Digital Scholarship in the Humanities Advance Access published February 3, 2016

Using Models of Lexical Style to Quantify Free Indirect Discourse in Modernist Fiction

Julian Brooke
Department of Computing and Information Systems, University of Melbourne, Australia

Adam Hammond
Department of English and Comparative Literature, San Diego State University, USA

Graeme Hirst
Department of Computer Science, University of Toronto, Canada

Correspondence:
Julian Brooke, Department of Computing and Information Systems, University of Melbourne, Melbourne, Victoria 3010, Australia.
E-mail:
jabrooke@unimelb.edu.au

Abstract

Modernist authors such as Virginia Woolf and James Joyce greatly expanded the use of 'free indirect discourse', a form of third-person narration that is strongly influenced by the language of a viewpoint character. Unlike traditional approaches to analyzing characterization using common words, such as those based on Burrows (1987), the nature of free indirect discourse and the sparseness of our data require that we understand the stylistic connotations of rarer words and expressions which cannot be gleaned directly from our target texts. To this end, we apply methods introduced in our recent work to derive information with regards to six stylistic aspects from a large corpus of texts from Project Gutenberg. We thus build high-coverage, finely grained lexicons that include common multiword collocations. Using this information along with student annotations of two modernist texts, Woolf's *To The Lighthouse* and Joyce's *The Dead*, we confirm that free indirect discourse does, at a stylistic level, reflect a mixture of narration and direct speech, and we investigate the extent to which social attributes of the various characters (in particular age, class, and gender) are reflected in their lexical stylistic profile.

1 Introduction

In fiction, a reader's understanding of individual characters is typically gleaned either from narrative description of these characters or from the words and thoughts of the characters as presented in direct discourse. A third option, which has been present for centuries but only reached its full potential with the modernist work of the early twentieth century, is to mix these traditional forms into a combination known as free indirect discourse (FID), which has properties of both narration and speech. The work presented here aims to introduce quantitative evidence into modernist debates about FID, which have been conducted in a less-than-rigorous manner, propelled by intuition and driven entirely

doi:10.1093/llc/fqv072

by qualitative impressions. Rather than beginning from anecdotal or a priori definitions of the device, our approach works inductively, drawing general conclusions from annotated passages in two key modernist works, *To The Lighthouse* by Virginia Woolf (1927) and 'The Dead' by James Joyce (1914), and a model of lexical stylistic variation derived by applying state-of-the-art computational techniques in a corpus of Project Gutenberg texts that represents a wide range of literary expression. Some of the questions to which we seek quantitative insight include the following: Is our computational, yet human-interpretable, model of style useful for literary analysis? Is FID stylistically distinguishable from direct discourse and narration? If so, does it occupy a middle position between the two, between the stylistic extremes of the narrator's language and that of individual characters? Is it possible to distinguish individual characters' styles from one another in a novel like *To the Lighthouse*, where the vast majority of character speech comes in the form of FID and is thus influenced by the language of the narrator? If so, in what ways do the voices within a text differ from one another, and how does the variety of voices within a text reflect its themes and the background of the author who produced it?

2 Free Indirect Discourse

Our work proceeds from an interest in voice[1] in fictional narratives. As Abbott (2008) argues, voice is one facet of what is commonly referred to as point of view, the other facet being focalization. Whereas focalization provides orientation in a story by giving the reader access to a character's personal experience, letting the reader see the world through the character's eyes, voice offers further insight by allowing us to hear a character's distinctive manner of expression. Our interest here lies in a particular technique for introducing character voice, FID (also known as free indirect style, among many other names), which consists of third-person narration into which personalized aspects of a character's subjective expression are introduced without being offset by quotation marks. Consider the following example from Jane Austen's *Pride and Prejudice*:

> [Elizabeth's] astonishment, as she reflected on what had passed, was increased by every review of it. That she should receive an offer of marriage from Mr. Darcy! that he should have been in love with her for so many months!

While the use of the past tense and the third person identify the passage as a span of narration, the exclamation points and the breathless, uncapitalized transition between clauses—neither of which is placed in quotation marks—clearly derive from Elizabeth's subjective consciousness, and not from the perspective of the narrator. A common way of conceptualizing FID (what we will call the 'classical' definition of FID) is as a mixture of direct and indirect discourse. In direct discourse, the narrator of a fictional work reproduces verbatim the actual words that a character speaks or thinks (their unfiltered voice), usually placing these words in quotation marks. For instance, if Austen had employed direct discourse in the above scene, she might have rendered it as follows: 'Elizabeth was astonished. Her thoughts ran wild: ''An offer of marriage from Mr. Darcy! Him, of all people, in love with me all these months!''' In indirect discourse, by contrast, the narrator summarizes the character's words rather than reproducing them verbatim: 'Elizabeth was astonished at Mr. Darcy's offer of marriage. She was particularly shocked that he had been in love with her for several months, despite his seeming indifference.' Austen's passage of FID, quoted above, conforms to this 'classical' definition by combining elements of direct discourse (the exclamation points and quick movement from sentence to sentence) with elements of indirect discourse (the past tense, the third person, and summaries rather than faithful renderings of Elizabeth's thoughts).

FID is today a universally recognized element of narrative theory. Yet, as the curious history of FID demonstrates, it has taken centuries for it to gain this widespread acceptance. Direct and indirect discourse were employed in the first Western novels, such as *Don Quixote*, published around 1600; it was not until some 200 years later, however, that FID

began to appear in fiction, Goethe and Jane Austen being among its earliest adopters (Pascal, 1977). Following the invention of FID, moreover, it took another century for critics to notice it. It is generally agreed that the German critic Adolph Tobler was the first to identify and describe FID, referring to it in an 1892 article as a 'peculiar mixture of direct and indirect speech' (qtd. in Pechey, 2007), and thus initiating the 'classical' definition of FID. In the years that followed Tobler's discovery, FID became the focus of sustained critical interest. The intensity of the discussion, as well as its uncertainty, is evidenced in the many names that 20th-century critics coined for FID: to name only a few, Theodor Kalepky called it 'veiled speech' (1912), Charles Bally 'free indirect style' (1912), Etienne Lorck 'experienced speech' (1921), Leo Spitzer 'pseudo-objective speech' (1921), Otto Jesperson 'represented speech' (1924), Edouard Dujardin 'indirect interior monologue' (1931), and Mikhail Bakhtin 'pseudo-objective discourse' (1920s–30s) (Pechey, 2007). Working in the midst of this critical tumult, writers of the modernist period (roughly 1880–1950) employed FID more self-consciously than predecessors such as Austen, and consequently began to employ the device in distinctive manners. For our purposes, two major differences exist between modernist employments of FID and what came before: first, modernists exploited the device's potential for vocal uncertainty; second, they pushed beyond the 'classical' definition of FID by mixing the thoughts of their characters together with narration distinct from indirect discourse. As to the first point, Sotirova (2013) argues that 19th-century writers such as Austen used FID to alternate between the perspectives of the narrator and the character, yet the transitions between a character's words and a narrator's account generally remained smooth and grammatically coherent. One has little difficulty in the passage of FID from *Pride and Prejudice* (Austen, 1813), for instance, in identifying where Elizabeth's subjective impressions end and the narrator's objective account recommences. Modernist writers, on the other hand, turned to FID precisely in order to blur the lines between narrator and character (Sotirova, 2013, Rundquist, 2014). Both texts under consideration here fall into the self-conscious modernist phase of FID employment; although there are many differences between them—'The Dead' is much shorter than *To the Lighthouse*, uses FID much less frequently, and applies it to far fewer characters—both exploit the potential for uncertainty and incoherence in FID. This is evident from the opening lines of the two works. 'The Dead' begins as follows:

> Lily, the caretaker's daughter, was literally run off her feet. Hardly had she brought one gentleman into the little pantry behind the office on the ground floor and helped him off with his overcoat than the wheezy hall-door bell clanged again and she had to scamper along the bare hallway to let in another guest.

