Reading Dickens's characters: Employing psycholinguistic methods to investigate the cognitive reality of patterns in texts

Michaela Mahlberg
University of Nottingham, UK

Kathy Conklin
University of Nottingham, UK

Marie-Josée Bisson
Loughborough University, UK

Abstract

This article reports the findings of an empirical study that uses eye-tracking and follow-up interviews as methods to investigate how participants read body language clusters in novels by Charles Dickens. The study builds on previous corpus stylistic work that has identified patterns of body language presentation as techniques of characterisation in Dickens (Mahlberg, 2013). The article focuses on the reading of 'clusters', that is, repeated sequences of words. It is set in a research context that brings together observations from both corpus linguistics and psycholinguistics on the processing of repeated patterns. The results show that the body language clusters are read significantly faster than the overall sample extracts which suggests that the clusters are stored as units in the brain. This finding is complemented by the results of the follow-up questions which indicate that readers do not seem to refer to the clusters when talking about character information, although they are able to refer to clusters when biased prompts are used to elicit information. Beyond the specific results of the study, this article makes a contribution to the development of complementary methods in literary stylistics and it points to directions for further subclassifications of clusters that could not be achieved on the basis of corpus data alone.

Corresponding author:
Michaela Mahlberg, University of Nottingham, University Park, Nottingham NG7 2RD, UK.
Email: michaela.mahlberg@nottingham.ac.uk

1 Introduction

This article builds on corpus stylistic research on character information in the novels of Charles Dickens. It aims to take corpus stylistic work further by bringing in complementary psycholinguistic evidence for claims about stylistically relevant linguistic phenomena. Literary stylistics in general is concerned with the analysis of a text both as a sample of language and as a work of art (cf. Leech & Short 2007: 11f.). Recent work in corpus stylistics has helped to focus attention on the linguistic patterns that are considered to be of literary relevance. At the same time, corpus stylistic findings raise questions about the extent to which readers are actually aware of textual patterns that are interpreted in terms of their literary functions. The way in which real readers approach a text receives particular attention in cognitive stylistics where textual meaning is seen as being created in the course of the reading process. We want to argue that psycholinguistic methods can add a valuable dimension to the interpretation of corpus stylistic findings. Such methods can be used to look empirically at the processing of linguistic units. Hence they can help to link findings on linguistic patterns as identified by corpus methods and claims about the creation of meaning in the mind of the reader. This article uses data from texts by Charles Dickens. This is because we build on earlier work in corpus stylistics (specifically Mahlberg, 2013), but also because Dickens's characters are thought to have noticeable effects on readers, as reflected by the continuing popularity of Dickens. The article focuses on patterns of body language presentation. Our aim is to investigate relationships between repeatedly occurring patterns of body language presentation and the way such patterns are read. The link that we make between corpus findings and literary reading has not been systematically studied yet, so our second aim is to develop an approach for further studies in this area.

The article begins with an outline of the corpus stylistic background to the present study (Section 2). Section 3 highlights links between corpus linguistic and psycholinguistic approaches more generally. We introduce our methodology in Section 4 and present the results in Section 5 before concluding the article with an evaluation of our findings and suggestions for future work.

2 Body language presentation from a corpus stylistic point of view

Recent approaches to the linguistic description of literary characters see characterisation as a process whereby impressions of fictional people are created in the mind of the reader (Culpeper, 2001). Both top-down and bottom-up processes play a role in characterisation. Top down processing means that readers bring their knowledge of people in the real world to literary texts and model fictional characters to some extent on real-world examples (Stockwell, 2009). Real-world knowledge interacts with bottom-up processing whereby text-driven meanings are created during the reading process. Hence, textual

cues provide character information and at the same time trigger relevant real-world knowledge. This conceptualisation of characterisation can be seen as part of a more general approach to the language in literary texts: any configuration in a literary text shows features of both literary and non-literary language (Carter, 2004). Even textual examples that are perceived as highly creative need to incorporate some degree of commonly occurring linguistic patterns in order to create meaning for a reader. For the presentation of character information this means there will be patterns that do not strike readers as unusual, as well as patterns that draw attention to unusual character features. However, it still seems to be useful to refer to 'literary' patterns of character presentation that are part of the creation of fictional worlds. We use the term 'literary' here to refer to patterns in a text that the reader chooses to read in a literary way (cf. Carter, 2004: 69). Such patterns that are used to create fictional people differ to some extent from patterns used to talk about people in the real world in texts such as newspaper articles or casual conversations.

The patterns we want to concentrate on in the present study are based on the findings in Mahlberg (2013) for 'clusters' in Dickens. Clusters are repeatedly occurring sequences of words. They are defined purely on the basis of formal criteria, that is, the number of words considered in the repetition and the minimum frequency with which a cluster has to occur in a corpus in order to be included in a study. Clusters are not necessarily structurally complete phrases, as illustrated by the five-word cluster *his eyes fixed on the.* Other terms that are used to refer to repeated sequences of words are, for instance, '*n*-grams' or 'lexical bundles'. The choice of term is often motivated by the research context and can also entail specific requirements for frequency cut-offs. A detailed discussion of the various approaches is beyond the scope of the present article, but can be found in Mahlberg (2013).[1] Clusters are of interest in corpus linguistics because re-occurring sequences of words tend to be associated with linguistic functions in particular types of discourses. Table 1 shows the frequency of occurrence of *his eyes fixed on the* across some of the sections in the BNC (British National Corpus). The cluster's function of describing body language of fictional characters is clearly reflected by its distribution across the different sections of the corpus. Most hits are in 'Fiction and verse'.

In this article, we focus on patterns that are relevant to the presentation of body language in literary texts and their functions as textual cues for characterisation. Mahlberg (2013) argues that patterns of body language presentation function on a cline of contextualising and highlighting patterns. Patterns that are 'contextualising' present character information in an inconspicuous way as part of a larger textual picture, while patterns

Table 1. Distribution of his eyes fixed on the across some of the sections in the BNC.

