CLiC Dickens

Mahlberg, Michaela; Stockwell, Peter; De Joode, Johan; Smith, Catherine; O'Donnell, Matthew Brook

DOI:
10.3366/cor.2016.0102

CLiC Dickens: novel uses of concordances for the integration of corpus stylistics and cognitive poetics*

Michaela Mahlberg,[1] Peter Stockwell,[2] Johan de Joode,[1]
Catherine Smith[3] and Matthew Brook O'Donnell[4]

Abstract

This paper introduces the web application CLiC, which we developed as part of a research project bringing together insights from both cognitive poetics and corpus stylistics, with Dickens's novels as a case study. CLiC supports the analysis of discourse in narrative fiction with search options that make it possible to focus on stretches of text within and outside quotation marks. We argue that such search options open up novel ways of using concordances to link lexico-grammatical and textual patterns. We focus specifically on patterns for the creation of fictional characters. From a technical point of view, we explain the XML annotation that CLiC works with. Our discussion of textual examples focusses on phrases in fictional speech that illustrate significant differences between text within and outside quotation marks. In terms of theory, we argue that CLiC supports the identification of textual patterns that can provide insights into fictional minds and contribute to the exploration of readerly effects within the wider framework of mind-modelling.

Keywords: Dickens, fictional speech, suspensions, characterisation, mind-modelling.

* This work was supported by the Arts and Humanities Research Council grant reference AH/K005146/1.
[1] Department of English Language and Applied Linguistics, University of Birmingham, Edgbaston, B15 2TT, United Kingdom.
[2] School of English, University of Nottingham, University Park, Nottingham, NG7 2RD, United Kingdom.
[3] ITSEE, Department of Theology and Religion, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom.
[4] Annenberg School for Communication, University of Pennsylvania, 3620 Walnut Street, Philadelphia, PA 19104, USA.

Corpora 2016 Vol. 11 (3): 433–463
DOI: 10.3366/cor.2016.0102

1. Introduction: the use of the concordance

'The language looks rather different when you look at a lot of it at once', declared John Sinclair (1991: 100) famously, summarising a fundamental tenet of corpus linguistics. The central corpus linguistic tool to enable researchers to look at a lot of the language at once is the concordance. This is a display format, showing the search word in the centre with a specified amount of context on the left and on the right. Since the beginning of modern corpus linguistics, concordances have played a significant role in the field, and the revolutionary effect they have had on the study of language is probably best exemplified in the Cobuild project (Sinclair, 1987). In the pre-computer era, concordances were compiled manually and so were reserved for high status texts such as the Koran, the Bible and Shakespeare's works, for example. Computational power meant that any text could be displayed in a concordance format and, as computers became more commonly used, the ability to create and compile concordances came within the reach of more researchers of language. With corpus applications spreading across into other disciplines, an increasingly wide range of researchers are now drawing on the support of this display format to investigate the meanings of words in co-text and context, such as in medical sciences (Skelton *et al*., 2002), language teaching (Johns, 1990) and continuing in religious studies (Altmeyer *et al*., 2015). Furthermore, the usefulness of concordances in school education has been recognised (Giovanelli *et al*., 2015).

In current corpus software, the concordance display is a standard function, as illustrated by the leading general software packages WordSmith Tools (Scott, 2016) and AntConc (Anthony, 2016), but also by the more specific web applications, such as CQPweb (Hardie, 2012), WebCorp (Renouf *et al*., 2005), or the BYU interface (Davies, 2010). Given the centrality of the concordance, it is surprising that there have been few advances in developing this useful tool further. Indeed, the standard appearance of the computer-generated concordance is strikingly similar to medieval biblical concordances, the first of which, in the thirteenth century, was compiled by 500 monks under the direction of Friar Hugh of Saint-Cher (see Rouse and Rouse, 1991).

Though the concordance display itself seems historically stable, a fresh look at the tool reveals its potential for interdisciplinary work and highlights the need for different disciplines to try to tackle digital challenges through collaborative research. In this paper, we present the web application CLiC (Corpus Linguistics in Context).[5] CLiC was developed to study Charles Dickens's fiction; however, it also shows the wider potential of the concordance display if it is combined with input data that is created to investigate specific research questions. While CLiC uses standard concordance functionalities, the corpus it accesses is marked-up in such a

[5] CLiC was originally 'Corpus Linguistics in Cheshire', but we renamed the web app in the course of the project to make it less dependent on this particular database.

way that different parts of the text can be searched as standard options. The technical development in itself is not the feature that makes CLiC stand out from other corpus tools, but the power of CLiC becomes apparent through the novel way in which it enables a search of discourse in narrative fiction.

The CLiC Dickens project, within which the web application was developed, set out to be a collaboration that drew together insights from both cognitive poetics and corpus stylistics – the two fields that have been most productive in recent literary stylistic research. As 'corpus stylistics', corpus methods are increasingly used to study literary texts and readings. For the study of literature, a simple use of the concordance display is to support close reading. More challenging, however, is to combine corpus methods and approaches in literary linguistics in a truly integrated fashion. We developed CLiC as part of a project with the overall research question: how can corpus methods be combined with literary linguistic approaches to produce new insights into the creation of meanings in literary texts? The specific area of focus for our work was the investigation of textual patterns in the creation of fictional characters in prose fiction, particularly textual representations in terms of speech and body language, using Dickens's novels as a case study. Our objective here is not to provide an account of the entire CLiC project – this would be beyond the scope of a single paper. Our aims for this paper are as follows:

(*1*) To introduce the web application CLiC;
(*2*) To illustrate how corpus data helps to test and validate theoretical claims in cognitive poetics; and,
(*3*) To argue that the theoretical concerns that have driven the development of CLiC have wider implications for what has come to be known as the Digital Humanities.

Our emphasis is on the relationship between corpus tools, specifically the concordance, and the research questions they can help to answer. Insight is driven not by the tool but by the overall research questions that guide technological development. For this reason, Section 2 contextualises our work in a wider digital-humanities context. Section 3 presents our approach to the search options in CLiC, providing brief technical background. Section 4 illustrates the search options with textual examples focussing on fictional speech. The conclusions in Section 5 include a discussion of further implications and directions for further research.

2. Using CLiC for the study of characterisation

The principal research problem guiding our work is the readerly conceptualisation of characterisation in narrative fiction: a process that combines textuality and mentality (Culpeper, 2001; Stockwell, 2009; and Vermeule, 2010). Recent cognitive poetic approaches in literary linguistics emphasise the

relationship between top-down and bottom-up processes in creating textual meanings and aesthetic effects. A literary linguistic analysis is text-driven in that (bottom-up) patterns in the text function as cues for the (top-down) activation of schematic knowledge. This text-drivenness offers the crucial linking concept to propose a general theoretical integration between corpus linguistics and cognitive poetics, as we will explain in more detail.

Our work on characterisation highlights how, in the field of literary discourse, narrative fiction presents particular key problems of addressivity and layers of perceiving consciousnesses. While embedded narrators and characters can be discerned in narrative forms of poetry such as ballad, epic or dramatic monologue, for the most part the dominant poetic forms such as the lyric, sonnet, confessional, elegy, ode or panegyric seem to present an apparently direct perceiving voice speaking its mind. Similarly, in drama, characters on stage or screen seem to speak directly for themselves, with the playwright rarely appearing on-scene other than in the stage directions in the playscript. For these reasons, narrative discourse presents the reader with a complex set of dialogic relationships (Bakhtin, 1982) and embedded viewpoints, marked out textually by a variety of means of thought and speech presentation.

In traditional methods within literary stylistics, identifying these different layers of consciousness has required manual search and annotation of what are often long extents of text in novels. The speech and thoughts of different characters, narrators, implied authors and the authorial extra-fictional voice are marked out textually in a variety of ways. Corpus linguistic methods are allowing us to take a fresh view on how we identify these different layers of consciousness. In the present paper, we focus on the presentation of speech as a crucial aspect of characterisation. In Section 4, we will develop the argument of discourse presentation further with regard to the concept of mind-modelling. Here in Section 2, we begin by outlining theoretical and practical points about the analysis of discourse presentation that provide links to corpus linguistic concerns.

2.1 Corpus stylistic and related studies of speech

The application of corpus methods to analyse patterns in speech is most straightforward when studying drama, as the text clearly indicates the direct speech of the characters, and there are no reporting clauses surrounding the actor's turn. Stage directions in playscripts are often formatted distinctly and so can be easily differentiated. As illustrated by Culpeper (2009), this text format lends itself readily to the extraction of all the speech of a particular character and comparisons between speakers. In a similar way, Bednarek's (2010) work on dialogue in TV series benefits from the format in which TV scripts are readily available.

By contrast, other narrativised genres and modes of writing are more complex for corpus stylistics, unless a text displays very specific features.