Though the first sentence may look at first glance like objective narration, the word 'literally' troubles this reading, because Lily is not knocked literally off her feet in this scene, but only figuratively so (Kenner, 1978, pp. 15–16). Is the first sentence thus FID, mixing the lower-class Lily's subjective idiom with the more correct third-person language of the narrator? Or is it the narrator himself who employs this idiom? Joyce leaves this ambiguity unresolved. Similar uncertainty pervades the opening sentences of *To the Lighthouse*:

> 'Yes, of course, if it's fine tomorrow', said Mrs Ramsay. 'But you'll have to be up with the lark', she added. [¶] To her son these words conveyed an extraordinary joy, as if it were settled, the expedition were bound to take place, and the wonder to which he had looked forward, for years and years it seemed, was, after a night's darkness and a day's sail, within touch.

In this passage, we are presented with two spans of narration ('said Mrs Ramsay' and 'she added') and two passages of direct discourse in which the narrator introduces the verbatim words of Mrs Ramsay ('Yes, of course, if it's fine tomorrow' and 'But you'll have to be up with the lark'). The rest of the passage is presented in FID, mixing together the voices of the narrator, Mrs Ramsay, and her son James: while the use of third-person pronouns and the past tense clearly indicate the voice of the

narrator, phrases such as 'for years and years it seemed' clearly present a subjective perspective. Typically of modernist employments of FID, however, it is often difficult or impossible to point to a single word in the passage and determine its speaker definitively—and while we know that the characters' thoughts are not given verbatim, there is no means of identifying unequivocally which words are quoted and which are not. For instance, does James himself use the word 'extraordinary' to describe his joy at the prospect of a visit to the lighthouse; or does Mrs Ramsay (or the narrator) use the word to interpret James's feelings; or might the word come from the narrator's interpretation of Mrs Ramsay's interpretation of her son's feeling? Like Joyce, Woolf uses FID to raise these ambiguities of voice, but leaves them unresolved.

The second distinct feature of modernist FID is the broadening of the device beyond its 'classical' definition. As Sotirova (2013) argues, 19th-century FID was generally limited to mixing together the actual words of a character (direct discourse) with the narrator's summary of those words (indirect discourse). Modernist writers expanded the device by mixing the subjective experience of the characters (including but not limited to plausible verbalizations of their thoughts) into (a) passages of narration clearly distinct from indirect discourse (for instance, passages in which the narrator is describing an element of the physical world rather than a character's thoughts) and (b) passages in which the narrator relates the non-verbal emotions of a character (which do not qualify as indirect discourse, since indirect discourse is by definition a summary of verbalized speech or thought). Both the texts under consideration here employ FID in this broadened sense. The opening sentence of Chapter 11 in Part I of *To the Lighthouse* demonstrates type (a):

> No, she thought, putting together some of the pictures he had cut out—a refrigerator, a mowing machine, a gentleman in evening dress—children never forget.

This sentence contains clear examples of direct discourse ('No' and 'children never forget') and narration ('she thought, putting together some of the pictures he had cut out'); less clear is the exact source of the phrase 'a refrigerator, a mowing machine, a gentleman in evening dress', which, as Rundquist (2014, pp. 165–6) explains, describes the physical world (rather than what Mrs Ramsay is thinking or saying, as in indirect discourse), but does so in a style that imitates aspects of the character's consciousness—the visual input, a sequence of pictures, which is intruding on her verbalized thoughts. Rundquist argues that this passage, by mixing together narration with a representation of subjective experience, is indeed a form of FID, but a form of 'non-classical' FID that he names 'represented perception'. An employment of type (b) occurs in the passage from which Joyce takes the name of 'The Dead'. Following his wife Gretta's revelation of a passionate youthful relationship with a young man dying of tuberculosis, Gabriel, the protagonist, ponders the influence of the dead on the living:

> His soul had approached that region where dwell the vast hosts of the dead. He was conscious of, but could not apprehend, their wayward and flickering existence. His own identity was fading out into a grey impalpable world: the solid world itself, which these dead had one time reared and lived in, was dissolving and dwindling.

While this passage mixes together third-person narration with elements of Gabriel's subjective experience, it does not contain any indirect discourse, since it does not summarize any of Gabriel's conscious, verbal thoughts (Gabriel could not plausibly have thought, in the second sentence, 'I am conscious of, but I cannot apprehend, their wayward and flickering existence'). For Rundquist, this passage qualifies as yet another type of 'non-classical' FID, 'consonant psycho-narration'. Given the expanded modernist usage of FID, we proceed in our article with a broadened definition of FID as third-person narration into which personalized aspects of a character's subjective expression are introduced without being offset by quotation marks.

3 Annotation

Our annotation methodology for tagging types of discourse in *To The Lighthouse* and 'The Dead'

falls somewhere between traditional annotation strategies involving a small number of expert annotators and modern crowdsourcing techniques. Our annotators were three cohorts of roughly 160 students each, mostly English literature majors, who were enrolled in a class on digital forms of literature; as such, they were experienced in analysis of literature, but they were not experts on the particular texts. The annotation took the form of a marked assignment, with each student assigned a short passage for interpretation. Because multiple valid interpretations of a given passage are often possible, each was assigned to several students (generally four to five). This allowed us to track subjective differences in interpretation, as well as to identify passages on which there was interpretive consensus. The latter of these goals is more relevant to the present work: for the analysis presented here, a span of text is considered to have a certain tag (for instance, to be FID) if a majority of annotators tagged it as such, a standard approach that can result in highly reliable gold standards even when the quality of annotators or the difficultly of task (as in our case) results in only moderate annotator agreement (Beigman Klebanov and Beigman 2009).[2]

Our annotation guidelines instructed the students to tag the text using TEI (Text Encoding Initiative) XML markup, in particular the TEI 'said' tag, which students used to identify and interpret every instance of character discourse. Since both FID and direct discourse can manifest both below and above the sentence level, we allowed the spans to be as short as individual words and as long as a paragraph. For annotation of discourse type, we modified the TEI guidelines to allow for annotations of FID, in addition to direct and indirect discourse, which are covered in the TEI guidelines.[3] We also had the students annotate whether a particular span of character discourse was spoken aloud or thought silently, using the 'aloud' attribute, and to identify the character whose discourse was being introduced, using the 'who' attribute. Narration was left untagged. As part of their assignment, the students were asked to justify their choices. Between the first and second round of annotations of *To the Lighthouse*, we made some small changes to our guidelines: instead of using the TEI 'direct' attribute as we had with the first round, for clarity we switched to a 'discourse' attribute (with the same options, expressed as 'direct' for direct speech and 'free' for FID), added a 'group' value to account for multiple characters (in the first iteration students had simply tagged multiple characters), and allowed for embedded tags for instances of character discourse within spans of character discourse, though this option was rarely used. None of these changes greatly affects our analysis here, and, otherwise, the guidelines were consistent across the three rounds of annotation: in each case, the exact requirements were explained in class, reviewed in a tutorial session, and also provided in the form of written annotation guidelines. An example of a student annotation is given below:

> <said who="#lily" discourse="free" aloud="false">She looked at the steps; they were empty; she looked at her canvas; it was blurred. With a sudden intensity, as if she saw it clear for a second, she drew a line there, in the centre</said>. It was done; it was finished. <said who="#lily" discourse="direct" aloud="false">Yes</said>, she thought, laying down her brush in extreme fatigue, <said who="#lily" discourse="direct" aloud="false">I have had my vision</said>.