Text type	Hits	Frequency per million words
Fiction and verse	20	1.24
Non-academic prose and biography	1	0.04
Academic prose	0	0
Spoken conversation	0	0
Newspapers	0	0

Table 2. Body language clusters with their frequencies in DNC and 19C and the number of texts in which they occur.[4]

Body part clusters	DNC	DNC texts	19C	19C texts
and his nose came down	7	1	0	0
his eyes fixed on the	17	9	14	6
his hand to his forehead	16	5	2	2

with 'highlighting' functions give prominence to character information. Such prominence can be achieved by describing body language as habitual, commenting on its significance and/or associating a character with a specific cluster. An example is the cluster *and his nose came down* that is associated with Rigaud in *Little Dorrit*. When this sequence of words is used for the first time in the opening chapter of the novel, it is part of a description of Rigaud in which the narrator emphasises features of the outward appearance of this character and his habitual behaviour:

> (1) When Monsieur Rigaud laughed, a change took place in his face, that was more remarkable than prepossessing. His moustache went up under his nose, *and his nose came down* over his moustache, in a very sinister and cruel manner.
>
> (Charles Dickens, *Little Dorrit*, Book 1, Chapter 1)[2]

The cluster is repeated seven times in the novel and only occurs with Rigaud. Another aspect of its highlighting function is that it does not occur in any other Dickens novel or in the 19th-century reference corpus used in Mahlberg (2013), as shown in Table 2, where DNC refers the Dickens Novels Corpus and 19C to the reference corpus.[3]

The two other examples in Table 2, *his eyes fixed on the* and *his hand to his forehead*, are less text-specific. They occur in Dickens's novels as well as in the reference corpus and can serve to illustrate different degrees on the contextualising and highlighting cline. The cluster *his eyes fixed on the* illustrates a contextualising function in *Dombey and Son*. Example (2) shows how the cluster refers to Mr Dombey. The description of him sitting *with his eyes fixed on the table* is an external reflection of his thought processes that make him forget the situation around him. In addition to the function of the cluster that is visible in the short example, the fact that the cluster just occurs once to describe Mr Dombey is also part of its contextualising nature.

> (2) He sat with *his eyes fixed on the* table, so immersed in thought, that a far heavier tread than the light foot of his child could make, might have failed to rouse him.
>
> (Charles Dickens, *Dombey and Son*, Chapter 18)

> (3) Thus it had come about, that Mr Twemlow had said to himself in his lodgings, with *his hand to his forehead*: 'I must not think of this. This is enough to soften any man's brain,' – and yet was always thinking of it, and could never form a conclusion.
>
> (Charles Dickens, *Our Mutual Friend*, Book 1, Chapter 2)

In example (3), the cluster *his hand to his forehead* illustrates a pattern with a highlighting function. Concentrating on the two short extracts alone, examples (2) and (3) are similar in that in both the body language is a reflection of a mental process. Both Dombey and Twemlow are thinking about something. For Dombey the body language reflects the intensity of his thought process and for Twemlow the confusing nature of his thoughts. For Twemlow, however, the same body language description occurs repeatedly: the cluster *his hand to his forehead* occurs 11 times in *Our Mutual Friend* and each time refers to the same character. While *his hand to his forehead* is associated with Twemlow in *Our Mutual Friend* it is still less text-specific than *and his nose came down.*

With regard to the functional interpretation of clusters the important point to note here is the following: corpus linguistic findings, as outlined in Mahlberg (2013), show that clusters can reveal patterns of body language presentation that cover both highlighting and contextualising functions. However, in literary criticism it is typically the body language patterns with highlighting functions that receive attention. Specifically for Dickens, critics see 'character tags' as a key technique of Dickens's creation of characters (e.g. Brook, 1970; Paroissien, 2000). Mahlberg (2013) argues that a crucial contribution of corpus methods is to bring contextualising patterns of body language to the attention of the analyst too, and to stress that body language patterns function on a cline of contextualising and highlighting functions. Although especially contextualising patterns may not have attracted much attention from the critics, they play an important role in the creation of fictional characters and provide the more general textual building blocks of narrative fiction. They are crucial to the 'literary competence' of the reader in the sense that readers are familiar with typical patterns of presenting body language and can draw on background knowledge of the way in which fictional characters are constructed. For this aspect of the literary competence it is important to take the distribution of contextualising patterns across different texts into account. If patterns are not restricted to a particular text but occur more widely, they are more likely to be encountered by readers and so contribute to building the readers' literary competence.[5]

The observations for literary patterns of body language description are in line with more general findings in corpus linguistics. Corpus linguistic methods focus the analyst's attention on repeatedly occurring patterns, and thus the functions associated with these patterns. Very frequent or commonly occurring patterns have often been overlooked in language descriptions because they seem to escape the language user's intuition. While corpus methods can reveal patterns and raise questions about the extent to which readers might be aware of such patterns, psycholinguistic methods are necessary to actually investigate how real readers perceive patterns. In this article, we aim to make suggestions towards a methodology for bringing corpus findings and experimental research together. In the present study we focus on patterns that occur across different texts and that can fulfil contextualising functions.

3 Textual patterns from corpus linguistic and psycholinguistic perspectives

A crucial link between corpus and psycholinguistic findings is the fact that repeatedly occurring patterns require less processing effort. From a corpus linguistic point of view, the focus on clusters is set in a wider context, where the emphasis has shifted from the

word as a unit of meaning to extended units of meanings that span several words. The relationship between meaning and form is captured, for instance, in the concept of 'collocations', that is, the tendency of words to co-occur as in *hands* co-occurring with *pockets* across the DNC and 19C. Sinclair (1991) explains the existence of lexico-grammatical patterns in the language through the operation of the idiom principle. According to the idiom principle, pre-fabricated phrases are stored as units in the brain. The outputs of the idiom principle are complemented by phrases and structures that can be accounted for by the open choice principle: the principle that accounts for the rules and regularities of the language independent of lexical choices. Some pre-fabricated phrases may result from phenomena in the real world that repeatedly occur together, such as *fish and chips*. However, there are also patterns that less clearly relate to concrete entities in the world and are better described in terms of discourse functions, for example, *all that sort of thing*. Still, in both cases, the repeated occurrences reflect repeated communicative needs of a specific register.