Walker (2010) is able to compare different narrators in Julian Barnes' *Talking it Over*, as the text is particularly suited for the key comparison approach – different chapters of the novel are told from the point of view of different first-person narrators. In their study, Semino and Short (2004) draw on Leech and Short's (1981) distinctions between direct speech, indirect speech, free indirect speech, and so on, to compare the distribution of the discourse presentation categories across sub-corpora of twentieth-century fictional, journalistic and autobiographical/biographical narratives in a corpus amounting to about 250,000 words. One of the outcomes of this study is the revision and extension of the Leech and Short (1981) model, specifically by including a scale for writing presentation in addition to speech and thought presentation. Busse (2010) develops the model further by applying it to a corpus of nineteenth-century fiction, and McIntyre and Walker (2011) focus on Early Modern English prose fiction and news writing. Given the nature of the discourse presentation scales, corpus linguistic studies in this area require manual annotation.

More computational techniques have also been used for the study of speech. For non-literary texts, Krestel *et al*. (2008), for instance, focus on speech in news reports, automatically annotating elements such as the source of the speech or the reporting verb. Studies of speech in news discourse are particularly concerned with the way in which attitudes and opinions are expressed and negotiated (Bergler, 2005; Krestel, 2006; and Balahur *et al*., 2009). Such approaches seem to focus on the attribution of the speech to a speaker and the effects of this for the interpretation of what is said.

Similarly, the computational analysis of direct speech within literary texts seems to have focussed on the identification of speakers. Glass and Bangay (2007), for instance, identify speech-verb sequences and attribute these to particular speakers. Elson and McKeown (2010: 1014) define quoted speech as 'a block of text within a paragraph falling between quotation marks', although they do not explicitly explain how they extract quoted text from the corpus. The features they use for speaker attribution include the proximity of a candidate character to speech or the frequency of occurrence of characters in quotes to identify the most probable character for a given quote. Developing this work futher, Elson *et al*. (2010) use the speaker attribution algorithm to extract dialogue fragments and dialogue partners to look at social networks in nineteenth-century British literature. It seems that work on quote extraction in literary and non-literary texts has largely focussed on who is speaking and not so much on what is being said. While we do not deny the value of speech attribution, in the CLiC project our interest is in the actual represented speech.

2.2 Concordance tools in use

With regard to work in literary criticism, corpus stylistics might be seen as a way to make links to 'close reading' (after Richards, 1929) specifically with the help of concordances. However, more recently within literary studies,

```
138       of the table under the cloth, with both hands, and awaited my fate.
139      lding him prisoner by the coat with both hands. 'When you saw what a
140       which, she tugged at his coat with both hands, and pulled him all
141      p, and seize him by the collar with both hands! 'You know what I
142      ther, taking him by the collar with both hands, 'I'll draw upon you;
143      h, clutching him by the collar with both hands, and shaking him as
144      ; then, clutching the coverlid with both hands, muttered some
145       Spenlow, adjusting his cravat with both hands. 'Take a week, Mr.
146       his brush, and seized the dog with both hands by the collar.
147       his drowned cap over his ears with both hands, and making himself
148      ls, she applied it to her eyes with both hands at once. 'He was
149      t, for she held it to her eyes with both hands and sat so, shedding
150      -handkerchief against his eyes with both hands-- as such men always
151      s eye, and, shielding his face with both hands, protested, while he
152      asped his brawny throat firmly with both hands. His face grew purple;
[…]
171       the Chinaman, and seizing him with both hands by the throat, turns
[…]
188      reast, and clutched its throat with both hands. 'Villain!' cried Mr
[…]
1,095    ve it about me,' said Jonas, putting his hands to his throat, as
```

Figure 1: A sample of the 2,616 concordance lines for *hands* in Dickens's novels retrieved with WordSmith Tools (Scott, 2016).

Moretti (2013: 48) has argued that an approach of 'distant reading' is preferable, where the close texture of literary works is set aside in preference for large-scale trends and cross-novel patterns. Moretti (2013) also uses computational methods to produce visualisations across many texts in order to highlight generic developments and characteristics. Of course, it is not necessary to choose between close and distant forms of reading or analysis, and in fact the basic premise of cognitive poetics is that top-down and bottom-up processes work together. The use of a contextualised concordance can address both aspects of the process.

As Tognini-Bonelli (2001) points out, to find repeated co-occurrences of words, a concordance is read 'vertically', rather than horizontally, as a text is normally read. The patterns that become visible in a concordance provide information on the meanings of words. Figure 1 contains a sample of the concordance lines for *hands* in Dickens's fifteen novels. The concordance is sorted alphabetically on the first, second and third word to the left of *hands*. The sample focusses on the sequence *with both hands* resulting from this type of sorting. The three-word sequence is part of a longer sequence *him by the collar with both hands*. In WordSmith Tools such sequences are referred to as 'clusters'. With a span of five words to the left and right of the node, a cluster length of six and a minimum frequency of three, the software retrieves a list of fifty-one clusters from the concordance of *hands*. Figure 2 shows the bottom eleven clusters starting with *by the collar with both hands*.

Clusters are contiguous sequences of word forms, so the sequence in Line 146 of the concordance, *with both hands by the collar*, is not listed as part of the cluster *by the collar with both hands*. The concordance sample also shows that the patterns around *with both hands* contain verb forms that

```
41       BY THE COLLAR WITH BOTH HANDS
42       BOTH HIS HANDS AS IF HE
43       HER HANDS BEFORE HER FACE AND
44       HIS HANDS UPON HIS KNEES AND
45       HIS HANDS OUT OF HIS POCKETS
46       PUTTING HIS HANDS INTO HIS POCKETS
47       HIS HEAD RESTING ON HIS HANDS
48       HIS CHAIR WITH HIS HANDS IN
49       HER HANDS FOLDED ON ONE KNEE
50       HIS HANDS INTO THE POCKETS OF
51       HIS HANDS AS IF HE COULD
```

Figure 2: Final eleven clusters of length six and minimum frequency three derived from the concordance of *hands* in WordSmith Tools (Scott, 2016).

will not be picked up by clusters: *seize*, *taking* and *clutching* (Lines 141–3). Even the repetition of the same verb in Line 146 would not be reflected in a cluster as it is a different form. Nevertheless, *seize* is an important part of the patterns and shows the link between *collar* and *throat* in Figure 1.

This example illustrates the variety of co-occurrence patterns that create meaning in the language and the way in which the concordance display supports the identification of such meaningful patterns. However, a meaningful analysis of what a concordance shows is no straightforward matter. Sinclair (2003) exemplifies systematic strategies for approaching concordance lines, but generally the analysis of concordances tends to be seen as a more qualitative approach to corpus data. To account for the display of non-contiguous sequences, Cheng *et al*. (2006) propose the ConcGram tool.[6] They define a 'concgram' as 'all of the permutations of constituency variation and positional variation generated by the association of two or more words' (Cheng *et al*., 2006: 414). Such a definition would allow for *by the collar* to appear on both sides of *hands*. They argue that 'the notion of a concgram challenges the current view about word co-occurrences that underpins the KWIC display', suggesting that the practice of choosing a node as centre for the display can lead to a perception of hierarchy between the node and the context words associated with it.

Another proposal to develop the traditional corpus concordancer was put forward by O'Donnell (2008). He suggests a 'KWICgrouper' that supports the way in which a concordance analysis brings together lines with formal similarities so that 'meaning' or 'functional' groups (Mahlberg, 2005) can be identified. Applied to our *with both hands* example, KWICgrouper would support the sorting of concordances for instance in the following way: having identified *seize/seized* and *clutching* as verb forms to go with the pattern, KWICgrouper would check the concordance lines automatically for

[6] Following the ConcGram proposal, WordSmith Tools (Scott, 2016) also incorporates a concgrams function.

forms of these verbs grouping Lines 171 and 188 together with the examples in Lines 141, 143, 144 and 146 to suggest a functional group.

The example of *hands* shows how the display format of a concordance can support the identification of formal patterns that are associated with functions in the text. At the same time, association patterns can also be identified without recourse to a concordance display, as the various types of collocation measures, or techniques for generating clusters/n-grams, skipgrams, and so on, show. Following the initial identification of word associations, concordances can then serve as a way of providing contextual information for the units that have initially been identified without access to the wider co-text. In addition, the recent proposal of GraphColl (Brezina *et al*., 2015) demonstrates that there are other visualisations for collocations than concordance lines.

Our main argument here is that the display and sorting of concordances begins with the lexico-grammatical level, but it is also important to take into account the texts and text sections from which the concordance lines are generated and which affect the type of lexico-grammatical patterns that can be observed. Corpus studies of register variation clearly highlight the importance of this point. Highly frequent clusters, or 'lexical bundles', have been shown to play an important role in accounting for variation across registers (e.g., Biber *et al*., 1999; and Biber, 2006). The BNC and the way in which both the BNCweb and BYU-BNC support the analysis of patterns further highlights links between patterns and the types of texts they occur in. Our example of *hands* is illustrative of this point. The pattern *with both hands* identified in Figure 1, also appears in a concordance for *hands* in a reference corpus of nineteenth-century novels by authors other than Dickens. However, the only example that shows some similarity with the *collar* / *throat* pattern above is Example 1, from *Dracula*. Taking into account the wider context of the pattern highlights that not all examples from the Dickens corpus indicate violence in the same way as the example from *Dracula*. Examples with *throat* (Lines 152, 171 and 188 in Figure 1, from the 'Madman's Manuscript' in *Pickwick Papers*, *The Mystery of Edwin Drood* and *Barnaby Rudge*, respectively) seem to be more similar to Example 1 while examples with *collar* are still displaying strong emotion but seem to be of a less violent kind, as illustrated by Example 2, from *David Copperfield*. Here, David observes his aunt attacking Uriah Heep, though the action is presented in a comical way.

