Fiction—One Register or Two?
Narrative and Fictional Speech in Dickens's Novels

Jesse Egbert (Northern Arizona University, USA) and Michaela Mahlberg (University of Birmingham, UK)

Few registers have perplexed linguists more than narrative fiction. At times, fictional prose behaves like a historical or biographical narrative, and at other times it creates fictional people who interact – often through fictional speech represented in a variety of forms (e.g. Semino & Short 2004). Fiction is typically seen as a linguistic variety shaped by creative use of language. As a consequence, findings from fiction do not always fit general patterns and might be discarded as outliers or idiosyncratic usage. Equally, when the focus is on patterns shared across the register, it seems fiction displays features that are less striking or contrasting than features of other registers. De Haan (1996, p. 38) observes: "fiction takes sort of a middle position between more formal writing on the one hand, and face-to-face conversation on the other". One plausible explanation for this pattern is that a novel is made of different discourse levels (e.g. fictional speech and narration), but linguistic counts are taken across the whole text.

When non-fictional registers are compared, Biber's (1988) Dimension 1 (Involved versus Informational Production) very clearly shows a distinction between Face-to-Face Conversations, (M = 35.3), and the narrative register of Biographies (M = -12.4). Assuming that fictional speech mirrors actual spoken language, these findings suggest that treating fictional speech and narration as a single text will simply produce an 'average' of the dialogic and narrative features and will not necessarily represent the linguistic characteristics of either. Nevertheless, few scholars have attempted to distinguish between fictional speech and narrative when analyzing linguistic patterns in novels. This is to some extent due to a long-held belief that fictional speech is fundamentally different from 'real' spoken language (Page 1988) precisely because it is part of the narrative text as a whole.

Only a small number of studies have focused exclusively on fictional speech (e.g. Burrows, 1987; De Haan, 1996; Hubbard, 2002, Oostdijk, 1990). Others have analyzed samples of fictional speech 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). Some recent research has focused on analyzing what Lambert (1981) calls the 'suspended quotation', or the interruption of fictional speech by narration e.g. Mahlberg, 2013; Mahlberg & Smith, 2012; Mahlberg, Smith, & Preston, 2013). Although there have been calls to reconsider the status of fiction as a single register (see, e.g., De Haan, 1996, 38-39; Egbert, 2012, 189), no research to date has approached fiction in this way.

In the most basic sense, the text of a novel can be divided into two parts based on formal features: fictional speech - within quotes - and narrative - outside of quotes (Mahlberg et al. forthcoming). In order to get a better understanding of the linguistic features of fictional narrative, the goal of this study is to carry out a comprehensive

linguistic description of the style of Charles Dickens, with a specific focus on the differences between narrative and fictional speech. Specifically, we aim to answer three questions in this paper:

1. In what ways does Dickens's style differ between narrative and fictional speech?
2. In what ways does Dickens's narrative and fictional speech change over time?
3. Can the narrative and fictional speech in Dickens's novels be grouped in a meaningful way?

Methods

In this study we used a corpus comprising all 15 novels written by Charles Dickens (DNov), each of which was divided into two texts, one that contains all of the fictional speech (text inside of quotations) and one that contains all of the narrative (text outside of quotations), for a total of 30 texts. The DNov corpus contains a total of 3.8 million words of running text, with about 2.5 million words (66%) of narrative and 1.3 million words (33%) of fictional speech.

Each of the 30 texts in the DNov corpus was tagged using the Biber Tagger and processed to calculate normed rates of occurrence for 150+ linguistic features using Biber's TagCount program. We used these counts to compute scores for each text on four dimensions from two previous Multi-Dimensional analyses. The first dimension comes from Biber (1988), and will be referred to as Biber_D1 (Involved versus Informational Production). This dimension was based on a corpus of texts from the LOB and London-Lund corpora. The other dimensions used in this study come from a description of stylistic variation in the FABLE (Fiction of America and Britain from the Late Eighteen hundreds) corpus (Egbert, 2012). The three dimensions from that study are FABLE_D1 (Thought Presentation versus Description), FABLE_D2 (Abstract Exposition versus Concrete Action), and FABLE_D3 (Dialogue versus Narrative). 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 FABLE studies.

In order to answer RQ1, we compared the mean dimension scores for the speech and narrative sub-corpora along each of the four dimensions. These comparisons were made using *t*-tests to test for statistical significance and Cohen's *d* as a measure of effect size. RQ2 was answered using Pearson's correlations between time (year of publication) and each of the three dimension scores from the FABLE study. The narrative and speech sub-corpora were analyzed separately. This yielded 6 correlation coefficients (3 dimensions x 2 text types). RQ3 was answered using a cluster analysis of the thirty texts based on scores for the three dimensions in this study. Hierarchical clustering was used to determine the ideal number of clusters to extract, after which *k*-means clustering was used to determine cluster membership of each text (see, e.g., Staples & Biber, 2015).