From a psycholinguistic point of view, the existence of pre-fabricated phrases can be seen in terms of an increase in processing efficiency because the phrases are represented in the brain and do not need to be put together compositionally. The processing effort of phrases seems to be related to the frequency of occurrence of the phrases (cf. e.g. Arnon and Snider, 2010; Bannard and Matthews, 2008; Siyanova-Chanturia et al., 2011). Processing effort plays a role both for speakers and hearers. Pre-fabricated phrases relieve processing pressure for the speaker and at the same time reduce effort for the hearer. Wray (2008: 69) points out that formulaic expressions can therefore also be used by the speaker to have specific effects on the hearer. Greetings or forms of address are examples of pre-fabricated phrases by which the speaker can influence the hearer's perception of them.

Investigations into the processing effort of phrases often touch upon the 'magical number seven', which is sometimes called *Miller's Law*, and refers to Miller's seminal finding that our working memory holds seven plus or minus two items (Miller, 1956). Importantly, our working memory, which allows us to hold multiple items while they are being manipulated, appears to have a limit of around five for monosyllabic English words. These limitations on working memory have serious implications when we consider the task of reading sentences, which are themselves situated in larger texts. To comprehend a sentence, readers need to hold the individual words in working memory and integrate the meaning of the sentence with their developing understanding of the text. If the median sentence length in a text is more than 15 words, for instance (cf. Kornai, 2008 discussing journalistic prose), readers will very quickly exceed the capacity of their working memory.

In contrast to working memory, long-term memory is a relatively abundant resource. Thus, to compensate for a lack of working memory, the brain may make use of long-term memory by storing frequently used strings of words. To illustrate how this might work, we can look at the example of *fish and chips*. Americans visiting the UK for the first time will know the words *fish* and *and*, as well as *chips* (although they will need to learn that the referent of *chips* is different in American and British English). In the UK they are likely to hear sentences like *We're going down to the chippie. Do you want some fish and chips?*, both of which have more than five monosyllabic words, thereby taxing working

memory capacity. Crucially, the sequence of words *fish and chips* will occur frequently in the language that speakers encounter in the UK. (There are probably few high streets in the UK without a signboard saying *fish and chips*). Once the sequence of words has been encountered a sufficient number of times, the words may become associated in long-term memory. Thus, *fish and chips* could be retrieved as a chunk, instead of as individual words, making the sentence *Do you want some fish and chips?* fall within the bounds of working memory capacity.

We want to argue that the use of common phrases as well as the creation of patterns that are specific to a literary text play an important role for the reading experience and the kind of literary meanings created in the reader's mind. There seem to be several possible explanations for the relevance of repeated phrases in the presentation of character information. One of the issues in the interpretation of textual cues is whether the information presented may add to a permanent feature of a character or relates to situational features that are of lesser importance to characterisation. Patterns may describe habitual behaviour because they occur repeatedly in a specific text to describe one and the same character, but there are also literary patterns of character presentation that refer to more general features of fictional people. They are 'repeated' in the sense that they occur across a number of literary texts and thus are part of the reader's literary competence. Such patterns can also add to creating vivid fictional characters with recognisable habits. In corpus linguistic literature, the point has been made that language users are insufficiently aware of quantitative properties of the language (e.g. Biber et al., 1998: 3). Thus, there is a question about whether readers are consciously, or even unconsciously, aware of repeated patterns. More specifically, there are questions about the extent to which repeated patterns lead to readers' impressions of a character, and how these patterns work together with readers' knowledge of real people and fictional characters more generally.

Psycholinguists are also interested in repeated patterns, and there is an increasing body of research dealing with the question of whether frequently occurring sequences of words are represented in long-term memory. For example, Siyanova-Chanturia et al. (2011) used an eye-tracking apparatus to monitor readers' eye movements as they read sentences containing sequences like *bride and groom* or *groom and bride*. Such examples illustrate that even for phrases that refer to concrete entities, the co-occurrence of words is not merely due to repeated experience of the real world, but also is part of the patterning of the language. In the Siyanova-Chanturia et al. (2011) reading study, the sequences were identical in terms of the words used, the syntax and meaning, and only differed in frequency (e.g. *bride and groom* is significantly more frequent than *groom and bride*). If frequently occurring word sequences are stored in long-term memory, *bride and groom* should be retrieved from memory, and thus read more quickly than the otherwise identical but less frequent *groom and bride.* Indeed Siyanova-Chanturia et al. (2011) found that the frequently occurring word sequences were read more quickly than their less frequent counterparts. The Siyanova-Chanturia et al. (2011) study highlights two important things. First, it appears that frequently occurring sequences of words are stored in long-term memory. Secondly, psycholinguists use reading speed to determine whether a sequence is stored in long-term memory. The underlying assumption is that retrieving and manipulating three items (*fish, and, chips*) from long-term memory takes longer than a single chunk (*fish and chips*). Thus, reading times that are statistically

faster than expected are used as an indication that a sequence of words is represented in long-term memory.

In an eye-tracking study, participants merely read passages on a computer monitor at their own pace and the apparatus determines which words are being looked at, when, and for how long. So a big advantage of eye-tracking is that it allows comprehenders to read whole sentences or texts 'naturally', without the need of a secondary task such a pressing a button when they notice a specific phrase. Importantly, while the foregoing discussion indicates that readers may store frequently occurring word sequences in long-term memory, the research to date has been done on sentences or texts that have been carefully designed to investigate the processing of very particular and well-controlled sequences of words. There is no research, that we are aware of, that investigates the processing of such frequently occurring sequences in literary texts.