The first two rounds of annotation dealt with two separate sections of *To The Lighthouse*: the first four chapters and the last seven chapters. The text as a whole was too large to annotate by our method, and choosing the chapters at opposite ends of the novel allowed us a reasonably wide range of character representation in the FID, given that there is a significant shift in the viewpoint from older to younger characters as the novel progresses. For 'The Dead', our third round of annotation, we included the entire short story: since there is much less FID in the short story, the task was less difficult, and we were thus able to assign longer passages. The annotations of *To The Lighthouse* can be browsed on our Web site for the 'Brown Stocking' project,[4] and annotations of 'The Dead' can be viewed on 'The (Living) Dead' project Web site.[5]

Before we move on to our methodology for our analysis, we note that our annotation of these texts

could be applied as training or testing data to a number of computational linguistics tasks in the domain of literature, for instance tracking subjective viewpoint in narrative (Wiebe 1994), distinguishing between speech, thought, and narrative (Brunner 2013), classifying the source of direct speech (He *et al.*, 2013), classifying character gender (Hota *et al.*, 2006), and others. For various reasons, however, the modernist texts in question would serve as particularly difficult and rather non-typical data for these tasks, and we should also point out that none of these tasks are our immediate interest here, as pursuing full-fledged models of these various phenomena would require detours into areas beyond the scope of this article, e.g. the discourse structure of the text.

4 Method

The primary technical work in our analysis is the creation of high-coverage stylistic lexicons and their use in analyzing stylistic differences across kinds of discourse and across the FID of various characters in our target texts. The six stylistic aspects[6] we focus on here are as follows, with the definitions adapted from earlier work (Brooke and Hirst, 2013b):

Objective: Words that are emotionally distant, projecting a sense of disinterested authority. Examples include 'invariable', 'finalize', and 'ancillary'.

Abstract: Words that refer to something that requires major psychological or cultural knowledge to grasp, which cannot purely be defined in physical terms. Examples include 'sophism', 'alienation', and 'implicit'.

Literary: Words that one would expect to see more or less exclusively in literature; these words often feel old-fashioned or 'flowery'. Examples include 'yonder', 'revelry', and 'wanton'.

Colloquial: Words that are used primarily in informal contexts, such as slang words used among friends. Examples include 'booze', 'dodgy', and 'damn'.

Concrete: Words that primarily refer to events, objects, or properties of objects in the physical world that one is able to see, hear, smell, or touch. Examples include 'radish', 'sew', and 'freeze'.

Subjective: Words that are strongly emotional or reflect a personal opinion. Examples include 'ugly', 'worthy', and 'bastard'.

A few important notes about these stylistic aspects: First, all of them are expressed primarily through lexical choice, though in some cases there may be non-lexical features that may also play a role in the phenomenon (for instance archaic syntactic structures may also give texts a literary feel). Second, though we may speak informally about a word 'having' one style or another, these aspects are not strictly binary, since it is possible to speak of one word being more X than another (e.g. both 'pristine' and 'godawful' have a subjective quality, yet the latter is more clearly more intense, and thus could be considered more subjective), and since there are borderline cases (is 'pristine' literary?). Third, the very general linguistic distinctions captured by these aspects are represented in more lexical items than it is reasonable to annotate manually, particularly if multiword phenomena are taken into account. Finally, individual words may have significant weighting in several of these aspects, and there are in fact strong positive and negative correlations with respect to which aspects tend to appear together, or not, in the same lexical item. The strongest correlative effect—one that also interferes with accurate automatic lexicon creation—can be attributed to communicative purposes of language and social status. Leckie-Tarry (1995), in her theory of register variation, posits a main cline of register, with aspects of language associated with spoken, situated language on one end (e.g. colloquial, concrete, and subjective) and written, culturally influenced language on the other (e.g. objective, abstract, literary), with positive statistical correlations among aspects on the same pole, and negative correlations among aspects on opposing poles. We note here that use of words on the written end of the cline will generally reflect increased social power.

Some of these stylistic aspects have been addressed to some degree or another (though often with different names) in existing (manual) lexical resources, for instance the Urban Dictionary,[7] the MRC psycholinguistic database (Coltheart, 1980),

the General Inquirer lexicon (Stone *et al.*, 1966), and the LIWC word lists (Pennebaker *et al.*, 2001). They also correspond roughly to some of the poles of the dimensions in Biber's (1989) multidimensional analysis of text genre; indeed we have used automatically created lexicons with these styles to distinguish genres in the British National Corpus (Brooke and Hirst, 2013a). They also clearly play a role in sociolinguistic variation (Tagliamonte, 2011). Though informed by more descriptive approaches to style, our original choices were based primarily on an analysis of prescriptive writing manuals (e.g. Strunk and White, 1979) and the aspects of style that are commonly addressed in those contexts; we prefer this breakdown of the style space in part because it better corresponds to everyday intuition than does the jargon of specific academic disciplines, and is thus more interpretable to those from a variety of backgrounds. Our specific choice of terminology is not intended to be universal (in other work, for instance, we have used 'formal' and 'informal' roughly to mean what we here call 'objective' and 'colloquial'), nor are these six styles intended to be exhaustive. Yet, taken as a whole, our six stylistic aspects capture a significant range of the major lexical stylistic variation found in English as identified by writing experts as well as descriptive linguists, and our lexicon annotation project (Brooke and Hirst, 2013b) has confirmed that people have an intuitive sense of these distinctions.

The methods used to build these lexicons are documented in detail elsewhere (Brooke and Hirst, 2013a, b, 2014; Brooke *et al.*, 2014), and so here we will keep the description brief. To begin, we need a relatively small set of words to serve as initial examples of these stylistic aspects. The words used in this work are a modified set of the 900 words whose stylistic annotation is described by Brooke and Hirst (2013b). The only modification necessary was to remove a set of colloquial words that are too modern to appear in our training and target texts, particularly Internet acronyms (e.g. 'lol'). In some cases, we were able to replace these words with near equivalents that would have been in use 100 years ago, for instance 'dude' with 'chap', and 'screw-up' with 'muddle', but our lexicon still lost seventy-one entries, or roughly half our set of colloquial words. The annotated lexicon used here is therefore 829 words.