(*1*) , and catching him *by the neck* with both *hands*, dragged him back with

(*Dracula*)

(*2*) What was my astonishment when I beheld my aunt, who had been profoundly quiet and attentive, make a dart at Uriah Heep, and *seize him by the collar with both hands*!

> ‘You know what I want?’ said my aunt
> ‘A strait-waistcoat,’ said he.
> ‘No. My property!’ returned my aunt
> [. . .]
> Whether my aunt supposed, for the moment, that he kept her property in his neck-kerchief, I am sure I don’t know; but she certainly pulled at it as if she thought so. [. . .].
>
> (*David Copperfield*)

Our examples so far point to differences between authors and differences between individual novels. Moreover, lexico-grammatical patterns are also associated with text-internal variation, as shown by corpus studies into the distribution of phrases across sections within texts (e.g., Scott and Tribble, 2006; O’Donnell and Mahlberg, 2008; Mahlberg, 2009; Römer, 2010; and O’Donnell *et al*., 2012). Software that has been created to support the study of distributions within texts is Barlow’s (2016) WordSkew, which allows, for instance, a focus on the beginning or end of sentences. On a theoretical level, relationships between lexical and textual patterns have been described in terms of what Hoey (2005) calls ‘textual colligations’ (e.g., the tendency of a word to occur as the theme of sentences).

Another concept to capture the link between lexical and textual relations is the ‘local textual function’ (see Mahlberg, 2005, 2013), which describes the patterns of a (set of) lexical item(s) in a specific (set of) text(s). While the categories used to capture local textual functions are less neat than those used to express textual colligations, they are described with reference to the text at hand. At the same time, the concept of local textual functions highlights the need to better understand the textual properties that can usefully be related to lexico-grammatical patterns so that we can create corpus tools to support the investigation of such patterns. O’Donnell’s (2008) KWICgrouper was designed to support the analysis of the lexico-grammatical elements of local textual functions. In this paper, we emphasise the textual dimension. Mahlberg *et al*. (2013) already made a step in this direction by arguing that ‘suspensions’ are meaningful units in narrative fiction and deserve systematic attention using concordances (see also Mahlberg and Smith, 2010, 2012).

3. Creating CLiC

Direct speech might be seen as one of the more straightforward categories of speech and thought presentation. We designed CLiC with a focus on Dickens’s novels, where externalised techniques of characterisation (John, 2001) mean that direct speech is not a simple matter though. Moreover, the annotation of discourse categories, as in Semino and Short (2004), is generally done manually, whereas CLiC[7] works on the basis of automatically

[7] CLiC is available online at: clic.bham.ac.uk/.

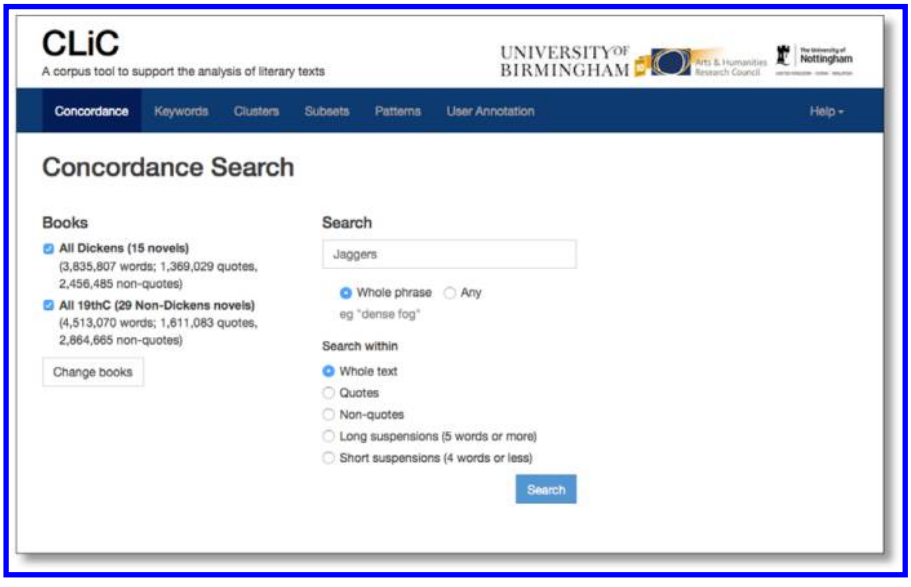


Figure 3: Search options in CLiC – with example *Jaggers* in 'Whole text'.

annotated texts. In Section 3.1 we will outline the sub-corpora-specific search options in CLiC and in Section 3.2 we give more background on the creation of the sub-corpora with precision and recall figures for the XML annotation on which the search options are based.

3.1 Search options and sub-corpora

In addition to Dickens's fifteen novels, CLiC also allows searches in a reference corpus of nineteenth-century novels written by other authors, based on the selection in Mahlberg (2013). Figure 3 shows the concordance function with *Jaggers* as search term for all these texts. The option 'Search within' requires the user to specify which discourse level to focus on. The result of this search returns a concordance for all instances of *Jaggers* in *Great Expectations*, as the only novel with a character of this name. It is important to note that CLiC includes results that are followed by punctuation (and so includes *Jaggers's*).

The 'Whole text' search is the norm for standard concordance tools. The four other options that CLiC provides can be illustrated with Example 3, which shows three consecutive paragraphs from *Great Expectations* (GE): 3a, 3b and 3c. Text between quotation marks is part of the sub-corpus 'Quotes', illustrated by the first two paragraphs in the example, 3a and 3b. In most cases Quotes will be the same as Direct Speech, although text within quotation marks can also be thought or writing. Given that these are less frequent options, we do not attempt to make this distinction. The

Quotes	DNov	19C	**Total**
	1,375,593	1,642,745	**3,018,338**
Non-quotes	2,463,585	2,860,409	**5,323,994**
Suspensions	*109,985*	*42,762*	*152,747*
Long suspensions	*84,613*	*28,951*	*113,564*
Total	**3,839,178**	**4,503,154**	**8,342,332**

Table 1: Word counts for subcorpora.[8]

third paragraph, 3c, exemplifies a 'Non-quote', defined as text that does not appear within quotation marks. Paragraphs 3a and 3b illustrate a sub-type of Non-quotes, called a 'suspension' – an interruption of a character's speech by narrator text, following Lambert (1981). For Lambert (1981), such an interruption has to be at least five words long – this would be a 'long suspension' for CLiC, as in Example 3a (whereas Example 3b is a short suspension). Suspensions, italicised in the example below, appear in the same sentence as Quotes. So if there were a full stop after *nose* in Example 3a, there would not be a suspension. With these definitions, a search for *Jaggers* in Non-quotes finds both Examples 3a and 3c, whereas a search in long suspensions only returns Example 3a.

(*3a*) "And on what evidence, Pip," *asked Mr. Jaggers, very coolly, as he paused with his handkerchief half way to his nose*, "does Provis make this claim?"

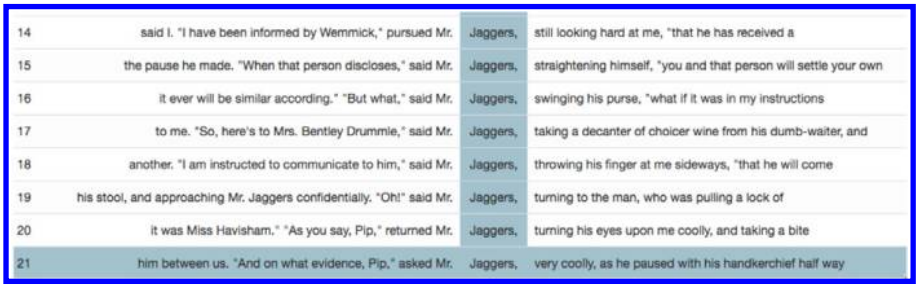
(*3b*) "He does not make it," *said I*, "and has never made it, and has no knowledge or belief that his daughter is in existence."

(*3c*) For once, the powerful pocket-handkerchief failed. My reply was so unexpected that Mr. **Jaggers** put the handkerchief back into his pocket without completing the usual performance, folded his arms, and looked with stern attention at me, though with an immovable face.

(GE)

Word counts for the resulting sub-corpora are shown in Table 1. This suggests a number of points for comparisons between Dickens and the reference corpus – which will be of specific interest from a distant reading point of view. Irrespective of detailed quantitative information, however, a crucial observation is already the following: while the literature suggests that the suspended quotation is a technique of Dickens's style (e.g., Lambert, 1981; Newsom, 2001; and Horne, 2013), being able to search suspensions across other novels underlines the prevalence of this phenomenon. If a more

[8] These counts have been generated offline because the Cheshire3 counts are inconsistent. For this reason future releases of CLiC will be replacing the current Cheshire3 database.