4 Method

The foregoing discussion highlights the fact that in both corpus stylistics and psycholinguistics, researchers are interested in frequently occurring linguistic patterns. However, corpus stylisticians have been concerned with readers' awareness of these patterns in the sense of the effects that such patterns can potentially create and how this relates to their understanding of characters, while psycholinguists are interested in whether these patterns are entrenched in long-term memory and how this relates to language processing. The goal of the current research is to bring these two domains together. Thus, in the current study:

(a) speeded reading times will be taken as an indication that patterns are entrenched in memory,
(b) results from (a) will be related to readers' awareness of these patterns.

To answer (a) we measure reading times in an eye-tracking experiment, and to answer (b) we elicit information from the participants about what they have read. In the present article, we work with textual examples that are relatively short and only contain one occurrence of a pattern.

4.1 The eye-tracking study

For the eye-tracking experiment we selected examples from Dickens's novels on the basis of the findings in Mahlberg (2013). Eight participants were given 16 text extracts to read where each extract included at least one example of the types of patterns identified in Mahlberg (2013). Not all of the extracts included examples of body language presentation – which are the focus of this article. The present study is part of a larger study for which the participants also read extracts with patterns other than those discussed here. For the purpose of the present study these other passages served as distractors to avoid that participants guess the aims of the experiment. In the following we concentrate on explaining the methodological considerations for the study of body language patterns.

4.1.1 Materials. Of the 16 extracts, seven are covered by the present study. Each of these seven extracts contains an example of a pattern that relates to one of the body language clusters identified in Mahlberg (2013). The starting points for Mahlberg (2013) were five-word clusters. Such five-word clusters can also be part of longer clusters, so that different five-word clusters are related. In example (3), repeated here for ease of reference, *his hand to his forehead* is a five-word cluster, and so is *with his hand to his.*

> (3) Thus it had come about, that Mr Twemlow had said to himself in his lodgings, *with his hand to his forehead*: 'I must not think of this. This is enough to soften any man's brain,' – and yet was always thinking of it, and could never form a conclusion.

(Charles Dickens, *Our Mutual Friend*, Book 1, Chapter 2)

Table 3 includes the nine five-word clusters that relate to the seven extracts of the present study. The table provides the frequency of a cluster in the DNC and in 19C (cf. also Section 2). Table 3 also indicates in how many different texts in each corpus a cluster is found. Mahlberg (2013) interprets these figures with regard to the contextualising and highlighting cline outlined in Section 2. All clusters occur repeatedly and are not limited to one text, and thus can potentially fulfil contextualising functions, in other words, they are not only highlighted features of one particular character. The clusters are not restricted to Dickens only, as shown by the figures for 19C, and are therefore more likely part of the literary competence of readers.

Based on the clusters listed in Table 3, Table 4 shows the regions of interest (ROIs) used for the reading times results reported in Section 5. We use the term 'regions of interest' and not 'clusters' for the linguistic units that we study in our experiments, because ROIs are not automatically the same as five-word clusters. ROIs can be longer to reflect, for instance, the relationship between the two clusters *his hand to his forehead* and *with his hand to his*. For *his eyes fixed on the* we included the *with* that introduced the phrase in the extract. However, some ROIs are the same as five-word clusters, as in *with his hands in his* where we did not add *pockets* to create a complete phrase. We opted for this

Table 3. Body part clusters with their frequencies in DNC and 19C and the number of texts in which they occur.

Body part clusters	DNC	DNC texts	19C	19C texts
his eyes fixed on the	17	9	14	6
with his eyes on the	16	7	4	3
his hand to his forehead	16	5	2	2
with his hand to his	30	9	2	2
his hands in his pockets	77	14	13	8
with his hands in his	51	14	12	7
laying his hand upon his	20	6	1	1
his hand upon his shoulder	10	4	2	1
with his back to the	41	12	22	11

Table 4. ROIs used in the reading times study.

with his eyes fixed on the
with his eyes on the
laying his hand upon his shoulder
with his hands in his
with his back to the
with his hand to his forehead
pressing his head between his hands

variation to see whether we can gain any insights that will help with refining the design for future studies. We also included an example that is similar in its textual function, but is not a cluster: *pressing his head between his hands* only appears once in Dickens (and not at all in 19C), that is to say it is not a cluster in the corpus under investigation.

As pointed out in Section 2, clusters can be incomplete phrases. As a cluster such as *his eyes fixed on the* is difficult or impossible to elicit, because the noun phrase is incomplete, we will redefine the ROIs for the follow-up questions (see section 4.2), so that for example (2) from Section 2, *table* will be part of the ROI: *his eyes fixed on the table*. However, any speed up in processing for the cluster due to frequency and entrenchment in memory should be due to the frequently occurring portion of the cluster, *his eyes fixed on the*.

4.1.2 Participants. Eight native English speakers from the University of Nottingham were paid £6 for their participation. All participants had normal or corrected-to-normal vision. Eight participants might seem like a small number. However, in traditional psycholinguistic studies, linguistic manipulations are presented across different experimental lists. Thus, a participant assigned to one list would see *bride and groom*, while a participant on another list would see the alternative *groom and bride*. Such a study would typically have 14–25 participants, meaning that only 7–12 participants saw any one linguistic form. Thus, 8 participants seems to be a reasonable number for a study where there is no manipulation of forms across versions of an experiment.

4.1.3. Eye-tracking apparatus and procedure. Each participant read the texts on a computer screen. Texts were presented double-spaced, in size 16 Courier New font. The order of presentation of the extracts was randomised across participants, that is, the order was different for each participant to avoid effect due to the order of the presentation. Eye movements were monitored using an EyeLink I eye-tracker (SR research, 250Hz). Before starting, participants were given a set of instructions and an explanation of the eye-tracking procedure. Participants were told that they would answer questions about what they had read. A nine-point calibration procedure was then done to ensure that the eye-tracker was accurately capturing the eye's position. Before presentation of a new passage, a calibration check was performed to ensure that the eye-tracker was still capturing the eye's exact position, and when necessary, recalibration was done.

