such little value, indeed, that you may have forgotten it. Do you remember anything of the kind – such as a bracelet now, for instance?’

Hugh with a muttered oath ***thrust his*** *hand into* ***his*** *breast, and drawing the bracelet forth,* ***wrapped*** *in a scrap of hay,* ***was*** *about to lay it on the table likewise, when* ***his*** *patron* ***stopped his*** *hand and* ***bade*** *him* ***put*** *it up again.*

‘You **took** that for yourself my excellent friend,’ ***he said***, ‘and may keep it. I am neither a thief nor a receiver. Don’t show it to me. You had better hide it again, and lose no time. Don’t let me see where you put it either,’ ***he added****, turning away* ***his*** *head.*

‘You’re not a receiver!’ ***said*** *Hugh bluntly, despite the increasing awe in which* ***he held him.*** ‘What do you call THAT, master?’ *striking the letter with* ***his*** *heavy hand.*

The difference between the discourse levels of fictional speech and narration on this dimension is stark. This difference, combined with the other situational and linguistic differences, provides clear evidence that fictional speech and narration in 19th-century novels appear to behave in some ways almost like two entirely different registers.

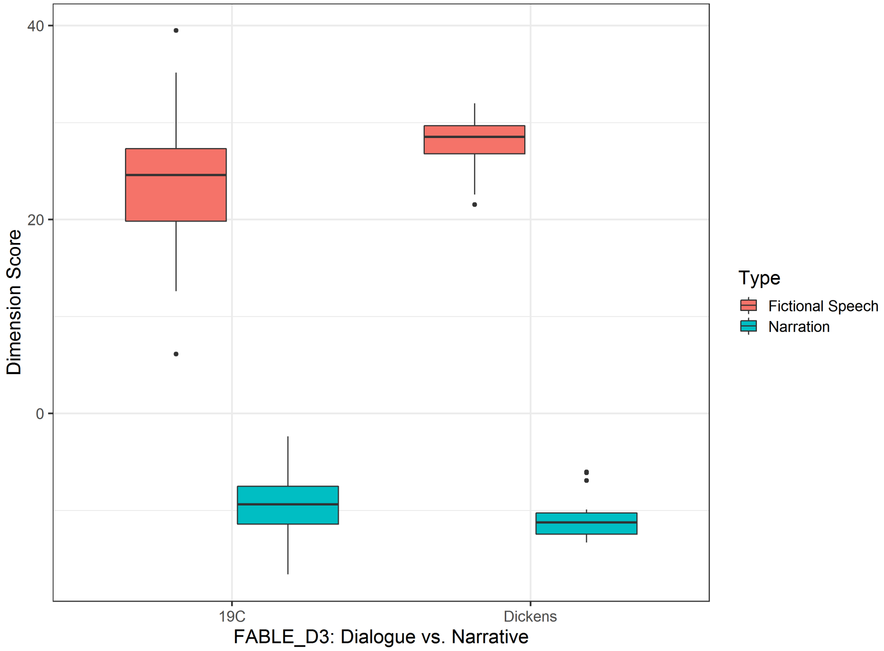

Figure 7. Boxplots of Egbert_D3 scores for narration and fictional speech in DNov and 19C. Positive scores are associated with "Dialogue"; negative scores are associated with "Narrative."