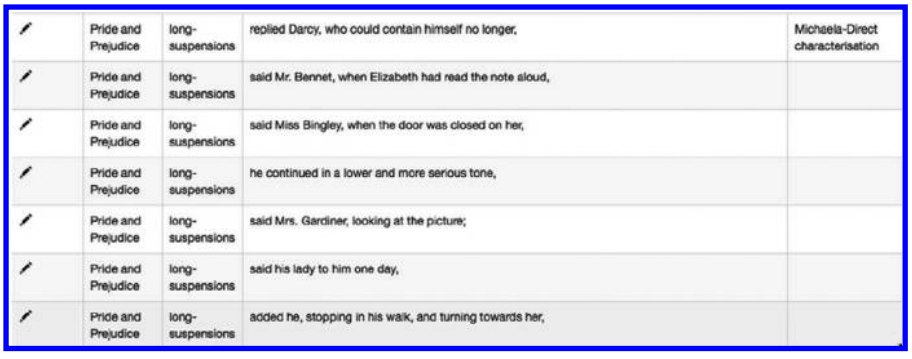
14	said I. "I have been informed by Wemmick," pursued Mr.	Jaggers,	still looking hard at me, "that he has received a
15	the pause he made. "When that person discloses," said Mr.	Jaggers,	straightening himself, "you and that person will settle your own
16	it ever will be similar according." "But what," said Mr.	Jaggers,	swinging his purse, "what if it was in my instructions
17	to me. "So, here's to Mrs. Bentley Drummle," said Mr.	Jaggers,	taking a decanter of choicer wine from his dumb-waiter, and
18	another. "I am instructed to communicate to him," said Mr.	Jaggers,	throwing his finger at me sideways, "that he will come
19	his stool, and approaching Mr. Jaggers confidentially. "Oh!" said Mr.	Jaggers,	turning to the man, who was pulling a lock of
20	it was Miss Havisham." "As you say, Pip," returned Mr.	Jaggers,	turning his eyes upon me coolly, and taking a bite
21	him between us. "And on what evidence, Pip," asked Mr.	Jaggers,	very coolly, as he paused with his handkerchief half way

Figure 4: Sample from the twenty-two lines for *Jaggers* in long suspensions.

fine-grained break-down is used, it becomes apparent that suspensions can be found in every text in the two corpora.

With CLiC's sub-corpora, concordance searches make it possible to complement the description of lexico-grammatical patterns by taking further textual dimensions into account. Figure 4 shows a selection of the twenty-two lines that are the result of a search for *Jaggers* in long suspensions. This is a relatively small sample compared to more than 2,600 examples for *hands* in Section 2.2, and this reflects the difference between focussing on a single text and looking across a range of texts. The more narrow focus has implications for the kind of patterns that a concordance can show. The patterns to the left (*said Mr. Jaggers*) are more similar to the formal patterns outlined in Section 2.2. The concordance is sorted on the first word to the right, which highlights the repetition of *turning*. In addition, *coolly*, which appeared in Example 3, corresponding to Line 21 in the concordance, is repeated in Line 20. If patterns are not restricted to verbatim repetition, there are several lines showing Jaggers as a cool and focussed character (e.g., in Line 20 he is turning his eyes *coolly* on Pip and in Line 14 he is *looking hard* at him).

By running concordances for character names, suspensions are a potentially useful place to check a text for character information, especially in the form of descriptions of body language. This point is illustrated in more detail in Mahlberg and Smith (2012), Stockwell and Mahlberg (2015) and Mahlberg and Stockwell (2016). Furthermore, a type of concordance can also be run without a node word (such as a character name) to start with. CLiC makes it possible to list all the suspensions in a text for closer analysis. Figure 5 is an example retrieved with CLiC's User-annotation component focussing on long-suspensions in *Pride and Prejudice*. The User-annotation makes it possible to add user-defined tags to help classify concordance lines (e.g., in Figure 5 a tag 'direct characterisation' is added by user 'Michaela' to mark-up suspensions that contain relative clauses). In the annotation view, suspensions can also be filtered to find further examples containing the relative pronoun *who*. This way of using concordances significantly improves the practicalities of a study like Mahlberg and Smith (2010) which classified all suspensions in *Pride and Prejudice* – but at the time did not have the functionality of CLiC to support the analysis.



✎	Pride and Prejudice	long-suspensions	replied Darcy, who could contain himself no longer,	Michaela-Direct characterisation
✎	Pride and Prejudice	long-suspensions	said Mr. Bennet, when Elizabeth had read the note aloud,	
✎	Pride and Prejudice	long-suspensions	said Miss Bingley, when the door was closed on her,	
✎	Pride and Prejudice	long-suspensions	he continued in a lower and more serious tone,	
✎	Pride and Prejudice	long-suspensions	said Mrs. Gardiner, looking at the picture;	
✎	Pride and Prejudice	long-suspensions	said his lady to him one day,	
✎	Pride and Prejudice	long-suspensions	added he, stopping in his walk, and turning towards her,	

Figure 5: Sample of long-suspensions in *Pride and Prejudice* viewed in User-annotation.

3.2 Precision and recall for the annotation

To create the CLiC corpora, we used plain text files from Project Gutenberg[9] and converted them to XML files with the help of a series of Python scripts. The XML database we use is Cheshire3.[10] Cheshire3 is also queried with Python scripts. Figure 6 illustrates the overall workflow. In this paper, we focus on the conversion from txt to XML. The CLiC code and the XML corpora are available online,[11] see also Appendix A.

The initial conversion of the text files to XML marks chapter divisions, paragraphs and sentences using the structure of the text files themselves. To identify the quoted passages in the texts we used an algorithm centred around two regular expressions: one for identifying quotations using double quotation marks and another for single quotation marks. The transcriptions available on project Gutenberg typically use either single or double quotes for an entire book (although there are errors in Gutenberg) so the transcriptions could be split into single or double quotation transcriptions and the appropriate regular expression used.

While chapters, paragraphs and sentences form a neat, nested hierarchy that can be dealt with easily in XML, the same cannot be said of quotations. These can span sentence and paragraph[12] boundaries and thus pose a problem for XML which does not allow such overlapping hierarchies. It is common to circumvent this limitation of XML by using empty elements, known as milestones (<milestone/>), as place markers rather than XML elements (<element></element>) (Marinelli *et al*., 2008; and Iacob *et al*., 2004). Hence, the XML elements that form our nested hierarchy, such

[9] See: https://www.gutenberg.org/.
[10] See: http://cheshire3.org/.
[11] See: clic.bham.ac.uk (current release 1.4).
[12] Our approach differs from Elson and McKeown (2010), see Section 2.1, who define quoted speech as blocks of text within a paragraph.

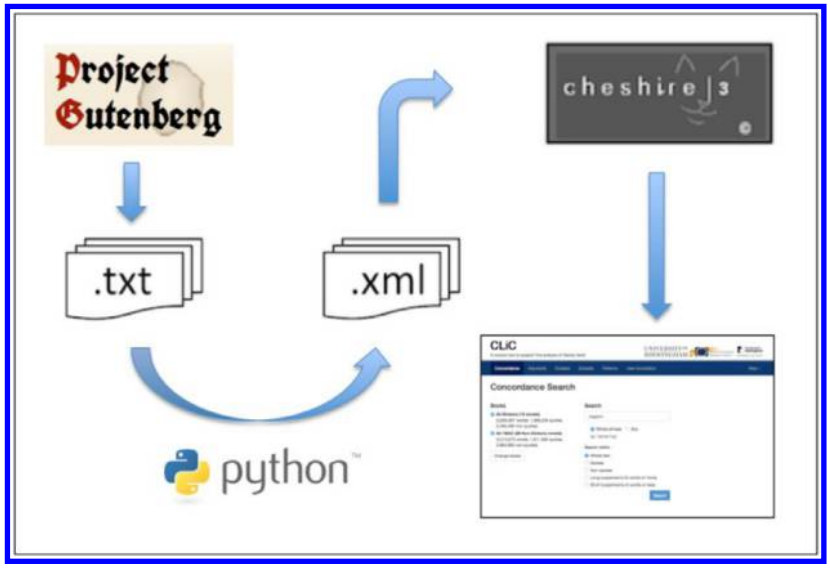


Figure 6: Workflow for Gutenberg to CLiC.

as sentences, contain the text of the sentence between an opening element (< s >) and a closing element (< /s >), as in Example 4:

(*4*) < s > "And on what evidence, Pip," asked Mr. Jaggers, very coolly, as he paused with his handkerchief half way to his nose, "does Provis make this claim?" < /s >

In contrast, the start and end of each quotation and each suspension is marked with an empty element used simply as a marker of the position. Once the sentence above is fully annotated the result is as per Example 5:

(*5*) < s >
< qs/ > "And on what evidence, Pip," < qe/ >
< sls/ > asked Mr. Jaggers, very coolly, as he paused with his handkerchief half way to his nose, < sle/ >
< qs/ > "does Provis make this claim?" < qe/ >
< /s >