4.2 Follow-up questions

Directly after reading an extract, participants were asked questions about what they had read. They answered these questions from memory, as the extract was not displayed on the screen any more. For example, participants were asked the following after reading the extract from *Dombey and Son* containing example (2) illustrated in Section 2, the full extract is provided in the Appendix as extract (1).

Q1: Describe the father in terms of adjectives and short phrases.
Q2: What time of day is it?
Q3: What is the father doing?
Q4: What is the father looking at?
Q5: How many fictional characters are mentioned by name in this passage?
Q6: Have you ever read *Dombey and Son?*
Q7: Or seen the film?

Questions after each extract were designed to be similar. For the first question, we adopted an approach similar to Culpeper (2001), who conducted an experiment to elicit character information. Question 1 asks for adjectives and short phrases, and aims to elicit the most condensed character information, which is potentially far removed from the actual words in the text. Questions 3 and 4 are increasingly biased prompts that are directed towards the ROI. These questions will help to ascertain if participants had any awareness of the clusters. Questions 2 and 5 were included merely as distractors to prevent participants from accurately guessing the intent of the study. Questions 6 and 7 ask for background information that is potentially relevant to recognising highlighting patterns. In other words, anyone who has familiarity with the text should have had increased exposure to the ROI, which could further speed reading times. If participants have read the book/seen the film they will not be excluded from the study.

Before starting the study itself, participants were given an example of a question like Question 1 to ensure they understood what was being asked for:

Example: Describe Harry Potter in terms of adjectives and short phrases.

Answer: Wizard, wears glasses, has a scar on his forehead, speaks Parseltongue, fights against Voldemort.

The analysis of the responses to these questions requires a more qualitative approach because we have to assess the varied information that the participants provided. For this qualitative analysis we focus on three extracts and redefine our regions of interests so that they include more phraseologically complete units. This approach will be explained in more detail in Section 5.2 with reference to the actual examples.

5 Results

We begin by reporting reading times for the seven ROIs (Section 5.1) and then focus on three clusters in particular to discuss results of the follow-up questions (Section 5.2).

5.1 Reading times

Prior to analysis, the eye-movement data from each participant were subject to the following procedures. Fixations less than 80 ms (milliseconds) and within 0.5° visual angle of another fixation and fixations less than 40 ms and within 1.25° visual angle of another fixation were merged with those fixations. Any remaining fixations less than 80 ms or greater than 1200 ms were then deleted, because short fixations reflect oculomotor programming, and long fixations are due to momentary track loss or blinks (Morrison, 1984). This procedure resulted in the loss of 6% of the data.

To provide a metric of overall comprehension ease/difficulty, analyses were done on 'total reading time' and 'total fixation count'. The former measure sums the duration of all fixations, while the later sums the number of fixations.[6] In each passage, ROIs (cf. Tables 4 and 5) were identified. Total reading times and numbers of fixations were collected for these regions as well as for the entire passage. Importantly, the ROIs and the extracts differed in length, making them difficult to compare. To correct for length differences, a total reading time/character calculation was done (see Sanford and Filik, 2007 who use this calculation to account for length differences in their reading study). For example, the ROI *with his hands in his* is 22 characters long (including spaces) and it is situated in a passage that is 376 characters long (including spaces). We can calculate in milliseconds how long a participant spent reading the passage as a whole and how much time was spent reading the ROI. If we look at a single participant reading this passage and the embedded ROI, we see that they spent 23,256 ms on the 376-character long passage, giving a reading speed of 61.85 ms/character. In the ROI, they spent 290 ms reading the 22 characters, giving them a reading speed of 13.18 ms/character. This calculation was done for each participant for every passage and the embedded ROI. This single example shows that, for this particular participant, reading was considerably faster for the ROI than for the rest of the passage.

Table 5 summarises the average number of fixations (FC) and the average reading times (RT) for each ROI and the passages in which they occur. In order to determine whether the differences between the ROIs and the whole extracts were significantly reliable, statistical analyses were done. Overall, participants read passages at the rate of 38.18 ms/character (SD = 5.31) and the ROIs at a rate of 26.46 ms/character (SD = 5.94). Paired sample t-tests were conducted by participants (t_1) and by items (t_2) to determine whether differences were significant, and thus generalisable to a similar set of participants and items. Analyses revealed that average reading speed was significantly faster for the ROIs than for the text as a whole ($t_1(7) = -5.02$, p < .01; $t_2(6) = -3.36$, p <0.02). The effect size for this analysis ($d = 2.08$) exceeds Cohen's (1988) convention for a large effect size ($d = .80$). Analyses of fixation count showed that there were on average fewer fixations per character in each word for ROIs ($M = .14$, $SD = .03$) than in the text as a whole ($M = .19$, $SD = .03$) and that this difference was significant ($t_1(7) = -3.62$, p < .01; $t_2(6) = -2.80$, p <0.03). Again the effect size ($d = 1.6$) exceeds what is classified as a large effect. As Table 5 shows, the ROI that is not a cluster (*pressing his head between his hands*) has the smallest difference between the ROI and the whole passage – which seems to provide even further support for the assumption that the reading time is affected by the repeated occurrence of a sequence. However, Table 5 also indicates that for the

Table 5. Regions of Interest (ROI), the average number of fixations (FC) and the average reading time (RT) for them, as well as for the passages in which they occur. Differences in the FC and RT for the ROI and passage, with positive values indicating fewer fixations or less reading time per character/letter in the ROI.