5. Conclusions

In this study we have used subcorpora of text within quotation marks and outside of quotation marks as approximations for speech and narration in novels. Through a comparison of the subcorpora based on dimension scores for individual texts, we identified register differences between speech and narration along the dimension of 'Involved' versus 'Informational' production. This dimension, Biber_D1, which previous research has shown to be robust over a large number of studies, also enabled us to compare fictional speech and narration with other registers. On Biber_D1, fictional speech sits between face-to-face conversation and spoken interviews, and narration takes a middle position between general fiction and biographies. In addition to the general register comparison based on Biber_D1, comparisons with three dimensions that are specific to narrative fiction enabled us to describe the register features of speech versus narration in more detail. Fictional speech and narration differ most clearly along the dimensions of 'Dialogue' versus 'Narration' (Egbert_D3) and 'Thought Presentation' versus 'Description' (Egbert_D1). Both show similarities with Biber_D1. The relevant features taken together clearly contrast the speaker-listener world in which fictional characters interact with one another with the perspective of the narrator who observes their interactions, as well as the immediacy of speech with the presentation of events in the past tense. Along the dimension of "Abstract Exposition" versus 'Concrete Action' (Egbert_D2), the distinction between speech and narration overall is less clear. But there is a distinction between DNov and 19C – and especially the narration in Dickens is significantly different from the narration in the corpus of other nineteenth-century writers. The features that describe the 'Abstract Exposition' versus 'Concrete Action' dimension include concrete nouns, such as body part nouns. Linking to previous research on body part nouns in Dickens, the use of such patterns seems to be associated with Dickens's particular skill of presenting fictional worlds and the characters therein. For the study of speech versus narration these nouns also point to the relationship between speech presentation and the contextual information of non-verbal communication that accompanies speech. In the present paper, we could only hint at this relationship, but it points to the wider significance of our findings.

The distinction between speech and narration that we were able to systematically describe in terms of register features, raises bigger questions for the study of registers, as well as for the linguistic study of fiction. As we pointed out in the introduction, Biber and Conrad (2009) stressed the complexity of fiction as a register because of the relationship between the real-world and the fictional-world context. Most register studies to date, however, have ignored this situational variation and complex context within fiction generally and novels in particular. Our suggestion here is a crucial step towards a better understanding of features reflecting context in fictional worlds. But what it will also need are complementary perspectives from other areas of research. In Text World Theory (Gavins 2007), for instance, linguistic devices and their functions for 'world-building' can be described. For Dickens in particular, corpus stylistic research has shown lexico-grammatical patterns that function to create fictional worlds. Such insights enable refinements of the ways in which we identify and measure register features. The features we have counted in this study, as in any register study, depend on the tagger that is used to identify them. In corpus linguistics in general, we use taggers that are well suited to capture general language features, but to capture the complexity of the fictional situational context we need tagsets that account for the building blocks of fictional worlds. Such tags would include patterns of body language to add precision to what our findings on concrete nouns point towards. In this way, we can provide greater detail to capture features of fictional worlds.

On the other hand, our findings also raise questions about the specific functions that the features we identified fulfil in their specific contexts. Based on the dimension scores, we can say that fictional speech shares features with face-to-face conversation. But more detailed analysis will be required to account for the lexico-grammatical patterns of first and second person pronouns, for instance, that are shared between fictional speech and face-to-face conversation. Pronouns alone are only part of the patterns. Again, this will help us to better describe how the context affects the linguistic functions we are able to observe through a study of register features.

One observation that we made in this paper refers to the need to consider context over time. Assuming that the real world context is of little relevance to register features in fiction would not only ignore the impact of the social context on the production and reception of fiction, but also functional change over time. Based on the results reported here, we provide a reference point based on Charles Dickens and novels of the 19th-century. To determine the extent to which our findings remain consistent or change over larger periods of time will require the replication of the method on other corpora (as well as the consideration of stylistic changes in the conventions of speech presentation).

Based on our results, we might conclude that a novel represents a sort of hybrid text composed of two registers. This runs counter to the traditional view that texts are nested within registers, but registers are not nested within texts. It also complements recent findings from corpus-based research on register variation online, as illustrated by Egbert et al. (2015). While one might suppose that intra-textual register variation is a new phenomenon restricted to the domain of the internet, Biber and Egbert (2018) hypothesize that hybrid registers may have existed all along and that we have simply been ignoring them (see Chapter 8). Our study provides evidence that examples of registers within texts are not restricted to register variation online and that novels provide particular challenges for a register approach.