Here < qs/ > marks the start of a quotation and < qe/ > the end, < sls/ > and < sle/ > mark the beginning and end of a long suspension, respectively. We also use < sss/ > and < sse/ > to mark the beginning and end of short suspensions (see also Mahlberg and Smith, 2012: 54–5). Longer quotations which span paragraphs are marked with the same < qs/ > and < qe/ > tags but further attributes are added to the paragraph tags that mark the paragraph as being a part of an extended quotation and to give an indication as to

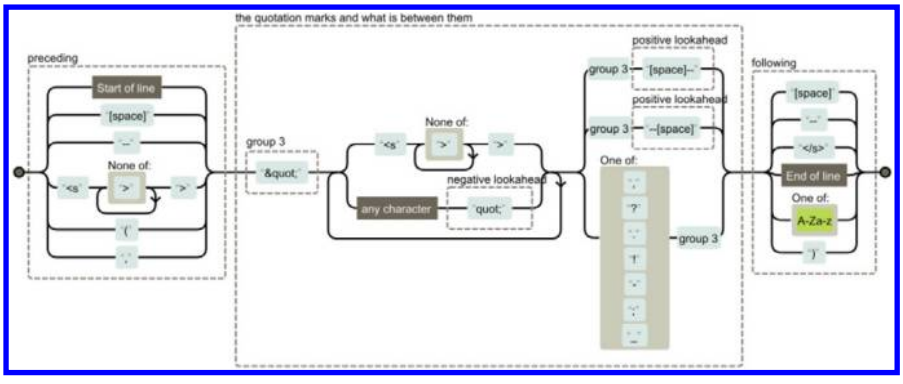


Figure 7: Regular expression for double quotation marks.

whether it is the first, last or an intermediate paragraph in the extended quotation. The indexing software used to process the XML is able to treat the text between our start and end markers as though it were contained within an element like the sentence example given above.

Figure 7 illustrates the regular expressions for double quotation marks.[13] The quoted text can be preceded by a space, a sentence tag, a double hyphen (which mimics an en-dash), a left bracket or a comma. It can be followed by a space, double dash, sentence end tag, end of line, or an alphanumeric character, or a right bracket. The regular expression for single quotes is essentially the same, but it contains more complicated exceptions which match the content of the quote itself because single quotes are also used as apostrophes.

We manually cleaned and corrected a large set of quotes in a gold standard text[14] (see also Appendix A) to be able to compute precision and recall figures. The gold standard consists of 1,033 randomly selected paragraphs equally distributed over DNov and 19C. Each paragraph contains at least one quote or is part of an extended Quote (i.e., a Quote crossing paragraph boundaries). Each book is represented by at least three speech paragraphs. Precision characterises the proportion of annotated Quotes that are genuine Quotes. Recall refers to the proportion of annotated Quotes in relation to the total number of actual Quotes found in the text, indicating how complete the Quote annotation is (Manning and Schütze, 2000: 267–8, 534–5). Precision and recall are interdependent, as illustrated by Example 6 from *Little Dorrit*, which lacks a quotation mark at the start. The annotation wrongly extracts a Quote starting at *said Mr Meagles*, and ending with *a tight one*. So it misses the two actual Quotes (recall) and wrongly identifies a Non-quote as a Quote (precision). This is the only mistake in the gold standard for *Little Dorrit* (which contains thirty-nine quotes), but it results in

[13] This is based on https://regexper.com/ which has been adapted to our purposes.
[14] Available online at: clic.bham.ac.uk.

97.4 precision and 94.9 recall. The effect is therefore that if the automated annotation mistakenly identifies a Non-quote as a Quote (a mistake reflected in precision) this regularly results in a Quote not being identified (a mistake reflected in recall).

(*6*) I require a deal of pulling through, Arthur,' said Mr Meagles, shaking his head, 'a deal of pulling through. I stick at everything beyond a noun-substantive–and I stick at him, if he's at all a tight one.'

(*Little Dorrit*)

Frequent mistakes are due to inconsistencies in the input text file which either lacks a space or a quotation mark, or adds a quotation mark where one would not expect one, or does not alternate the type of quotation mark used in embedded quotes (for an example of the latter, see the first line in Figure 8 of *portable property* in Section 4). Example 7 shows how single and double quotation marks are mixed:

(*7*) "Ain't I ollays quiet, miss? Did anybody ever hear me rampage? If you please, ma'am, the squire's come home.'

(*The Small House at Allington*)

As Table 2 (p. 449) indicates, precision and recall are higher for the Dickens corpus. This is partly a result of our focus on this author and partly a reflection of the reference corpus containing texts by a variety of authors. The values are also higher for novels that use double quotation marks. Single quotation marks are more complex because of the need to disambiguate their other use as apostrophes. Precision and recall figures are specifically affected by the number of extended speech paragraphs. *Frankenstein* and *Armadale* dominate the gold standard for 19C because they contain many, long extended quotes (sometimes as long as an entire chapter). Hence, Table 2 also presents figures without these books. The gold standard contains a selection of speech paragraphs regardless of whether they are part of extended quotes. If, therefore, a randomly selected speech paragraph is part of an extended quote, the paragraphs before and after that are part of the extended quote were added to the gold standard and any inaccuracy found within the entire extended quote has an effect on precision and recall. The high accuracy of quotation annotation also leads to high accuracy of suspension annotation.

4. Enabling new insights into fictional speech

In this section, we focus on fictional speech to illustrate how a tool like CLiC can contribute to the exploration and testing of the notion of 'mind-modelling'. This cognitive poetic notion has its origins in cognitive

	Subcorpus	Precision (%)	Recall (%)
DNov	Quotes	98.59	96.39
	Quotes single quotation marks	98.30	95.67
	Quotes double quotation marks	100	100
	Suspensions	100	94.53
19C without Frankenstein and Armadale	Quotes	99.39	96.29
	Quotes single quotation marks	98.82	93.33
	Quotes double quotation marks	99.51	96.92
	Suspensions	96.0	94.12
19C (italicised values same as for 19C without two novels)	Quotes	93.32	79.29
	Quotes single quotation marks	*98.82*	*93.33*
	Quotes double quotation marks	92.49	77.39
	Suspensions	96.0	57.14

Table 2: Precision and recall figures.

psychological research on 'Theory of Mind'. This is an explanation of the phenomenon whereby a person seems to be able to make assumptions about other people's beliefs, dispositions, states of mind and intentions. Beginning from around the age of three, and developing rapidly during adolescence, neurotypical children become increasingly adept at understanding and predicting the states of mind of others. This is based on a presumption (a 'Theory') that those other people are people just like oneself, with a conscious awareness, and a similar palette of perceptions and human conditions (see Premack and Woodruff, 1978; Baron-Cohen, 1997; Carpendale and Lewis, 2006; and Apperly, 2011).

Adapted to the peculiar, displaced scenario of literary reading, the presumption of a Theory of Mind (ToM) is applied by projection to imaginary and fictional minds, just as actual, real minds are rendered psychologically (see Leverage *et al*., 2011). As in real life, running our ToM capacity is what allows us to form conclusions about the knowledge and beliefs of characters, and to engage in empathetic relationships with fictional people. Since this process, especially in a literary experience, is active and creative, it has been called 'mind-modelling' (Stockwell, 2009; see also the term 'mind-reading' in Turner, 1992, and Zunshine, 2006).

It is clear that a reader has the following textual patterns (among others) generally available as the raw material for mind-modelling a character's mind (see Stockwell and Mahlberg, 2015: 134, for a more comprehensive list):


Searched for **portable property** within **non-quotes**.

1 to 3 of 3 entries CSV Print Toggle metadata Filter concordance:

	Left	Node	Right	Book	Ch	Par	Sen	In bk
1	... myself, my guidingstar always is, "Get hold of	portable property"."	When I had rendered homage to this light, he ...	Great Expecta...	24	42	108	
2	... Wemmand I judged her to stand possessed of	portable property.	The cut of her dress from the waist upward, b...	Great Expecta...	37	12	32	
3	nose and a very new moon, was a piece of	portable property	that had been given her by Wemmick. We ate ...	Great Expecta...	37	25	68	

1 to 3 of 3 entries Filter concordance:

Figure 8: Results for search of *portable property* in Non-quotes.

(*1*) Direct descriptions of physical appearance and manner, gestures and body language; and,
(*2*) The presentation of speech for an apparently autonomous sense of characters' personality, mood and perspective.

In a long novel, the textual markers of character-building and mind-modelling are almost always diffused across the entire text. CLiC can help to identify them and group them for close analysis, illustrating the potential of the concordance display to 'zoom in' on places that provide character information. The two types of textual patterns listed here are only examples from a much more extensive list, but they are examples which seem to be particularly suited for study with CLiC. CLiC's capacity for differentiating between speech and non-speech narratorial framing, as well as its identification of suspensions of varying lengths between speech, offer an opportunity for pinpointing features from Point 1, as we argued in Section 3.1.