ROI	ROI average FC	Passage average FC	Difference in passage and ROI FC	ROI average RT	Passage average RT	Difference in passage and ROI RT
with his eyes fixed on the	0.15	0.21	0.05	27.81	41.09	13.28
with his eyes on the	0.12	0.24	0.12	20.43	48.87	28.45
laying his hand upon his shoulder	0.16	0.18	0.03	27.41	36.41	9.00
with his hands in his	0.13	0.19	0.07	23.91	37.39	13.48
with his back to the	0.19	0.17	−0.02	33.48	35.10	1.62
with his hand to his forehead	0.09	0.17	0.08	18.45	33.31	14.86
pressing his head between his hands	0.16	0.18	0.02	33.73	35.12	1.39
Mean =	0.14	0.19	0.05	26.46	38.18	11.72

cluster *with his back to the*, the difference is relatively small. A reason for this might be that body language that refers to the eyes and hands is different from body language that describes posture and positions. This is an observation that will need further investigation, but it might point to a further subclassification of body language clusters, revealed with the help of reading-time information.

5.2 Responses to open questions

We report qualitative and quantitative results for three of the 16 extracts. The three extracts include examples (2) and (3) given earlier, as well as the extract that contains the following example (4). For all three examples the full extracts and the follow-up questions used in the study are included in the Appendix.

> (4) 'Hush!' said Nicholas, *laying his hand upon his shoulder*. 'Be a man; you are nearly one by years, God help you.'
>
> (Charles Dickens, *Nicholas Nickleby*, Chapter 8)

In the analysis of the answers to our follow-up questions we were interested in the extent to which the answers suggested that the readers had consciously noticed the ROIs

used in the reading time experiment. However, as outlined in Section 4, for this part of the study we redefined some ROIs to account for the fact that participants are unlikely to refer to incomplete phrases. The following are the ROIs for the analysis of the answers to the open questions. For extracts 2 and 3 the ROIs are in fact six-word clusters resulting from the two respective clusters in Table 3. For extract 1, *with* and *table* are added to the cluster because they complete the phrase in the example.

> Extract 1: *with his eyes fixed on the table*
> Extract 2: *with his hand to his forehead*
> Extract 3: *laying his hand upon his shoulder*

In the classification of the answers given by the participants a distinction was made between 'unbiased' and 'biased' prompts. For extract 1, *Q1: Describe the father in terms of adjectives and short phrases* is an unbiased prompt, whereas *Q3: What is the father doing?* and *Q4: What is the father looking at?* are biased because these questions focus on the characters' actions described by the ROI. The answers were assessed according to the following binary criteria:
Unbiased prompt:

(1) Is part of the ROI elicited? Y/N
(2) Is part of the context of the ROI elicited? Y/N
(3) Is 'wider' interpretation provided/has the participant noticed character information? Y/N

Biased prompt:

(1) Is part of the ROI elicited? Y/N
(2) Is part of the body language from the ROI elicited? Y/N
(3) Is part of the context of the ROI elicited?

For these categories, the 'context' of the ROI is defined as the whole sentence around the ROI. The results are summarised in Table 6. Because three examples are included in the analysis and eight participants answered the questions, the total number of responses for each category is 24.

Our hypothesis for the analysis is that reading times of the clusters are shorter, because readers do not pay as much attention to them as they do to other parts of the text. This is particularly the case for clusters with contextualising functions. When participants only see the single text extract for each of the clusters, all are potentially contextualising. As Mahlberg (2013) argues, what adds to the highlighting function of *(with) his hand to his forehead* is that the cluster appears repeatedly throughout the novel always referring to the same character. In the experiment, however, readers do not see this context. The offline responses about the text provide another indication of what readers attended to while reading the text. Table 6 shows that when a question simply asks participants about the passage in general, information from the ROI is rarely elicited, nor is part of its context. Importantly, participants do give wider information about the character. It seems

Table 6. Results of the analysis of the open questions using a Binomial test (two-tailed significance), which is used for dichotomous (e.g. yes/no) data.

Prompt type and information elicited	Response type	Significance
Unbiased prompt		
(1) ROI	Yes = 2 No = 22	0.001
(2) context	Yes = 3 No = 21	0.001
(3) wider character information	Yes = 19 No = 5	0.01
Biased prompt		
(1) ROI	Yes = 8 No = 16	n.s.
(2) body language	Yes = 5 No = 19	0.01
(3) context	Yes = 1 No = 23	0.001

that comprehenders remember important character information, but not necessarily the linguistic form in which this information is presented. For example, Mr Dombey ('the father' in example 2) is described as 'distracted', 'concerned' or 'disconnected'. This finding is in line with more general psycholinguistic findings that comprehenders remember the gist of what they read, but not the specific form of the text (Bobrow, 1970).

The answers to the biased questions show that more specific forms/parts of the ROIs are elicited when prompted. The number of participants who reproduce (part of) the ROIs is not significantly different from those who do not. This shows that readers actually do pay attention to the clusters but do not seem to be fully aware of them unless their attention is drawn to the ROI.

6 Conclusions

Our findings indicate that there are body language clusters that are read faster than other parts of the text, which suggests that these clusters are (part of) sequences that readers store as units in long-term memory. Against the background of research in psycholinguistics and corpus linguistics, this suggests that such body language clusters are part of readers' literary competence that is built up through exposure to a number of literary texts. The results of the follow-up questions that used different degrees of biased prompts further suggest that the clusters receive less attention when reading because the body language is not perceived as the most important information in the extract. Even if readers are able to remember body language when prompted, they rarely mention it in the description of characters, which is in line with the contextualising functions of the patterns. Part of the contextualising functions of the patterns is also that the meaning of the body language is mutually dependent on the context in which it occurs.

Together these findings make an important contribution to the discussion of the concept of literariness. Although narrative fiction creates fictional worlds that are different from the real world, there is a set of patterns that function as building blocks of fictional texts and that create a background against which the more specific characteristics of a fictional world are set. Although these characteristic features tend to receive the main conscious attention of readers, the less noticeable patterns can also be identified. So far, these patterns have mainly been discussed on the basis of corpus findings. By linking

corpus findings with psycholinguistic data, there is further support for the patterns as relevant textual units.