The other important input to our model is a corpus in which the co-occurrence of words and expressions reflects the stylistic variation in which we are interested. In previous work, we used a social media corpus, which is not suitable to our literary interests. Here we choose instead to work from the collection of out-of-copyright texts in Project Gutenberg.[8] In the present work, we used all the English texts in the 2010 image of the Web site, approximately 24,000 texts of varying genres and styles. Although this may appear to be a large number of texts, it is in fact quite small relative to the social media corpus used in our previous work (2.5 million blogs)—though in raw word count the Gutenberg corpus is larger. We were also concerned that the stylistic diversity of literary texts—which, unlike much non-fiction, is made up of a great number of speaking voices—would hamper identification of stylistic aspects. We therefore increased the number of texts and the degree of variation by identifying novels and plays and including the speech of individual characters—when it could be identified in the nearby surrounding context and when there was sufficient quantities of it—as a 'text' separate from narration or stage directions, etc. When a specific character could not be identified, we also distinguished speech according to the pronoun used (i.e. 'he' or 'she'), with all other speech (e.g. text in quotation marks but with no surrounding attribution) included in an 'other' category separate from the narrator. We also removed the Project Gutenberg information from the beginning and the end of the texts, and removed titles, illustrations, tables of contents, and prefaces. All of this was accomplished using heuristic, rule-based methods; our expectation was not to perform these tasks perfectly, but rather to accomplish the basic goal of increasing variation across 'texts' in the corpus. The ability to divide literary texts in this manner is a feature of our GutenTag tool for digital humanities research in the Project Gutenberg corpus (Brooke *et al.*, 2015).[9]

A third important difference between this approach and our previous approach to lexicon

acquisition is the inclusion of multiword units in the analysis. We applied the method introduced by Brooke *et al.* (2014) to segment the Gutenberg corpus into (potentially) multiword segments based on *n*-gram corpus statistics, which were shown in that work to correspond reasonably well to known multiword expressions (in the Gutenberg corpus as well as other corpora). The algorithm itself is quite complex, including an initial round of segmentation to resolve high-frequency *n*-gram overlaps, another step that combines the segments as a vocabulary for further splitting, and a final round of segmentation that applies these splits to the initial segmentation. Yet, the basic idea of the approach is simple: it seeks to preserve segments which show high internal predictability (i.e. conditional probability) based on the *n*-gram statistics collected across the whole corpus. The same statistics were used to segment the texts under analysis, and these segments were used as the base lexical units for building our lexicon, rather than individual words. Examples of phrases that were found in our target texts include 'tell you the truth', 'looked as if', 'nothing to do but', 'absorbed in', 'point of view', 'shame that', 'fat of the land', 'such people', 'heart would break', 'found himself', 'keeping house', 'were bound to', 'hung round with', 'cast a gloom over', 'week or so', 'passed away', 'years and years', 'gone for ever', 'young lady', 'could not bear', 'lighted his pipe', 'starting off', 'came to a stop', and 'anything at all'. It is clear that many of these phrases have stylistic connotations that are not available from the individual words (or, at the very least, they offer word sense disambiguation), so using phrases rather than individual words should be preferable for stylistic analysis (provided the phrases appear reasonably often in the corpus, which is guaranteed by the methodology of Brooke *et al.*, 2014).

The stylistic lexicons are created by a combination of two methods introduced elsewhere. First, each stylistic aspect is addressed independently using the continuous lexical spectrum method of Brooke and Hirst (2014), which was shown to be superior to other corpus-based techniques for building lexicons of this kind. This approach is supervised. First, feature vectors for each word or expression are derived, with each element of the vector corresponding to the number of times the target word or expression co-occurs with one of a large set of profile words (words of moderate frequency in the corpus, in this case appearing not more than once per 10 documents but no less than once per 100), normalized so that the vector sums to 1. For each stylistic dimension, we use an off-the-shelf machine learning algorithm, SVM Rank (Joachims, 2002), to derive a weight vector which allows us to predict style scores for unseen words or expressions. Next, we use the initial scores as the input to the method of Brooke and Hirst (2013b), which improves the scores of the individual styles by considering all the styles together in a single, graphical model—with edge weights corresponding to the cosine distance in the six-dimensional stylistic space created by the initial values—and updating them using a simple, one-step label propagation function. The main benefit of this approach lies in distinguishing styles that are otherwise strongly influenced by one another (due to being on the same pole of the cline of register, etc.). We evaluated the quality of the resulting lexicon by the same method as Brooke and Hirst (2013b), i.e. using the pairwise accuracy of our annotated words, using five-fold cross-validation to derive the scores. Pairwise accuracy is calculated by exhaustively pairing off all words with opposing style annotations (i.e. for each pair, there is one word that has been judged to have the style, and one that has not), and then counting the percentage whose automatic style scores have the correct relative orientation (i.e. the word judged to have the style does indeed have a higher score for that style). The result is shown in Table 1. Although the two evaluations are not identical due to the changes in the lexicon, we also include the best (average) results from Brooke and Hirst (2013b) for comparison. Performance is strikingly similar, despite the various differences (in particular, the use of the Project Gutenberg corpus instead of blogs), and is reasonably high, all above 90% except for subjectivity, which is more difficult to identify automatically for reasons discussed in Brooke and Hirst (2013b).

Having confirmed that this method of lexical style acquisition also works for Project Gutenberg

texts, we derived raw style scores for every word or expression found in our annotations, excluding only character names. The resulting lexicon was then normalized in the usual way, namely by subtracting from each style score the mean of all the style scores, and then dividing by the standard deviation. To assign style scores beyond the word, we average the (normalized) style scores for all types appearing within spans with a particular tag (i.e. for all those passages with the same 'discourse' or 'direct' attribute, or, for characters, with the same 'who' character attribute). We will refer to these six numbers together as a 'stylistic profile'. To give a hypothetical example, suppose our span is the sentence 'The wanton destruction of the tome was the devil's work' and our normalized stylistic lexicon contains the entries the = [0.1, 0, 0, 0, 0, 0], wanton destruction = [0.1, −0.1, 0.2, −0.1, 0.2, 0.4], of = [0, 0, 0.1, 0, 0, 0], tome = [0.2, 0, 0.3, −0.2, 0, −0.1], was = [0, 0, 0, 0, 0, 0], and the devil's work = [−0.2, 0.1, 0.3, 0.1, 0.1, 0.5]. We can calculate a stylistic profile for this span, [0.02, 0, 0.13, −0.3, 0.05, 0.13], by summing these vectors and dividing by 6, the number of types in the passage (the two occurrences of the outside a multiword phrase only count once). We use types rather than tokens so that the style scores of function words and other very common words do not unduly influence the results[10]; in this regard, our analysis is more similar to that of Ramsay and Steger (2006), who analyze Woolf's *The Waves* by looking at hapax legomena, than it is to more traditional approaches, which are driven by function words, e.g. Burrows (1987). The reason we can look at rare words and expressions is, of course, that we are deriving knowledge about the stylistic use of these words from a much larger corpus which has sufficient instances of their use, whereas traditional analysis (including Burrows's approach) is carried out only within the context of the individual text or texts under investigation. Finally, to improve the readability of our results, we carried out a second normalization on the tag-level stylistic profiles by making narration in *To The Lighthouse* the zero-point of our stylistic space, shifting all the results by subtracting its stylistic profile from each. In the case of the FID in *To The Lighthouse*, for instance, the initial stylistic profile was [−0.10, 0.01, −0.06, −0.13, −0.06, −0.21], so to calculate the profile presented in Table 1, we subtract the initial stylistic profile for narration, [−0.18, −0.16, −0.08, −0.11, 0.09, −0.22], which results in a more interpretable profile, [0.08, 0.17, 0.02, −0.02, −0.15, 0.02].