In this section, we focus specifically on Point 2: the presentation of speech. Dickens is well known for his use of repeated phrases or 'speech tics' as a technique of characterisation; for instance, the habitual phrase *portable property* associated with Wemmick in GE. The point of such habitual phrases is that they are striking and noticeable. Brook (1970: 143) observes: 'It may be [. . .] that part of the secret of Dickens's success is that he makes things easy for his readers by his constant repetitions, and his habitual phrases are remembered by readers who are not used to reading with close attention'. A concordance can be used to trace such repeated phrases throughout the text and in this sense support the literary critic's close reading. Given the strikingness of such phrases it might be argued that it is not even necessary to run a concordance for them – concordances are generally seen to support the identification of less obvious patterns. However, what is less obvious about fictional characters' habitual phrases is how they are used by the narrator. The phrase *portable property* occurs thirteen times in GE. A search in Non-quotes returns three lines (see Figure 8).

Below are the two examples from Chapter 37. In Example 8, the narrator, Pip, sees Miss Skiffins for the first time. His assessment of her as standing 'possessed of portable property' is a reflection of her relationship

with Wemmick, who is obsessed with 'portable property'. In a similar way, in Example 9, Pip comments on the brooch that Miss Skiffins is wearing as 'portable property' because it is a present from Wemmick.

(*8*) Miss Skiffins was of a wooden appearance, and was, like her escort, in the post-office branch of the service. She might have been some two or three years younger than Wemmick, and I judged her to stand possessed of *portable property*.

(*9*) I inferred from the methodical nature of Miss Skiffins's arrangements that she made tea there every Sunday night; and I rather suspected that a classic brooch she wore, representing the profile of an undesirable female with a very straight nose and a very new moon, was a piece of *portable property* that had been given her by Wemmick.

(GE)

Both of these examples can be seen as instances of free indirect discourse (FID) in that Wemmick's characteristic verbal tic is assimilated into the narratorial discourse of Pip. Given that the homodiegetic Pip is not like a heterodiegetic omniscient author–narrator, the use of FID might seem surprising – and it might then motivate a search for further examples of this intriguing narratorial style in the novel. Furthermore, the examples here demonstrate the narratological argument that the different elements comprising the FID are not an improbable form of blended consciousness: instead, one mind is presented as being deflected through another (as Stockwell, 2013: 273, suggests). In this example, CLiC provides material for further textual research on the novel, and it validates theoretical claims made in general.

The example of *portable property* also illustrates a point affecting our precision and recall figures. Line 1 in Figure 8 is listed as Non-quote; however, the extended context in Example 10 shows that the example is one of quotation marks within Quotes. This is the first instance of Wemmick using the phrase, and in effect explaining its relevance.

(*10*) "Oh yes," he returned, "these are all gifts of that kind. [. . .] It don't signify to you with your brilliant look-out, but as to myself, my guidingstar always is, "Get hold of *portable property*"."

(GE)

As the example of *portable property* illustrates, when speech is discussed with regard to the creation of fictional characters, the focus tends to be on how speech individualises characters in the sense of making them different from other fictional people. This is also underlined through corpus studies that use key words to compare the speech of different characters (see discussion in Section 2). Specifically in terms of Dickens, the idiolects or

speech tics (Brook, 1970) of his characters have received much attention. The annotation we describe in Section 3 is not designed to mark-up the speech of individual characters, but focusses on speech across characters. This is where CLiC creates another theoretical link between mind-modelling and corpus linguistics. The similarity between fictional people and real people that is fundamental to the concept of mind-modelling means that features in the text can function to differentiate a fictional character away from the reader's model of a person. Foregrounded features in the text are evidence of idiolects and individual speech behaviour. At the same time, patterns in the text can also function to strengthen similarities across characters and the impression of naturalness of a fictional character's speech. Such backgrounded features connect to the reader's background knowledge in the top-down activation of knowledge.

The innovative contribution CLiC makes to the study of fictional speech becomes even clearer, when we consider how fictional speech has mainly been approached so far. Page (1988: 7) observes: 'there is an inevitable gap – wider or narrower at different times, but never disappearing entirely – between speech, especially in informal situations, and even the most "realistic" dialogue in a work of literature'. Page (1988: 7ff.) further argues that there are at least three reasons for this:

(*1*) Differences in the medium of spoken form and written representation of speech;
(*2*) The context of situation which is crucial for spoken language is only partially presented in a fictional text; and,
(*3*) The phonological component of spoken language contributes to meaning, too.

However, Page (1988: 3ff.) also makes another observation that is crucial to our argument, questioning whether the notion of realism in fictional speech is 'often based on an inadequate or inaccurate notion of what spontaneous speech is really like'. A major achievement of corpus linguistics has been to find evidence of what spontaneous speech is like. This has led to rather radical changes in the way in which we describe spoken language (Carter and McCarthy, 2010; and Leech, 2000). In particular, what are called 'chunks', 'clusters' or 'lexical bundles' in speech have contributed to our understanding of the way in which spoken language works in context. So while the situational context might still only be partially presented in fictional texts (as Page, 1988, suggests), the occurrence of such speech patterns does reflect this context.

Together with corpus linguistic findings based on real spoken language, CLiC illustrates how 'general' fictional speech can be studied. The general speech patterns that are identified in this way are relevant to those aspects of mind-modelling that enforce the naturalness of fictional characters. To study such speech patterns, generating clusters can be a useful starting point. Whereas concordances need a search word, or a 'node' to

begin the exploration, clusters are a way of listing patterns irrespective of a node. To narrow down an initial overview of clusters, Table 3 shows the top fifteen 5-word key clusters for, firstly a comparison of Quotes against Non-quotes and, secondly a comparison of Non-quotes against Quotes. This illustrates clear phraseological differences between the fictional speech among characters and the way in which narrators describe the fictional world. In terms of mind-modelling, this also shows a distinction between the different fictional minds (character *versus* narrator) with whom a reader must engage.

Table 3 shows that key clusters in Quotes reflect the speaker–listener world of the characters – indicated by the first- and second-person pronouns. Clusters that are key in Non-quotes, however, illustrate the narrator's role in describing characters' body language (e.g., *his hands in his pockets*, *leaning back in his chair* and *with his back to the*), in commenting on and interpreting the characters' behaviour as reflected in *as if* clusters (*as if he had been*), and in locating the narrative with reference to place and time (*up and down the room*). Mahlberg (2013) discussed these groups, focussing on the surface features of the actual clusters, such as the presence of *as if*, the occurrence of pronouns or body-part nouns. In Mahlberg (2013) the clusters were generated across the texts as a whole and the groups of clusters indicated differences between the ways in which characters and narrators contribute to creating different aspects of the fictional world. The key comparison of clusters in Quotes and Non-quotes now provides further support for this classification.

With the help of text-internal comparisons, the classification can also be extended. The cluster *and all the rest of* cannot be as neatly classified as the examples in Table 3, if features like pronouns, body-part nouns, and so on, are drawn on alone. However, it is a key cluster when Quotes are compared to Non-quotes ($LL = 14.92$, $p < 0.001$) – underlining the spokenness of the cluster so that it can be grouped with the speech clusters. At the same time, it is an example of what Carter and McCarthy (2006: 202) refer to as 'purposeful' vagueness.

(*11*) From the village school of Chesney Wold, intact as it is this minute, to the whole framework of society; from the whole framework of society, to the aforesaid framework receiving tremendous cracks in consequence of people (iron-masters, lead-mistresses, and what not) not minding their catechism, and getting out of the station unto which they are called–necessarily and for ever, according to Sir Leicester's rapid logic, the first station in which they happen to find themselves; and from that, to their educating other people out of THEIR stations, and so obliterating the landmarks, and opening the floodgates, *and all the rest of* it; this is the swift progress of the Dedlock mind.

(*Bleak House*)

n-gram	Q	NQ	LL	*n*-gram	NQ	Q	LL
what do you mean by	67	0	151.97	*his hands in his pockets*	77	0	59.78
what do you think of	57	1	119.96	*with his hands in his*	51	0	39.59
i beg your pardon sir	49	0	111.14	*as if he had been*	71	3	36.82
very much obliged to you	37	2	69.7	*up and down the room*	42	0	32.61
i should like to know	30	0	68.05	*as if it were a*	57	2	31.32
to tell you the truth	33	1	66.6	*as if he would have*	37	0	28.73
i want to speak to	29	0	65.78	*leaning back in his chair*	37	0	28.73
i am glad to see	28	0	63.51	*hands in his pockets and*	36	0	27.95
you don't mean to say	28	0	63.51	*with the air of a*	41	1	24.65
do me the favour to	30	1	59.99	*with his back to the*	40	1	23.92
i am very glad to	26	0	58.97	*the opposite side of the*	51	3	23.23
i am not going to	29	1	57.79	*was in a state of*	29	0	22.51
upon my word and honour	25	0	56.7	*in a corner of the*	29	0	22.51
how do you do mr	25	0	56.7	*back in his chair and*	26	0	20.19
am glad to see you	24	0	54.44	*the gen l m n*[1]	25	0	19.41

[1] The 5-gram 'the gen l m n' is retrieved because CLiC splits tokens on whitespace and punctuation.

Table 3: Top fifteen 5-word key clusters for Quotes *versus* Non-quotes and Non-quotes *versus* Quote, $p < 0.0001$.