The combination of corpus and psycholinguistic data also raises questions about further subcategories of clusters. On the basis of the corpus data alone, the cluster *with his back to the* could be regarded as a cluster whose main function is contextualising. This cluster occurs across both Dickens and other 19th-century writers. However, in terms of reading time, it stands out from the other body language clusters. Hence a question for future studies is whether readers perceive clusters that refer to different parts of the body in different ways.

In addition to findings about the selected set of patterns, this study also makes a contribution to complementing methods in corpus linguistics with psycholinguistic approaches. Questions that we have dealt with in this study relate to the way in which we can design experiments in such a way that they can help investigate issues raised by corpus findings. Our considerations about the ROIs show that units that are revealed by corpus findings are not necessarily the same as the units that can be investigated through psycholinguistic experiments. Especially the categories we used for the biased and unbiased prompts can serve as a starting-point for future work that aims to specify the relationship between clusters and the contexts in which they are found.

In the present article the focus is on patterns that can function as contextualising. To investigate the relationship between contextualising and highlighting, future work needs to look at the use of patterns across the same text. If a particular sequence of words occurs repeatedly in a text, these words should become associated in the reader's long-term memory. Thus, readers could represent very specific sequences of words, which occur several times in a single text, in long-term memory. We would expect to see evidence of this in reading times, such that over the course of reading a text, comprehenders should become faster at reading the frequent sequences. Another point for future work is to investigate the effect of previous knowledge of the book or film. In the present study there were only a few participants who actually had read the book/seen the film. When we analysed their results separately, they did not pattern significantly, so we need to collect more data to be able to make any claims in this regard.

Funding

Notes

1. In Mahlberg (2013) the threshold for clusters requires that there is at least one text in which the cluster occurs five times. See Chapter 3 in Mahlberg (2013) for a discussion of the definition of 'clusters' vs 'lexical bundles'.
2. Page numbers are not provided, as the examples are taken from e-texts from Project Gutenberg, http://www.gutenberg.org (e-texts downloaded 2006).
3. DNC contains about 3.9 million words, 19C contains about 4.5 million words.
4. The clusters are based on Mahlberg (2013), but note that the frequencies in the present article (cf. also Table 3) only refer to Dickens's novels, whereas the respective table in Mahlberg (2013: 106f.) refers to DCorp, which includes the novels and other texts written by Dickens.

5.   The reader's experience of encountering certain patterns also plays a role in Hoey's (2005) more general theory of 'lexical priming'.
6.   In eye-tracking, a number of measures can be analysed. 'First fixation duration' (duration of the first fixation on a word), 'total reading time' (sum of all fixation durations), 'total fixation count' (sum the number of fixations) are often reported. However, early measures, which count initial fixations and durations on an individual word, have not been found to be of interest when looking at longer stretches of text beyond one or two words (e.g. Siyanova-Chanturia et al., 2011).

References

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

Bannard C and Matthews D (2008) Stored word sequences in language learning: The effect of familiarity on children's repetition of four-word combinations. *Psychological Science* 19(3): 241–248.

Biber D, Conrad S and Reppen R (1998) *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Bobrow S (1970) Memory for words in sentences. *Journal of Verbal Learning and Verbal Behavior* 9(4): 363–372.

Brook GL (1970) *The Language of Dickens*. London: Andre Deutsch.

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

Cohen J (1988) *Statistical Power Analysis for the Behavioral Sciences* (2nd edn). Hillsdale, NJ: Lawrence Earlbaum Associates.

Culpeper J (2001) *Language and Characterisation. People in Plays and Other Texts*. Harlow: Pearson Education.

Hoey M (2005) *Lexical Priming: A New Theory of Words and Language*. London: Routledge.

Kornai A (2008) *Mathematical Linguistics*. New York: Springer Verlag.

Leech G and Short M (2007) *Style in Fiction. A Linguistic Introduction to English Fictional Prose* (2nd edn). Harlow: Pearson Education.

Mahlberg M (2013) *Corpus Stylistics and Dickens's Fiction*. New York and London: Routledge.

Miller GA (1956) The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63(2): 81–97.

Morrison R (1984) Manipulation of stimulus onset delay in reading: Evidence for parallel programming of saccades. *Journal of Experimental Psychology: Human Perception and Performance* 10(5): 667–682.

Paroissien D (2000) Characterization. In: Schlicke P (ed.) *Oxford Reader's Companion to Dickens*. Oxford: Oxford University Press, 76–82.

Sanford AJ and Filik R (2007) 'They' as a gender-unspecified singular pronoun: Eye-tracking reveals a processing cost. *The Quarterly Journal of Experimental Psychology* 60(2): 171–178.

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

Siyanova-Chanturia A, Conklin K and van Heuven WJB (2011) Seeing a phrase 'time and again' matters: The role of phrasal frequency in the processing of multiword sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 37(3): 776–784.

Stockwell P (2009) *Texture. A Cognitive Aesthetics of Reading*. Edinburgh: Edinburgh University Press.

Wray A (2008) *Formulaic Language: Pushing the Boundaries*. Oxford: Oxford University Press.

Author biographies

Michaela Mahlberg is Professor of English Language and Linguistics at the University of Nottingham, UK. She is the editor of the *International Journal of Corpus Linguistics* (John Benjamins) and co-editor of the book series *Corpus & Discourse* (Bloomsbury). She has published on the relationship between lexis and grammar, the way in which phraseology contributes to cohesion in texts and the use of corpus methods in the analysis of discourse. Her recent publications include the book *Corpus Stylistics and Dickens's Fiction* (Routledge, 2013).

Kathy Conklin is an Associate Professor in psycholinguistics at the University of Nottingham, UK. She is interested in multi-word units, automatic word activation in an L1 vs. L2, bilingual/monolingual cognitive control, and implicit L2 learning. To investigate L1/L2 language representation and processing, she makes use of behavioral measures, eye-tracking, and EEG. She has published articles in journals such as *Bilingualism: Language and Cognition; Applied Psycholinguistics* and *Brain Research*.