5 Results

Table 2 shows a comparison between the stylistic profiles of different types of discourse. With respect to the classic dichotomy between narration and direct discourse, we see the basic patterns we would expect: narration mostly describes action in a dispassionate manner; therefore, its values on the concrete and objective dimensions are relatively high, and those on the colloquial and subjective dimensions are consistently low. This simply indicates that our lexicons are working as they should. With respect to the differences between speech and thought, however, the results are more equivocal: although in *To the Lighthouse* thought is more abstract and less colloquial, as we might expect, the opposite pattern is true in 'The Dead', where the low type counts for thought might be influencing our results, creating extreme values in some cases.

With respect to our main interest, the status of FID, in *To the Lighthouse* we found that in four of

Table 1. Comparison of stylistic lexicon acquisition performance (pairwise accuracy) between current and previous work

Corpus used	Pairwise accuracy by style					
	Objective	Abstract	Literary	Colloquial	Concrete	Subjective
Project Gutenberg (current article)	98.0	93.7	94.1	97.5	95.1	85.2
Blog (Brooke and Hirst, 2013b)	97.2	94.0	92.7	97.7	94.9	86.4

the six styles, FID occupies a middle position: it is more abstract than narration, but less abstract than directly rendered thought; more literary than narration but less so than direct speech or thought; less concrete than narration but more so than direct speech or thought; and so on. There are two exceptions, colloquial and objective, where FID is in the extreme position: for colloquial, FID nonetheless tracks very closely with narration, and in fact both exceptions may perhaps merely reflect a single peculiarity of Woolf's narrator: that, when she mixes her language with that of her characters, she tends not to admit their colloquialisms, but instead to elevate their language to a (relatively) higher register. In 'The Dead', FID occupies a middle position in all six styles, with the clearest cases being colloquial and subjective: while Joyce's characters are very colloquial and quite subjective in their direct discourse, and his narrator is relatively neutral in both, FID falls consistently in between.

Although our results show that FID functions similarly in the two texts with respect to style, they nonetheless capture some of the peculiarities of the texts relative to one another. For example, while the values for the narrators of the two works are close to one another—and quite neutral, as we would expect from works with 'reliable' extradiegetic third-person narrators who relate the story but are not present in it—the slight differences are telling. Where Woolf's narrator is extremely flat, detached, and objective, Joyce's narrator has a clear personality, which is reflected in his higher values on the literary, colloquial, and subjective dimensions.

Although FID is less 'extreme' stylistically than direct discourse, it nonetheless clearly manifests individual characters' particular styles and personalities. Table 3 contains the style scores for the FID of various characters in the two texts. Though FID is employed relatively sparingly in 'The Dead', and nearly all of it is given to a single character, Gabriel's personal linguistic manner comes across clearly in his FID: compared to all other characters, Gabriel—reserved, given to pondering big ideas, prone to quoting literature—is notably less colloquial, more literary, and more abstract in his FID. In *To the Lighthouse*, where FID is by far the most prevalent means for introducing character discourse, individual personalities likewise shine through in FID, despite the fact that characters' individualized expression is mixed together with the language of the narrator. Compared with his wife, Mr Ramsay, a professor of philosophy and himself fond of literary quotations, is much more abstract, much more literary, and far less concrete in his FID. Despite their dissimilarities, Mr and Mrs Ramsay nevertheless have more in common with one another stylistically than they do with their children: Cam and James are notably less objective, less abstract, and more colloquial. Stylistic profiles can also reveal relationships between characters. Lily Briscoe and Mrs Ramsay (the two female protagonists, the latter much influenced by the former) have remarkably similar stylistic profiles, while Charles

Table 2. Stylistic profiles for various types of discourse in *To the Lighthouse* and 'The Dead'. 'Narration' refers to passages identified as the narrator's words; 'FID' to passages of free indirect discourse; 'thought' to passages of silent direct discourse; 'speech' to passages of direct discourse spoken aloud

Text	Discourse	Type count	Styles					
			Objective	Abstract	Literary	Colloquial	Concrete	Subjective
To The Lighthouse	Narration	765	0.00	0.00	0.00	0.00	0.00	0.00
	FID	2,916	0.08	0.17	0.02	−0.02	−0.15	0.02
	Thought	212	−0.15	0.21	0.07	0.30	−0.20	0.08
	Speech	172	−0.32	0.14	0.06	0.49	−0.20	0.11
'The Dead'	Narration	1,325	−0.01	0.02	0.08	0.04	0.00	0.09
	FID	400	−0.13	0.19	0.10	0.19	−0.15	0.11
	Thought	57	−0.43	0.18	0.12	0.74	−0.30	0.22
	Speech	651	−0.11	0.23	0.06	0.27	−0.19	0.16

Table 3. Stylistic profiles for the FID of various characters in *To the Lighthouse* and 'The Dead'. 'Other' refers to all characters in 'The Dead' other than Gabriel considered together

Text	Character	Type count	Styles					
			Objective	Abstract	Literary	Colloquial	Concrete	Subjective
To The Lighthouse	Mrs Ramsay	805	0.07	0.24	0.00	0.03	−0.22	0.03
	Mr Ramsay	70	0.09	0.58	0.27	0.01	−0.49	0.00
	William Bankes	248	−0.01	0.19	0.03	0.14	−0.17	0.08
	Lily Briscoe	1,485	0.06	0.17	0.03	−0.02	−0.15	0.01
	James Ramsay	540	−0.06	0.03	0.08	0.06	−0.03	0.02
	Cam Ramsay	381	−0.10	0.04	0.04	0.10	−0.06	0.00
	Charles Tansley	138	−0.07	0.21	−0.07	0.22	−0.23	0.05
'The Dead'	Gabriel	358	−0.12	0.21	0.12	0.17	−0.17	0.10
	Other	85	−0.30	0.06	0.02	0.44	−0.07	0.14

Tansley and Mr Ramsay (both philosophers, though marked by differences of age and class) are dissimilar in every aspect but abstract. William Bankes, though similar in age to Mr and Mrs Ramsay, is a more unassuming, approachable character, which is reflected in his lower values for the high status styles such as objectiveness and abstractness.