The differences between the ‘texts within registers’ and ‘registers within texts’ paradigms, can be illustrated with the ‘texts are trees’ metaphor introduced by Egbert and Schnur (2018). In the ‘texts within registers’ case, we would expect texts from a particular register to share many characteristics, just as trees of the same species have much in common. However, the situation is not so simple in the ‘registers within texts’ case, where the text contains segments of discourse from more than one register. We might compare this situation to a grafted tree. Just as a horticulturalist can carefully graft stems from multiple tree species onto a single rootstock, we argue that authors and speakers have the ability to graft language from more than one register into a single text. The dichotomy between narration and speech is not unique to novels. Another similarly ‘grafted’ register that has not yet been discussed in these terms is news reporting. Newspaper articles combine narrative prose with direct quotations from sources, interviewees, and eye witnesses. While functions of both the prose and the quotations can be interpreted in terms of their contribution to a story's news values, situational differences between report and quotations suggest that we are likely to find linguistic differences, too.

If we regard novels as composed of many segments of narration and fictional speech that are intertwined – or 'grafted' – into a single text, this presents a challenge for the way we operationalize the unit of analysis in research on fiction. This situation is different from the 'hybrid texts' from Egbert et al. (2015) where the language in many of the web documents had the characteristics of more than one register, or fell in between more than one register, but without necessarily delimitable units. The situation is also not as simple as the move structures (e.g. IMRD in research articles) identified by genre researchers because in most cases these moves only occur once in a text, in a predictable order. For novels, there is a different challenge in that segments of the text can be classified into (at least) two situationally and linguistically distinct discourse level (register) categories – fictional speech or narration – that are interspersed throughout a text.

While a register studies approach enabled us to treat speech and narration in novels separately, the approach does not provide detail on how features are associated across the two different data sets. It creates an artificial situation in which segments of narration and fictional speech are collected in separate corpora and so placed adjacent to each other in a single text that is not coherent and may not even be comprehensible. Our results for the 'Abstract Exposition' versus 'Concrete Action' dimension suggest that the relationship that in reality exists between the two data sets needs to be accounted for, too. While speech and narration are less clearly distinguishable along this dimension, the difference becomes clearer for Dickens versus the corpus of other nineteenth-century writers. Focusing on speech and narration within texts by the same author can better highlight differences that are the result of systematic, stylistic patterns, as we have argued for examples of body language that accompany speech. So while our approach in this study reveals fundamental differences between speech and narration from a register studies point of view, it needs to be seen as complementing detailed textual analyses.

Overall, our study has shown that engaging with the complexity of fiction does not only contribute to a more systematic understanding of the features that build fictional worlds, but also to further development of approaches in corpus and register studies. We hope the findings in this study serve as a springboard for future research on the complex nature of fictional texts, as well as the complexities of other texts with hybrid register characteristics.

Funding

This work was supported by the Arts and Humanities Research Council, grant number AH/P504634/1.

References

Axelsson, K. (2009). Research on fiction dialogue: Problems and possible solutions. In A. H. Jucker, D. Schreier and M. Hundt (Eds.), *Corpora: Pragmatics and discourse*. Amsterdam: Rodopi, 189–201. https://doi.org/10.1163/9789042029101_011

Biber, D. & Conrad, S. (2009). *Register, Genre, and Style*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108686136

Barlow, M. (2016). WordSkew. Linking corpus data and discourse structure. *International Journal of Corpus Linguistics*, 21(1), 105–115. https://doi.org/10.1075/ijcl.21.1.05bar

Bhatia, V. K. (2014). *Analysing genre: Language use in professional settings*. London: Routledge. https://doi.org/10.4324/9781315844992

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511621024

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, 8(4), 243–257. https://doi.org/10.1093/llc/8.4.243

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.23

Biber, D. (2014). Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Languages in Contrast*, 14(1), 7–34. https://doi.org/10.1075/lic.14.1.02bib

Biber, D., Connor, U. & Upton, T. A. (2007). *Discourse on the move: Using corpus analysis to describe discourse structure*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.28

Biber, D., & Egbert, J. (2018). *Register variation online*. Cambridge: Cambridge University Press.