Marie-Josée Bisson is a Postdoctoral Research Associate at Loughborough University, UK. She completed a PhD in Psychology at the University of Nottingham in 2013 and has published articles in *Applied Psycholinguistics, PLoS One* and *Language Learning.* Her research interests include incidental vocabulary learning, multi-modalities, eye-tracking, and bilingualism.

Appendix

Extract 1 (Charles Dickens, Dombey and Son, Chapter 18)

In its standing open, even by so much as that chink, there seemed to be hope. There was encouragement in seeing a ray of light from within, stealing through the dark stern doorway, and falling in a thread upon the marble floor. She turned back, hardly knowing what she did, but urged on by the love within her, and the trial they had undergone together, but not shared: and with her hands a little raised and trembling, glided in.

Her father sat at his old table in the middle room. He had been arranging some papers, and destroying others, and the latter lay in fragile ruins before him. The rain dripped heavily upon the glass panes in the outer room, where he had so often watched poor Paul, a baby; and the low complainings of the wind were heard without.

But not by him. He sat *with his eyes fixed on the table*, so immersed in thought, that a far heavier tread than the light foot of his child could make, might have failed to rouse him. His face was turned towards her. By the waning lamp, and at that haggard hour, it looked worn and dejected; and in the utter loneliness surrounding him, there was an appeal to Florence that struck home.

'Papa! Papa! speak to me, dear Papa!'

He started at her voice, and leaped up from his seat. She was close before him, with extended arms, but he fell back.

Q1: Describe the father in terms of adjectives and short phrases.
Q2: What time of day is it?
Q3: What is the father doing?

Q4: What is the father looking at?
Q5: How many fictional characters are mentioned by name in this passage?
Q6: Have you ever read *Dombey and Son?*
Q7: Or seen the film?

Extract 2 (Charles Dickens, Our Mutual Friend, Book 1, Chapter 2)

But, it was not this which steeped the feeble soul of Twemlow in confusion. This he was used to, and could take soundings of. The abyss to which he could find no bottom, and from which started forth the engrossing and ever-swelling difficulty of his life, was the insoluble question whether he was Veneering's oldest friend, or newest friend. To the excogitation of this problem, the harmless gentleman had devoted many anxious hours, both in his lodgings over the livery stable-yard, and in the cold gloom, favourable to meditation, of Saint James's Square. Thus. Twemlow had first known Veneering at his club, where Veneering then knew nobody but the man who made them known to one another, who seemed to be the most intimate friend he had in the world, and whom he had known two days – the bond of union between their souls, the nefarious conduct of the committee respecting the cookery of a fillet of veal, having been accidentally cemented at that date. Immediately upon this, Twemlow received an invitation to dine with Veneering, and dined: the man being of the party. Immediately upon that, Twemlow received an invitation to dine with the man, and dined: Veneering being of the party. At the man's were a Member, an Engineer, a Payer-off of the National Debt, a Poem on Shakespeare, a Grievance, and a Public Office, who all seem to be utter strangers to Veneering. And yet immediately after that, Twemlow received an invitation to dine at Veneerings, expressly to meet the Member, the Engineer, the Payer-off of the National Debt, the Poem on Shakespeare, the Grievance, and the Public Office, and, dining, discovered that all of them were the most intimate friends Veneering had in the world, and that the wives of all of them (who were all there) were the objects of Mrs Veneering's most devoted affection and tender confidence.

Thus it had come about, that Mr Twemlow had said to himself in his lodgings, *with his hand to his forehead*: 'I must not think of this. This is enough to soften any man's brain,' – and yet was always thinking of it, and could never form a conclusion.

This evening the Veneerings give a banquet. Eleven leaves in the Twemlow; fourteen in company all told. Four pigeon-breasted retainers in plain clothes stand in line in the hall. A fifth retainer, proceeding up the staircase with a mournful air – as who should say, 'Here is another wretched creature come to dinner; such is life!' – announces, 'Mis-ter Twemlow!'

Q1: Describe Twemlow with adjectives and short phrases.
Q2: Can you remember any of his gestures?
Q3: What does he do with his hand?
Q4: How many fictional characters are mentioned by name in this passage?
Q5: Have you ever read *Our Mutual Friend?*
Q6: Or seen the film?

Extract 3 (Charles Dickens, Nicholas Nickleby, Chapter 8)

There was such an obvious fear of giving offence in his manner, and he was such a timid, broken-spirited creature, that Nicholas could not help exclaiming, 'Poor fellow!'

If he had struck the drudge, he would have slunk away without a word. But, now, he burst into tears.

'Oh dear, oh dear!' he cried, covering his face with his cracked and horny hands. 'My heart will break. It will, it will.'

'Hush!' said Nicholas, *laying his hand upon his shoulder*. 'Be a man; you are nearly one by years, God help you.'

'By years!' cried Smike. 'Oh dear, dear, how many of them! How many of them since I was a little child, younger than any that are here now! Where are they all!'

'Whom do you speak of?' inquired Nicholas, wishing to rouse the poor half-witted creature to reason. 'Tell me.'

'My friends,' he replied, 'myself – my – oh! what sufferings mine have been!'

'There is always hope,' said Nicholas; he knew not what to say.

'No,' rejoined the other, 'no; none for me. Do you remember the boy that died here?'

'I was not here, you know,' said Nicholas gently; 'but what of him?'

'Why,' replied the youth, drawing closer to his questioner's side, 'I was with him at night, and when it was all silent he cried no more for friends he wished to come and sit with him, but began to see faces round his bed that came from home; he said they smiled, and talked to him; and he died at last lifting his head to kiss them. Do you hear?'

Q1: Describe Nicholas in terms of adjectives and short phrases.
Q2: Describe Smike in terms of adjectives and short phrases.
Q3: What does Nicholas do to encourage Smike?
Q4: What does Nicolas do with his hands?
Q5: How many fictional characters are mentioned by name in this passage?
Q6: Have you ever read *Nicholas Nickleby?*
Q7: Or seen the film?