In Table 4, we collapse characters from *To the Lighthouse* into groups based on relevant social categories.[11] Since we are directly comparing two categories in this table, we carried out *t*-tests; gray shading indicates that the differences between two categories of the same factor are statistically significant at the $P < 0.01$ level. Charles Tansley, of working-class origins, is the only character of lower class to be given any FID in *To the Lighthouse*; other characters, such as Macalister and Mrs McNab, speak directly or not at all. Comparing his limited FID with all that of characters of higher class, however, reveals a quite conventional power dynamic: the higher-class characters in *To the Lighthouse* are more objective, more literary, more concrete, less subjective, and far less colloquial. Though the difference between young and old is broadly similar to the difference between classes, there are important distinctions that justify our more detailed stylistic profile: for both class and age, objectiveness (i.e. the projection of authority) is important, but the other key stylistic distinction for age is use of abstract terminology (words that require significant cultural knowledge), whereas the other key stylistic distinctions for class are literariness and colloquialness. In any case, while *To the Lighthouse* exhibits conventional power dynamics in terms of age and class, it almost completely reverses the conventional dynamics for gender. Compared to male characters, female characters in *To the Lighthouse* are generally more objective, more abstract, less colloquial, and less subjective; further, Mr Ramsay's extreme values for literariness and concreteness possibly explain why men rank slightly above women in these categories.

6 Discussion

Our research provides quantitative support for two long-held but seldom-tested hypotheses about FID: that it is an identifiable mode of discourse distinct both from narration and direct discourse, and that it falls stylistically in between these two poles. While a sample of only two texts is of course very limited, the fact that FID functions so similarly in two such dissimilar texts—a novel and a short story; one with an experimental, pervasive use of the FID and one with a more limited, more conventional deployment of the device; one by a female English writer, the other by an Irish male—provides reason to believe that the 'in-betweenness' of FID will be found to apply more generally. We have further shown that is it possible, using our method, to distinguish individual characters' styles in FID, where their personalized expression is presented in an 'impure' mixture with the words of the narrator.

Table 4. Stylistic profiles for the FID of various social groups in *To the Lighthouse*. Gray shading indicates statistically significant difference at the $P < 0.01$ level between the two categories of the same factor

Social identity		Type count[a]	Styles					
Factor	Category		Objective	Abstract	Literary	Colloquial	Concrete	Subjective
Age	Young	969	−0.03	0.06	0.04	0.04	−0.06	0.01
	Old	2,248	0.09	0.21	0.02	−0.02	−0.19	0.02
Class	Low	138	−0.07	0.21	−0.07	0.22	−0.23	0.05
	High	2,844	0.08	0.16	0.03	−0.02	−0.15	0.02
Gender	Female	2,356	0.08	0.18	0.02	−0.02	−0.17	0.01
	Male	878	0.02	0.14	0.06	0.03	−0.12	0.03

[a]Note that when dealing with types, counts cannot be summed, since types may be duplicated across two partitions of the data; this explains why the sum of the types in the two categories is not the same for the various factors in this table, nor do they correspond to the total types for FID in Table 2.

Our approach to character style in particular promises to enrich literary discussion. Our analysis of individual characters' FID in *To the Lighthouse*, for instance, reveals thematically relevant patterns of influence among characters. One of the major themes of the novel is influence across generations: Cam and James, the Ramsays' youngest children, spend much of Part III of the novel brooding on the authority and influence of their parents, wondering what lingering impact their deceased mother has in their life and bristling consciously against the authority of their surviving father. Our stylistic profiles reveal that this generational conflict is expressed at the level of language: the Ramsays and their children speak quite different languages. Lily, the painter who is the central consciousness of Part III, also ponders the Ramsays' influence. Though her feelings toward Mr Ramsay are largely negative, she is ambivalent toward Mrs Ramsay, admiring her deeply yet resisting the conventional gender role she adopted in her family. Though she questions Mrs Ramsay's way of living, however, our stylistic profile reveals that Lily is unquestionably Mrs Ramsay's linguistic heir: in all six categories, their styles are nearly identical. By contrast, the stylistic relationship between Mr Ramsay and his own would-be disciple, Charles Tansley, is one of extreme disparity. Tansley, a young philosopher and a student of Mr Ramsay's, is desperate to belong to the Ramsays' social and intellectual sphere, yet is acutely and bitterly aware of his differences, particularly of social class. Stylistically, all that he shares with Mr Ramsay is a certain philosophical penchant for the abstract; in all other categories, his language is clearly marked by his class.

Moving outside the text itself to the level of authorship, our analysis of sociolinguistic categories in Woolf's FID likewise has the potential to inform long-standing critical debates about Woolf's stylistic politics.[12] That Woolf reverses the conventional stylistic power dynamics for gender—rendering females' FID as more objective, more abstract, less colloquial, and less subjective than that of males—is not unexpected, given that she was among the most prominent and outspoken of early-20th-century feminists. (One of her most important works of feminist non-fiction, *A Room of One's Own*, was published 2 years after *To the Lighthouse*.) That she upholds conventional stylistic power dynamics for class, moreover, may be taken to bolster the widespread criticism that Woolf, born into an upper-middle class London family, was insensitive in her representations of working-class characters (Childers, 1992; Light, 2007). As we have argued elsewhere (Hammond *et al.*, 2013), however, Woolf's representation of class is consistent with what Mikhail Bakhtin calls 'dialogism': the modernist practice of allowing characters to speak in their own distinctive manners, without altering them to suit the particular linguistic practices and prejudices of the author (Bakhtin, 1981). In his landmark analysis of *To the Lighthouse*, Erich Auerbach indeed sees FID as the essential narrative device by which Woolf achieves the 'multi-personal representation of consciousness',

relating the story through the perspectives of the characters themselves with minimal interference from the narrator (Auerbach, 1953).

Turning now to comparison with other approaches, the work of Burrows (1987), which relies on variation in the use of common words to build a low-dimensional vector space using principal component analysis (PCA), has become a standard in the field of literary computing for looking at how authors differentiate characters within literature: examples of this include the work of McKenna and Antonia (1996) and Rybicki (2006). Our approach is distinct in a number of ways, not the least of which is that our dimensions of stylistic variation are designed to be human interpretable, allowing for kinds of analysis that are unavailable with a technique like PCA.[13] For the task of authorship attribution, the origin of these common-word approaches, it is arguably sufficient to simply differentiate. For literary analysis, however, we are interested in how authors use language to shape their readers' perception of characters, and the pathway to this influence is far more likely to lie in the choice of (eye-catching) rarer words and expressions, rather than in the statistics of common ones, which, individually, are uninteresting to the reader. In the context of our interest in social variables, we might also pursue some adaptation of variationalist sociolinguistic methodology (Tagliamonte, 2011), but sociolinguists too tend to focus only on common, individual phenomena where there are sufficient instances in the corpus in question to make statistical generalizations. This, we believe, is simply untenable in the context of literature, where the data are limited and collecting more data is not possible; thus, it is crucial that stylistic information be brought in from an external source. Finally, we note that in the context of a mixed discourse like FID, relying on common words (e.g. 'he', 'she') is problematic since they would be more indicative of the syntactic status of FID (i.e. third-person narration), which is not of interest in our analysis.