Results

The analysis for RQ1 resulted in significant differences and extremely large effect sizes between speech and narrative along Biber_D1 ($p < .0001$, $d = 9.47$), FABLE_D1 ($p < .0001$, $d = 6.49$) and FABLE_D3 ($p < .0001$, $d = 14.44$). FABLE_D2 on the other hand showed no significant difference and a very small effect size ($p = .59$, $d = .20$). These results reveal that Dickens's speech uses more linguistic features associated with involvement (Biber_D1), thought presentation (FABLE_D1) and, unsurprisingly, dialogue (FABLE_D3). Dickens's narrative, on the other hand, uses more features associated with an informational focus (Biber_D1), description (FABLE_D1), and narrative (FABLE_D3). Figure 1 contains boxplots that display each of these differences.

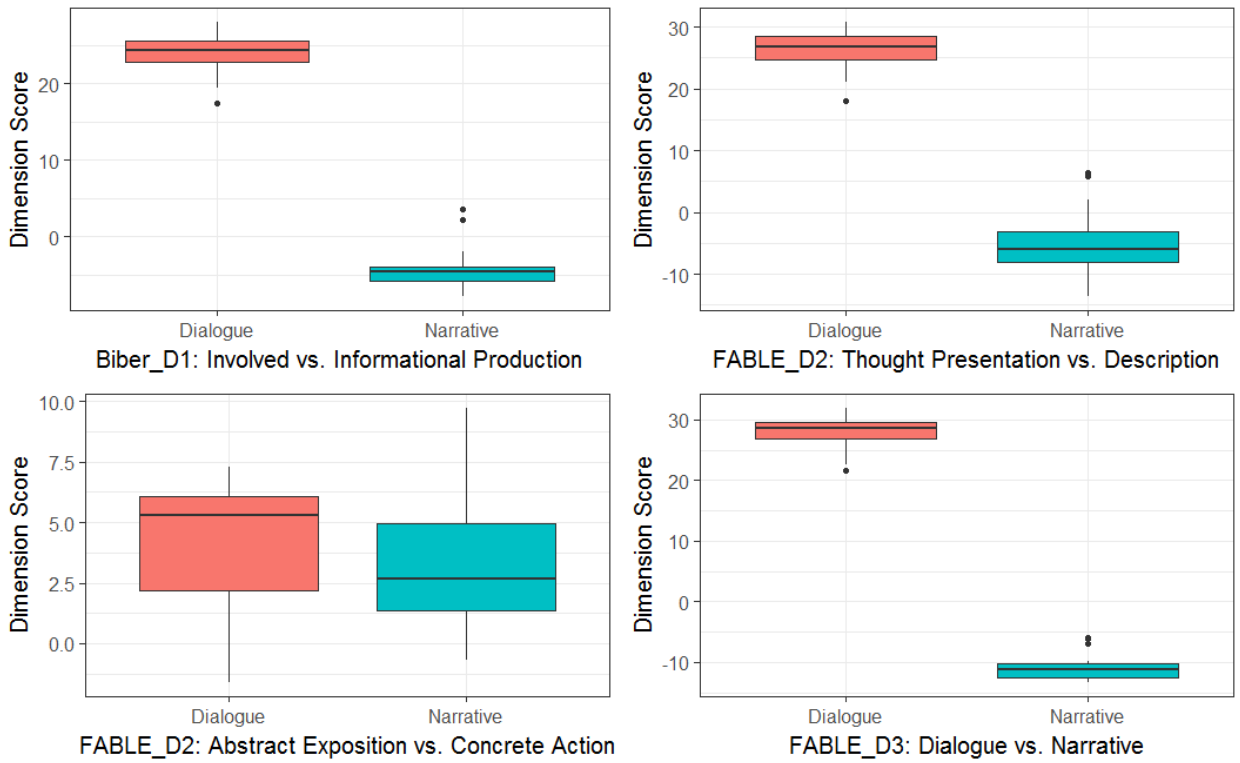

Figure 1. Dimension scores results for fictional speech and narrative texts.

In answer to RQ2, there were no statistically significant or noteworthy correlations between time and dimension scores in Dickens's fictional speech. While there is a great deal of linguistic variation in Dickens's fictional speech, this cannot be attributed to the variable of time using the linguistic measures used here. In the narrative sub-corpus, on the other hand, there was a strong and significant negative correlation between time and FABLE_D2 ($r = -.82$, $p = .0002$). This shows that more than 67% of the variance in FABLE_D2 scores can be accounted for by the variable of publication year (see Figure 2). This shows that over time Dickens's narrative prose used more features related to concrete action and fewer features related to abstract exposition.