Biber, D., & Finegan, E. (1994). Multi-dimensional analyses of authors' styles: Some case studies from the eigtheenth century. In D. Ross & D. Brink (Eds.), *Research in humanities computing* (pp. 3–17). Oxford: Oxford University Press.

Biber, D., & Finegan, E. (2001). Diachronic relations among speech-based and written registers in English. In S. Conrad & D. Biber (Eds.), *Variation in English: Multi-dimensional studies* (pp. 66–83). New York, NY: Routledge.

Burrows, J. F. (1987). *Computation into criticism: A study of Jane Austen's novels and an experiment in method*. Oxford: Clarendon.

Clarke, I., & Grieve, J. (2017). Dimensions of abusive language on Twitter. In *Proceedings of the First Workshop on Abusive Language Online* (pp. 1–10), Vancouver, Canada (August 4, 2017). https://doi.org/10.18653/v1/W17-3001

Conrad, S. & Biber, D. (2001). Variation in English: Multi-dimensional studies. New York: Routledge.

Culpeper, J. & Kytö, M. (2010). *Early modern English dialogues: Spoken interactions as writing*. Cambridge: Cambridge University Press.

Davies, M., & Gardner, D. (2010). *A frequency dictionary of contemporary American English*. New York, NY: Routledge.

De Haan, P. (1996). More on the language of dialogue in fiction. *ICAME Journal*, 20, 23–40.

Egbert, J. (2012). Style in nineteenth century fiction: A multi-dimensional analysis. *Scientific Study of Literature*, 2(2), 167–198. https://doi.org/10.1075/ssol.2.2.01egb

Egbert, J. (2015). Publication type and discipline variation in published academic writing: Investigating statistical interaction in corpus data. *International Journal of Corpus Linguistics*, 20(1), 1–29. https://doi.org/10.1075/ijcl.20.1.01egb

Egbert, J., Biber, D., & Davies, M. (2015). Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology*. 66(9): 1817–1831. https://doi.org/10.1002/asi.23308

Egbert, J., & Schnur, E. (2018). Missing the trees for the forest: The role of the text in corpus and discourse analysis. In A. Marchi & C. Taylor (Eds.), *Corpus approaches to discourse: A critical review*. New York, NY: Routledge.

Flowerdew, L. (2003). A combined corpus and systemic-functional analysis of the problem-solution pattern in a student and professional corpus of technical writing. *TESOL Quarterly*, 37(3), 489–511. https://doi.org/10.2307/3588401

Flowerdew, L. (2008). *Corpus-based analyses of the problem-solution pattern: A phraseological approach*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.29

Gavins, J. (2007). *Text world theory. An introduction*. Edinburgh: Edinburgh University Press. https://doi.org/10.3366/edinburgh/9780748622993.001.0001

Gray, B. (2011). Exploring academic writing through corpus linguistics: When discipline tells only part of the story (Unpublished doctoral dissertation). Northern Arizona University.

Hoey, M. (2001). *Textual interaction: An introduction to written discourse anlaysis*. London: Routledge.

Hubbard, E. H. (2002). Conversation, characterization and corpus linguistics: Dialogue in Jane Austen's *Sense and Sensibility*. *Literator*, 23(2), 67–85. https://doi.org/10.4102/lit.v23i2.331

Leech, G. & Short, M. (2007). *Style in fiction: A linguistic introduction to English fictional prose*. 2nd edition. Harlow: Pearson.

Kyto, M., Rudanko, J., & Smitterberg, E. (2000). Building a bridge between the present and the past: A corpus of 19th-century English. *ICAME journal*, 24, 85–98.