More directly comparable to our method is the investigation of character by using pre-existing word classes, particularly syntactic and semantic labels (Balossi, 2014; Culpeper, 2009). One problem with this type of approach is that it is somewhat scattershot, involving large sets of labels from which a few interesting examples are hand-selected for further manual analysis. From a statistical perspective, however, this is troubling, since there are so many categories that at least a few of them are bound to show some statistically significant difference. As Culpeper (2009, p. 52) points out, semantic categories are sometimes dominated by a single keyword, which completely invalidates the use of categories and can lead to false generalizations. Another problem with using existing word classes is that many are simply inappropriate for the analysis at hand, and coverage may be poor in a domain like literature. By contrast, our approach uses a small set of styles that are extremely relevant to characterization, and our method derives a stylistic profile for every word (or expression) appearing in the text based on their distribution in a literary corpus; this profile is included in the sum only once for each word (or expression) type. A more focused word-class approach, perhaps the most similar work to our own, is offered by Deforest and Johnson (2000), who use density of Latinate terms in the speech of Jane Austen's characters to identify characters that Austen wished to characterize as pretentious. We note that, like us, they rely on somewhat imprecise methods for classifying their words, but nonetheless offer fairly compelling results. Our work, however, benefits from having a variety of styles (for instance, being able to distinguish the effects of class from age); though in English many Latinate words have connotations that place them on the written end of the cline of register, this distinction is not at all categorical, and word co-occurrence has been shown to be a more powerful tool for quantifying lexical formality in general (Brooke *et al.*, 2010).

Along with some of the work mentioned above, our approach is notable for using the output of a model based on computational linguistics for literary analysis, which we have argued elsewhere is an important direction for both fields (Hammond *et al.*, 2013). Other recent work in this vein includes that of Kao and Jurafsky (2012), Brooke *et al.* (2012), He *et al.* (2013), Voigt and Jurafsky (2013), and Bamman *et al.* (2014). There is an obvious tension here, however, between typical methodologies of computational linguistics, where the

goal is to demonstrate that a particular approach is successful through comparison to objective gold standards, and those of literary analysis, which succeed when they offer interesting insights beyond what is obvious or generally accepted. Although we have diffused some of this tension by using a method of stylistic lexicon acquisition that has been independently evaluated at both the word and sentence level (Brooke *et al.*, 2014b), many of our results are indeed fairly predictable, and should be viewed primarily as additional evidence that our automatically generated lexicons are reliable enough to do useful analysis. The ultimate goal of this article is to contribute to literary analysis of the works in question and offer a better understanding of FID; still, the various implicit evaluations that result from applying these automatic lexicons in this context is an important aspect of this work, and is conducive to furthering such interdisciplinary efforts.

Finally, although there is necessarily some degree of noise introduced when applying more complex computational techniques to literary analysis, we also argue they have more long-term potential than work which is restricted purely to raw corpus statistics (e.g. word counts), particularly in the context of large corpus analysis (which is one direction for future work). Though imperfect, these sorts of models combine the objectivity of quantificational methods with human-interpretable generalizations of linguistic phenomena—a powerful combination for bringing literary study closer to ideals of scientific research.

7 Conclusion

The work presented here represents the first computational, quantitative analysis of the phenomena of FID in its modernist form. Our method is also an important step forward, in that we rely not only on annotation of texts of interest, but also consider the lexical stylistic information contained in tens of thousands of other texts, which allows us to derive a human-interpretable stylistic profile of even fairly minor characters. Using these lexical stylistic profiles, we showed that FID does indeed deserve its reputation as a mixture of narration and direct discourse, and that Woolf and Joyce quite clearly used lexical choice to distinguish the social backgrounds of their characters, even when the character's viewpoint is being presented in a mixture with third-person narration.

References

Abbott, H. P. (2008). *The Cambridge Introduction to Narrative*. 2nd edn. Cambridge: Cambridge University Press.

Auerbach, E. (1953). *Mimesis: The Representation of Reality in Western Literature*. Princeton, NJ: Princeton University Press.

Austen, J. (1813). *Pride and Prejudice*. London: T. Egerton.

Bakhtin, M. M. (1981). Discourse in the novel. In Holquist, M. (ed.), *The Dialogic Imagination: Four Essays*. Austin: University of Texas Press, pp. 259–422.

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Bamman, D., Underwood, T., and Smith, N. (2014). A bayesian mixed effects model of literary character. In *Proceedings of the 52st Annual Meeting of the Association for Computational Linguistics*. Baltimore, MA, June 2014, pp. 370–9.

Balossi, G. (2014). *A Corpus Linguistic Approach to Literary Language and Characterization: Virginia Woolf's the Waves*. Philadelphia, PA: John Benjamins.

Beigman Klebanov, B. and Beigman, E. (2009). From annotator agreement to noise models. *Computational Linguistics*, 35(4): 495–503.

Brooke, J., Hammond, A., and Hirst, G. (2012). Unsupervised stylistic segmentation of poetry with change curves and extrinsic features. In *Proceedings of the NAACL '12 Workshop on Computational Linguistics for Literature*. Montreal, QC, June 2012, pp. 26–35.

Brooke, J., Hammond, A., and Hirst, G. (2015). GutenTag: an NLP-driven tool for digital humanities research in the Project Gutenberg corpus. In *Proceedings of the NAACL '15 Workshop on Computational Linguistics for Literature*. Denver, CO, June 2015, pp. 42–7.

Brooke, J. and Hirst, G. (2013a). A multi-dimensional Bayesian approach to lexical style. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, GA, June 2013, pp. 673–9.

Brooke, J. and Hirst, G. (2013b). Hybrid models for lexical acquisition of correlated styles. In *Proceedings of the International Joint Conference on Natural Language Processing*. (IJCNLP '13). Nagoya, Japan, October 2013, pp. 82–90.

Brooke, J. and Hirst G. (2014). Supervised ranking of co-occurrence profiles for acquisition of continuous lexical attributes. In *Proceedings of the 25th International Conference on Computational Linguistics*. Dublin, Ireland, August 2014, pp. 753–61.

Brooke, J., Tsang, V., Hirst, G., and Shein, F. (2014). Unsupervised Multiword Segmentation of large corpora using prediction-driven decomposition of *n*-grams. In *Proceedings of the 25th International Conference on Computational Linguistics*. Dublin, Ireland, August 2014, pp. 2172–83.

Brooke, J., Wang, T., and Hirst, G. (2010). Automatic acquisition of lexical formality. In *Proceedings of the 23rd International Conference on Computational Linguistics: Poster Volume*. . Beijing, China, August 2010, pp. 90–8.

Burrows, J. F. (1987). *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*. Oxford: Clarendon Press.

Brunner, A. (2013). Automatic recognition of speech, thought, and writing representation in German narrative texts. *Literary and Linguistic Computing*, 28(4): 563–75.

Childers, M. M. (1992). Virginia Woolf on the outside looking down: reflections on the class of women. *Modern Fiction Studies*, 38(1): 61–79.

Coltheart, M. (1980). *MRC Psycholinguistic Database User Manual: Version 1*. London: Birkbeck College.

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): 29–59.

DeForest, M. M. and Johnson, E. (2000). Computing latinate word usage in Jane Austen's novels. *Computers and Texts*, 18/19 24–5.

Hammond A., Brooke J. and Hirst G. (2013). A tale of two cultures: bringing literary analysis and computational linguistics together. In *Proceedings of the NAACL 13 Workshop on Computational Linguistics for Literature*. Atlanta, GA, June 2013, pp. 1–8.

Hota, S., Argamon, S., Koppel, M., and Zigdon, I. (2006). Performing gender: automatic stylistic analysis of Shakespeare's characters. In *Digital Humanities*, Paris, France, July 2006, pp. 82–8.

He, H., Barbosa, D., and Kondrak, G. (2013). Identification of speakers in novels. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria, August 2013, pp. 1312–20.