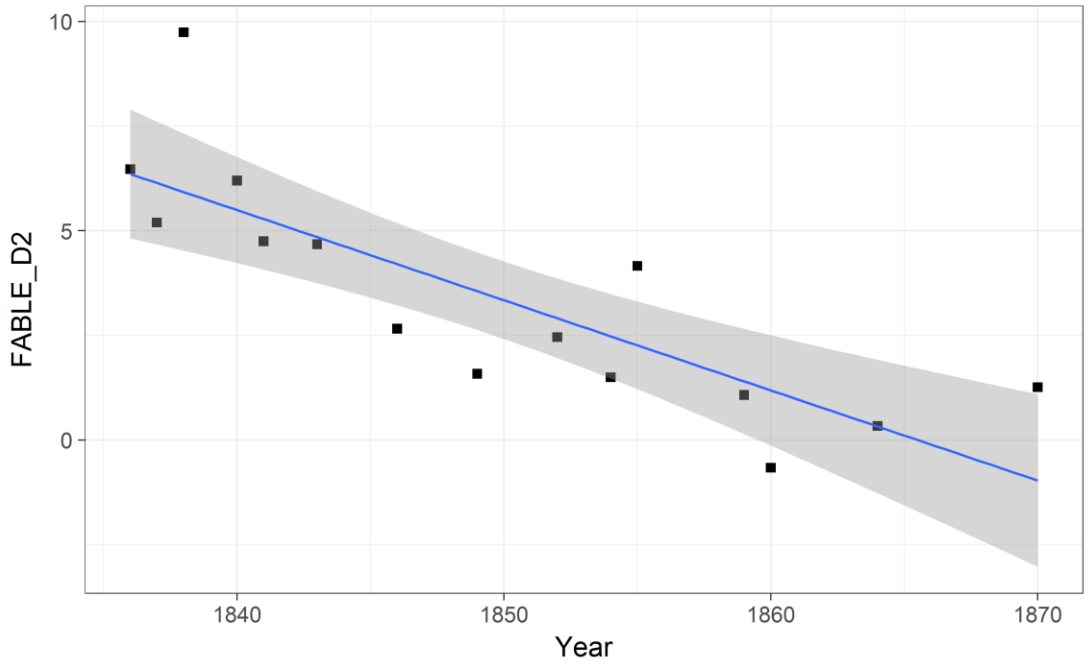

Figure 2. Relationship between year of publication and FABLE_D2 scores for narrative (shaded area represents 95% confidence region)

The cluster analysis revealed two clear top level clusters, with 15 texts in each that correspond perfectly to the narrative vs. speech categories. There are five clusters on the next level of the hierarchy that reveal additional groupings within the narrative and speech clusters. These reveal that although the most important predictor of Dickens's style is the speech vs. narrative distinction, there are other stylistic features that can group his novels in meaningful ways.

Conclusion

This study represents an important step forward for research in corpus stylistics and literary studies. We show stark differences between the narrative prose produced by Dickens and his representation of character speech. We also reveal a diachronic change in Dickens's narrative writing style that is not found in his fictional speech, suggesting that this pattern could not have been revealed had we not divided the novels into narrative and speech. Finally, we show that while the narrative-speech distinction is the strongest predictor of linguistic variation, there are other sub-clusters among Dickens's novels within the clusters of narrative and speech that we plan to explore further.

The approach we have adopted here of analyzing narrative and fictional speech separately offers much to the study of novels, a register that has presented major challenges to corpus linguists and literary scholars. Our results strongly suggest that the narrative-speech distinction has to be accounted for in future research, especially quantitative corpus stylistics. This can be accomplished using our method of treating novels as two texts rather than one, or by some other means. This paper also reveals

new insights into the complex literary style of Charles Dickens and has wider implications for the conceptualization of narrative fiction.

References

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

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

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

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

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.

Hubbard, E.H. 2002. Conversation, characterization and corpus linguistics: Dialogue in Jane Austen's *Sense and Sensibility. Literator, 23*(2), 67-85.

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.

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

Mahlberg, M., Smith, C., & Preston, S. (2013). Phrases in literary contexts: Patterns and distributions of suspensions in Dickens's novels. *International Journal of Corpus Linguistics, 18*(1), 35-56.

Mahlberg, M., Stockwell, P., de Joode, J., Smith, C., O'Donnell, M. Brook, (forthcoming) CLiC Dickens – Novel uses of concordances for the integration of corpus stylistics and cognitive poetics, *Corpora* 11.3.

Oostdijk, N. 1990. The language of dialogue in fiction. *Literary and Linguistic Computing*, *5*(3), 235-241.

Page, N. 1988. Speech in the English Novel. Atlantic Highlands: Humanities Press International.

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

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