Mahlberg, M. (2013). *Corpus stylistics and Dickens's fiction*. London: Routledge.https://doi.org/10.4324/9780203076088

Mahlberg, M & Stockwell, P (2016). Point and CLiC: Teaching literature with corpus stylistic tools. In M. Burke, O. Fialho & S. Zyngier (Eds.), *Scientific approaches to literature in learning environments* (pp. 251–267). Amsterdam: John Benjamins.

Mahlberg, M., Stockwell, P., de Joode, J., Smith, C., & O'Donnell, M. Brook. (2016). CLiC Dickens – Novel uses of concordances for the integration of corpus stylistics and cognitive poetics. *Corpora*, 11(3), 433–463. https://doi.org/10.3366/cor.2016.0102

Mahlberg, M., Wiegand, V., Stockwell, P., & A. Hennessey. (2019a). Speech-bundles in the 19th-century English Novel. *Language and Literature*, 28(4), 326–353. https://doi.org/10.1177/0963947019886754

Mahlberg, M., Wiegand, V., Lentin, J., et al. (2019b). CLiC User Guide v2.0.1 documentation. Retrieved from: <http://clic.bham.ac.uk/docs/> (19, August 2019).

Mooi, E., & Sarstedt, M. (2010). Data. In R. Stahlbock, S. F. Crone, & S. Lessmann (Eds.), *A concise guide to market research* (pp. 25–44). Berlin: Springer. https://doi.org/10.1007/978-3-642-12541-6_3

Norris, J. (2015). Discriminant analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research*. London: Routledge.

O'Donnell, M.B., Scott, M., Mahlberg, M., & Hoey, M. (2012). Exploring text-initial words, clusters and concgrams in a newspaper corpus. *Corpus Linguistics and Linguistic Theory*, 8(1), 73–101.

Oostdijk, N. (1990). The language of dialogue in fiction. *Literary and Linguistic Computing*, 5(3), 235–241. https://doi.org/10.1093/llc/5.3.235

Page, N. (1988). *Speech in the English novel.* Atlantic Highlands, NJ: Humanities Press International. https://doi.org/10.1007/978-1-349-19047-8

Paltridge, B. (1996). Genre, text type, and the language learning classroom. *ELT Journal*, 50(3), 237–243. https://doi.org/10.1093/elt/50.3.237

Poulsen, J., & French, A. (2008). Discriminant function analysis. Retrieved from: <http://userwww.sfsu.edu/~efc/classes/biol710/discrim/discrim.pdf> (23 October 2017).

Rosso, M.A. (2008). User-based identification of web genres. *Journal of the American Society for Information Science and Technology*, 59(7), 1053–1072. https://doi.org/10.1002/asi.20798

Römer, U. (2010). Establishing the phraseological profile of a text type: The construction of meaning in academic book reviews. *English Text Construction*, 3(1), 95–119. https://doi.org/10.1075/etc.3.1.06rom

Scott, M., & Tribble, C. (2006). *Textual patterns. Key words and corpus analysis in language education*. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.22

Semino, E., & Short, M. (2004). *Corpus stylistics: Speech, writing and thought presentation in a corpus of English writing*. London: Routledge. https://doi.org/10.4324/9780203494073

Staples, S., & Biber, D. (2015). Cluster analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research*. London: Routledge.

Sorensen, K. (1989). Narration and speech-rendering in Dickens. *Dickens Quarterly*, 6(4), 131.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Tabachnick, B.G., & Fidell, L.S. (2007). *Using multivariate statistics*. Boston, MA: Pearson Education.

Address for correspondence

Jesse Egbert
Northern Arizona University
English Department
Box 6032
Flagstaff, AZ 86011
USA

Jesse.Egbert@nau.edu

Co-author information

Michaela Mahlberg
University of Birmingham
M.A.Mahlberg@bham.ac.uk