Joachims, T. (2002). Optimizing search engines using click through data. In *Proceedings of the 9th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Edmonton, AB, July 2002, pp. 133–42.

Joyce, J. (1914). The dead. In *Dubliners*. London: Grant Richards, pp. 216–78.

Kao, J. and Jurafsky, D. (2012). A computational analysis of style, sentiment, and imagery in contemporary poetry. In *Proceedings of the NAACL '12 Workshop on Computational Linguistics for Literature*. Montreal, QC, June 2012, pp. 8–17.

Kenner, H. (1978). *Joyce's Voices*. Berkeley, CA: University of California Press.

Leckie-Tarry, H. (1995). *Language and Context: A Functional Linguistic Theory of Register*. London: Pinter.

Light, A. (2007). *Mrs Woolf and the Servants*. London: Fig Tree.

McKenna, C. W. F., and Antonia, A. (1996). 'A few simple words' of interior monologue in *Ulysses*: reconfiguring the evidence. *Literary and Linguistic Computing*, 11(2): 55–66.

Pascal, R. (1977). *The Dual Voice*. Manchester: Manchester University Press.

Pechey, G. (2007). *Mikhail Bakhtin: The Word in the World*. London: Routledge.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). *Linguistic Inquiry and Word Count (LIWC): LIWC2001 Manual*. Mahwah, NJ: Erlbaum Publishers.

Ramsay, S. and Steger, S. (2006). Distinguished speakers: keyword extractions and critical analysis with Virginia Woolf's the waves. In *Digital Humanities*, Paris: Sorbonne, pp. 5–9

Rundquist, E. (2014). How is Mrs. Ramsay thinking? The semantic effects of consciousness presentation categories within free indirect style. *Language and Literature*, 23(2): 159–74.

Rybicki, J. (2006). Burrowing into translation: character idiolects in Henryk Sienkiewicz's trilogy and its two English translations. *Literary and Linguistic Computing*, 21(1): 91–103.

Sotirova, V. (2013). *Consciousness in Modernist Fiction: A Stylistic Study*. New York: Palgrave Macmillan.

Stone, P. J., Dunphy, D. C., Smith, M. S., and Ogilivie, D. M. (1966). *The General Inquirer: A Computer Approach to Content Analysis*. Boston, MA: MIT Press.

Strunk, W. and White, E. B. (1979). *The Elements of Style*. 3rd edn. New York: Macmillan.

Tagliamonte, S. A. (2011). *Variationist Sociolinguistics: Change, Observation, Interpretation*. Toronto: Wiley-Blackwell Publishers.

Voigt, R. and Jurafsky, D. (2013). Tradition and modernity in 20th century Chinese poetry. In *Proceedings of the NAACL '13 Workshop on Computational Linguistics for Literature*. . Atlanta, GA, June 2013, pp. 17–22.

Wiebe, J. M. (1994). Tracking point of view in narrative. *Computational Linguistics*, 20(2): 233–87.

Woolf, V. (1927). *To the Lighthouse*. London: Hogarth Press.

Notes

1 Our use of the word 'voice' in this article should not be confounded with the linguistic meaning of 'voice', e.g. active voice.

2 We believe our agreement is indeed moderate, which is to say well above chance but probably below what is generally required for a single-annotator gold standard for a typical 'objective' annotation task. It is, however, not trivial to calculate a useful inter-annotator agreement score in this case. The most obvious choice, chance-corrected Fleiss's Kappa, assumes a simple (non-hierarchical) set of labels, a fixed set of judged instances, and fixed set of annotators, none of which is present here. Readers who wish to gain a better sense of the extent of variation among annotators are invited to visit the project Web site (http://brownstocking.org) where the different annotations for each span can be directly compared.

3 Although we initially allowed students to tag indirect discourse, there is very little indirect discourse in these texts, and we do not include it in our analysis.

4 http://brownstocking.org/

5 http://livingdead.ca/

6 This is not to be confused with the linguistic meaning of 'aspect'; we are referring to aspects (subcategories) of style.

7 http://www.urbandictionary.com

8 http://www.gutenberg.org

9 http://www.projectgutentag.org

10 We also tested the idea of simply removing words that appeared often and then using tokens, but this requires choosing an arbitrary cutoff point; and moreover when the term-frequency effect is removed, common words do not have much influence on our scores due to Zipfian effects, i.e. most of the types are in fact hapax legomena. Thus, using types seemed to us to be a more principled approach.

11 See the Appendix for details of our sociolinguistic categories.

12 Although the status of the author has been much questioned in 20th-century and contemporary literary theory, we embrace the assumption that the author is an autonomous subject outside the text who creates it and crafts the voices within it. This is not to say that we see any particular voice within the text (for instance, the narrator) as an authorial stand-in.

13 Although it is, of course possible in some cases to interpret PCA dimensions after the fact—for instance the dimensions of register identified by Biber (1988) using his multidimensional analysis approach—the use of common words as features does not lend itself to post hoc analysis.

Appendix

Our sociolinguistic distinctions are based on attributes manually recorded for each character in a TEI personography. Definitions of categories such as 'young' vary between texts: in *To the Lighthouse*, which has a younger cast of characters, 'young' means below 30 years of age, whereas in 'The Dead', where the characters are generally older, 'young' means below 40 years. Our classification of class is based on National Research Survey social grade distinctions current in mid-20th-century Britain. Our scale runs from 0 to 5; for both texts, the cutoff for 'lower' and 'higher' occurs between 2 (skilled manual workers; clerical; some junior managerial, administrative) and 3 (intermediate managerial, administrative, professional). For children, we assign the class of parents. Charts showing the relative numbers of characters belonging to these categories, with examples, are provided below. Where numbers vary it is because insufficient information is present in the text (i.e. in 'The Dead', where the sum of young and old characters is less than that of lower- and higher-class characters, this is because of the great number of minor characters whose age is not clearly presented).

Table A1 Sociolinguistic distinctions in *To the Lighthouse*

Social identity		Number	Examples
Factor	**Category**		
Age	Young	13	Minta Doyle, Macalister's boy, Marie, Andrew Ramsay
	Old	8	William Bankes, Mrs Beckwith, Mrs Ramsay, Augustus Carmichael
Class	Low	5	Charles Tansley, Marie, Mrs McNab, Macalister, Macalister's boy
	High	16	William Bankes, Mrs Beckwith, Augustus Carmichael, Mr Ramsay, Mrs Ramsay
Gender	Female	10	Mrs Beckwith, Lily Briscoe, Minta Doyle
	Male	11	William Bankes, Augustus Carmichael, Macalister

Table A2 Sociolinguistic distinctions in 'The Dead'

Social identity		Number	Examples
Factor	**Category**		
Age	Young	7	Gabriel Conroy, Gretta Conroy, Molly Ivors, Lily
	Old	5	Mr Browne, Julia Morkan, Kate Morkan
Class	Low	2	Michael Furey, Lily
	High	22	Gabriel Conroy, Molly Ivors, Julia Morkan, Kate Morkan
Gender	Female	14	Mrs Cassidy, Gretta Conroy, Miss Daly, Miss Furlong
	Male	10	Mr Bergin, Mr Browne, Mr Clancy