Example 11, which is from *Bleak House*, with its unusual omniscient third-person but present-tense narration, is a passage that slips from what seems at first to be a purely narratorial level, into Sir Leicester Dedlock's consciousness. Initial cues that the narration is slipping towards FID can perhaps be discerned in the spokenness of phrases such as *and what not* and *not minding*, and then signalled more strongly by the capitalisation for spoken emphasis of *THEIR*; but it is the speech cluster *and all the rest of* it that indicates finally we have moved into 'the Dedlock mind'. Dickens underlines the fact explicitly for his less sensitive readers at the end of the extract. Subtle texture such as this seems easy to be missed analytically without the sort of functionalities that CLiC offers.

5. Conclusion

Although in the space of this paper our discussion of examples had to be selective, we have made a number of far-ranging methodological and theoretical points. Relevant textual patterns of character information do not have to be verbatim repetitions, but also extend into more complex contexts. In this sense, their identification supports and complements claims in literary criticism (see the example of Wemmick) and adds systematicity to close reading (see the example of Jaggers). As per the arguments favoured by distant reading, the corpus view beyond Dickens showed that suspensions are a wider phenomenon and not just a Dickensian technique. When data across texts is accumulated, we see general, shared patterns, such as narratorial accounts of body language, but also shared speech phrases. This is an important point that highlights how corpus methods provide evidence for claims of mind-modelling. In particular, our discussion of fictional speech showed how individual characters do not only rely on idiosyncratic phrases. The naturalness of the characters' words reflected by shared speech phrases is equally important. Through the cumulative picture of fictional speech, corpus methods broaden the view from bottom-up cues in an individual text to a more general account of fictional speech patterns across texts that affect the top-down processes that are relevant to mind-modelling. Fictional characters are not only defined by features that differentiate them from others but also by features that make them similar to other characters and to other people. From cognitive poetics, we treat text-drivenness as the principle domain of analysis for our exploration of readerliness and the evidence from comparing textual patterns across different narrative texts can suggest similarities in readerly experiences. In line with other work in corpus linguistics, CLiC obviously concentrates on the retrieval of replicable textual data. More research is still needed to investigate how the patterns we identify are processed by readers (see Mahlberg *et al*., 2014, for an initial suggestion).

The sub-corpora we created and the way in which CLiC accesses them have wider implications on a theoretical level. The perspective provided

by concordances has traditionally maintained a focus on the lexical and phraseological level as the unit of analysis. The concept of local textual functions highlights the need to go beyond concordance lines. In our work with CLiC, we have tried to adopt an approach that follows a principle of language as discourse – in the applied linguistic rather than critical theoretical sense (see Cook, 1994; McCarthy and Carter, 1994; and Howarth, 2000). That is, we take text as the unit of analysis, and use corpus linguistic methods to explore the principle of text from its patterning. This takes into account lexical and phraseological units, but also descriptions of demeanour and body language, narratorial suspensions, and other textual traces of different levels of consciousness and fictionality. Hence our approach uses the concordance for an analysis of discourse. At present, CLiC does not distinguish between direct speech and thought (or writing). The exploration of this dimension can add even further detail to our view of discourse.

In developing CLiC, we were motivated by a desire to find a common ground between corpus stylistics and cognitive poetics – the two most innovative, productive and insightful developments in literary linguistic analysis and criticism of recent decades. We have found that using CLiC to explore readerly effects in Dickens has led to a greater integrative approach than we had expected. An initial presumption was that we would be able to use corpus linguistic tools and methods in order to test cognitive poetic claims about texture; and then validate, reject or revise those claims; and then produce a richer, more complex and more compelling account of the interaction of textual patterns and readerly effects than had previously been possible. This initial approach represents a use of corpus linguistics in the service of cognitive poetics.

A second line of inquiry presumed that we could use the subtle, speculative and complex work in cognitive poetics as a means of making corpus linguistic methods more complex, more discourse-focussed, and able to explore equally subtle and textually diffused features in literary works. In other words, we envisaged an interdisciplinary project in which one field was viewed from the vantage point of the other, with a trajectory one way or the other, respectively. In the course of our work, we have been able to assure ourselves that both of these interdisciplinary trajectories have been possible, and have rendered tangible results for the benefit of both disciplines.

However, we have also learned that there is more than an interdisciplinary common ground between corpus linguistics and cognitive poetics. A multidiscipline has emerged in which theory and technique from both sources can be integrated and developed together. The example of FID outlined briefly in this paper illustrates this: a traditional stylistic feature is interrogated by cognitive poetics, and then a concordance exploration suggests further theoretical complexity for the notion, and this can be verified with reference to principled cognitive psychological patterns, and in turn then explored further using concordance and cluster searches. There is a payoff for cognitive poetics, narratology, corpus stylistics and literary criticism.

A particular facility of CLiC is the ease with which it differentiates quoted material from non-quoted material, and can identify and present suspensions in character speech. This allows for a rich exploration of the narrative embedding of consciousness that can be applied to all forms of literary narrative, and of course to any instances of narrative recount in which the narrator is displaced from the time or place of the narrated story itself. These interactions of minds – actual and fictional – are not telepathic nor abstract, but are text-driven. There are textually manifest traces by which a reader can build worlds and fictional minds: they are already available for discovery and exploration; they are not self-generated phenomena nor artefacts of the analytical process or critical framework itself. Our integrated approach offers a method for discovery that is not a critical theory in this sense.

Our research for this project has necessarily had a proper focus on one literary domain – here, Dickens's prose fiction. We have been able to make a contribution to Dickensian literary criticism, particularly in relation to characterisation, which we hope is valuable and suggestive. Of course, another way of regarding this work is to see it as a case-study for research in narratology, poetics, literary theory and critical theoretical innovation in general. For example, we have been concerned to answer particular research questions about the uses of fictional speech and narratorial body language in prose fiction, about readerliness and readerly effects in engaging empathetically with fictional minds, and about the complexities involved in understanding the interplay of psychology and a text. In short, we have been interested in exploring issues that are authentic for all readers, and using our best current understanding of reading and textuality as our integrated analytical tool.

The way in which our work with CLiC has highlighted multidisciplinary concerns of corpus stylistics and cognitive poetics has implications for developments in the Digital Humanities more widely. We certainly need specific technical knowledge and skills to preserve, access and analyse electronic text or artefacts more generally. At the same time, research under the digital umbrella allows us to ask new research questions and provides new avenues for interdisciplinary work in the humanities.

References

Altmeyer, S., C. Klein, B. Keller, C.F. Silver, R.W. Hood and H. Streib. 2015. 'Subjective definitions of spirituality and religion: an exploratory study in Germany and the US', International Journal of Corpus Linguistics 20 (4), pp. 526–52.

Anthony, L. 2015. AntConc (Version 3.5) [Computer Software]. Tokyo: Waseda University. Accessed May 2016 at: http://www.laurenceanthony.net/.

Apperly, I. 2011. Mindreaders: The Cognitive Basis of 'Theory of Mind'. New York: Psychology Press.

Bakhtin, M. 1982. The Dialogic Imagination. (Edited and translated by M. Holquist and C. Emerson.) Austin: University of Texas Press.

Baron-Cohen, S. 1997. Mindblindness: An Essay on Autism and Theory of Mind. Cambridge: MIT Press.

Balahur, A., R. Steinberger, E. van der Goot, B. Pouliquen and M. Kabadjov. 2009. 'Opinion mining on newspaper quotations' in 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technologies Technology. Accessed May 2016 at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.156.9753.

Barlow, M. 2016. 'WordSkew: linking corpus data and discourse structure', International Journal of Corpus Linguistics 21 (1), pp. 104–14.

Bednarek, M. 2010. The Language of Fictional Television: Drama and Identity. London: Continuum.

Bergler, S. 2005. 'Conveying attitude with reported speech' in J.C. Shanahan, Y. Qu and J. Wiebe (eds) Computing Attitude and Affect in Text: Theory and Applications. Dordrecht: Springer.

Biber, D. 2006. University Language: A Corpus-Based Study of Spoken and Written Registers. Amsterdam: John Benjamins.

Biber, D., S. Conrad, E. Finegan, G. Leech and S. Johansson. 1999. Longman Grammar of Spoken and Written English. Harlow: Longman.

Brezina, V., T. McEnery and S. Wattam. 2015. 'Collocations in context: a new perspective on collocation networks', International Journal of Corpus Linguistics 20 (2), pp. 139–73.

Brook, G.L. 1970. The Language of Dickens. London: Andre Deutsch.

Busse, B. 2010. Speech, Writing and Thought Presentation in a Corpus of Nineteenth-Century English Narrative Fiction. Bern: University of Bern.

Carpendale, J.I.M. and C. Lewis. 2006. How Children Develop Social Understanding. Oxford: Blackwell.

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

Cheng, W., C. Greaves and M. Warren. 2006. 'From n-gram to skipgram to concgram', International Journal of Corpus Linguistics 11 (4), pp. 411–33.

Cook, G. 1994. Discourse and Literature: The Interplay of Form and Mind. Oxford: Oxford University Press.

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

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

Davies, M. 2010. 'More than a peephole: using large and diverse online corpora', International Journal of Corpus Linguistics 15, pp. 405–11.

Dekhtyar, A. and I.E. Iacob. 2005. 'Processing xml documents with overlapping hierarchies' in JCDL'05. Proceedings of the 5th ACM/IEEE-CS Joint Conference. Accessed May 2016, at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4118612.

Elson, D.K. and K.R. McKeown. 2010. 'Automatic attribution of quoted speech in literary narrative', Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2010). Atlanta, Georgia. Accessed May 2016 at: http://www.cs.columbia.edu/~delson/pubs/AAAI10-ElsonMcKeown.pdf.

Elson, D.K., N. Dames and K.R. McKeown. 2010. 'Extracting social networks from literary fiction', Proceedings ACL '10 Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 138–47. Uppsala, Sweden. Accessed December 2016 at: http://www.aclweb.org/anthology/P10-1015?CFID=699532838&CFTOKEN=79981142.

Giovanelli, M., A. Macrae, F. Titjen and I. Cushing. 2015. AQA English Language and Literature Student Book. Cambridge: Cambridge University Press.

Glass, K. and S. Bangay. 2007. 'A naive, salience-based method for speaker identification in fiction books' in Proceedings of the 18th Annual Symposium of the Pattern Recognition Association of South Africa (PRASA '07), pp. 1–6. Accessed December 2016 at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.494.3729.

Hardie, A. 2012. 'CQPweb – combining power, flexibility and usability in a corpus analysis tool', International Journal of Corpus Linguistics 17 (3), pp. 380–409.

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

Horne, P. 2013. 'Style and the making of character in Dickens', in D. Tyler (ed.) Dickens's Style. Cambridge: Cambridge University Press.

Howarth, D. 2000. Discourse. Philadelphia: Open University Press.

Iacob, I.E., A. Dekhtyar and W. Zhao. 2004. Xpath Extension for Querying Concurrent XML Markup, pp. 1–15. Department of Computer Science, University of Kentucky. See: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.116.4616&rep=rep1&type=pdf.

John, J. 2001. Dickens's Villains: Melodrama, Character, Popular Culture. Oxford: Oxford University Press.

Johns, T.F. 1990. 'Should you be persuaded', in T.F. Johns and P. King (eds) Concordancing in the Language Classroom. (English Language Research Journal, volume 3.) Birmingham: University of Birmingham.

Krestel, R. 2006. Automatic Analysis and Reasoning on Reported Speech in Newspaper Articles. PhD Thesis. Universität Karlsruhe (Fakultät für Informatik) and Concordia University (Department of Computer Science). Available online at: http://www.semanticsoftware.info/system/files/believer.pdf.

Krestel, R., S. Bergler and R. Witte. 2008. 'Minding the source: automatic tagging of reported speech in newspaper articles', Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC 2008). 28–30 May. Marrakech, Morocco.

Lambert, M. 1981. Dickens and the Suspended Quotation. New Haven and London: Yale University Press.

Leech, G. 2000. 'Grammars of spoken English: new outcomes of corpus-oriented research', Language Learning 50 (4), pp. 675–724.

Leech, G. and M. Short. 2007 [1981]. Style in Fiction: A Linguistic Introduction to English Fictional Prose. Harlow: Pearson Education.

Leverage, P., H. Mancing, R. Schweickert and J.M. William (eds). 2011. Theory of Mind and Literature. West Lafayette: Purdue University Press.

Marinelli, P., F. Vitali and S. Zacchiroli. 2008. 'Towards the unification of formats for overlapping markup', New Review of Hypermedia and Multimedia 14 (1), pp. 57–94. Accessed May 2016 at: http://doi.org/10.1080/13614560802316145.

Mahlberg, M. 2005. English General Nouns: A Corpus Theoretical Approach. Amsterdam: John Benjamins.

Mahlberg, M. 2009. 'Local textual functions of move in newspaper story patterns' in U. Römer and R. Schulze (eds) Exploring the Lexis–Grammar Interface, pp. 265–87. Amsterdam: John Benjamins.

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

Mahlberg, M. and M.B. O'Donnell. 2008. 'A fresh view of the structure of hard news stories' in S. Neumann and E. Steiner (eds) Online Proceedings of the 19th European Systemic Functional Linguistics Conference and Workshop, Saarbrücken. 23–25 July 2007. Accessed May 2016 at: http://scidok.sulb.uni-saarland.de/volltexte/2008/1700/.

Mahlberg, M. and P. Stockwell. 2016. 'Point and CLiC: teaching literature with corpus stylistic tools' in M. Burke *et al.* (eds) Scientific Approaches to Literature in Learning Environments, pp. 251–67. Amsterdam: John Benjamins.

Mahlberg, M. and C. Smith. 2010. 'Corpus approaches to prose fiction: Civility and body language in Pride and Prejudice' in D. McIntyre, and B. Busse (eds) Language and Style, pp. 449–67. Basingstoke: Palgrave Macmillan.

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

Mahlberg, M., K. Conklin and M.-J. Bisson. 2014. 'Reading Dickens's characters: textual patterns and their cognitive reality', Language and Literature 23 (4), pp. 369–88.

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

Manning, C.D. and H. Schuetze. 1999. Foundations of Statistical Natural Language Processing. Cambridge: MIT Press.

McCarthy, M. and R. Carter. 1994. Language as Discourse: Perspectives for Language Teaching. Harlow: Longman.

McIntyre, D. and B. Walker. 2011. 'Discourse presentation in Early Modern English writing: a preliminary corpus-based investigation', International Journal of Corpus Linguistics 16 (1), pp. 101–30.

Moretti, F. 2013. Distant Reading. London: Verso.

Newsom, R. 2000. 'Style of Dickens' in P. Schlicke (ed.) The Oxford Reader's Companion to Dickens, pp. 553–7. Oxford: Oxford University Press.

O'Donnell, M.B. 2008. 'KWICgrouper: designing a tool for corpus-driven concordance analysis' in Software-aided Analysis of Language: Special Issue of International Journal of English Studies 8 (1), pp. 107–22.

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

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

Premack, D.G. and G. Woodruff. 1978. 'Does the chimpanzee have a theory of mind?', Behavioral and Brain Sciences 1 (4), pp. 515–26.

Renouf, A., A. Kehoe and J. Banerjee. 2005. 'The WebCorp search engine: a holistic approach to web text search', in Electronic Proceedings of CL2005. University of Birmingham.

Römer, U. 2010. 'Establishing the phraseological profile of a text type: the construction of meaning in academic book reviews', English Text Construction 3 (1), pp. 95–119.

Richards, I.A. 1929. Practical Criticism. London: Kegan Paul, Trench, Trubner.

Rouse, R.H. and M.A. Rouse. 1991. 'The development of research tools in the thirteenth century', in Authentic Witness: Approaches to Medieval Texts and Manuscripts, pp. 221–55. Notre Dame: University of Notre Dame Press.

Scott, M. 2016. WordSmith Tools (Version 7). Stroud: Lexical Analysis Software.

Scott, M. and C. Tribble. 2006. Textual Patterns: Key Words and Corpus Analysis in Language Education. Amsterdam: John Benjamins.

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

Skelton, J.R., A.M. Wearn and F.D.R. Hobbs. 2002. '"I" and "we": a concordancing analysis of how doctors and patients use first person pronouns in primary care consultations', Family Practice 19 (5), pp. 484–8.

Sinclair, J. (ed.). 1987. Looking Up. An Account of the COBUILD Project in Lexical Computing. London: HarperCollins.

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

Sinclair, J. 2003. Reading Concordances: An Introduction. London: Pearson.

Sinclair, J. 2004. Trust the Text: Language, Corpus and Discourse. London: Routledge.

Stockwell, P. 2009. Texture: A Cognitive Aesthetics of Reading. Edinburgh: Edinburgh University Press.

Stockwell, P. 2013. 'The positioned reader', Language and Literature 22 (3), pp. 263–77.

Stockwell, P. and M. Mahlberg. 2015. 'Mind-modelling with corpus stylistics in David Copperfield', Language and Literature 24 (2), pp. 129–47.

Tognini-Bonelli, E. 2001. Corpus Linguistics at Work. Amsterdam: John Benjamins.

Turner, M. 1992. Reading Minds: The Study of English in the Age of Cognitive Science. Princeton: Princeton University Press.

Vermeule, B. 2010. Why Do We Care About Literary Characters? Baltimore: Johns Hopkins University Press.

Walker, B. 2010. 'Wmatrix, key concepts and the narrators in Julian Barnes's Talking it Over' in D. McIntyre and B. Busse (eds) Language and Style, pp. 364–87. Basingstoke: Palgrave Macmillan.

Zunshine, L. 2006. Why We Read Fiction: Theory of Mind and the Novel. Columbus: Ohio State University Press.

Appendix A

Data associated with this paper that can be downloaded from clic.bham.ac.uk:

- CLiC web app code;
- DNov corpus (xml version);
- 19C corpus (xml version);
- Annotation module (used to create DNov and 19C); and,
- Gold standard.