Dialogism in the novel: A computational model of the dialogic nature of narration and quotations

Grace Muzny
Department of Computer Science, Stanford University, USA

Mark Algee-Hewitt
Department of English, Stanford University, USA

Dan Jurafsky
Department of Computer Science, Stanford University, USA and
Department of Linguistics, Stanford University, USA

Abstract

Understanding how spoken language is represented in novels over time is a key question in the Digital Humanities. We propose a new metric for characterizing spoken dialogue in the novel, called dialogism, that instantiates Bakhtin's claim that all texts are fundamentally dialogic. This measurement uses abstract grammatical features in a span of text (such as the use of pronouns, mood, or subordinate clause structure) to measure the extent to which the span is dialogic, i.e. exhibits the grammatical structures common to natural spoken dialogue. We use this metric to explore the dialogism of 1,100 largely canonical English novels over 230 years. We combine quantitative and qualitative investigation of the dialogic properties of both dialogue and narration to show novel stylistic properties of literary innovation during three periods: the late 18th century, the turn of the 19th century, and the mid-20th century. We find that during these moments, certain authors reject literary conventions by changing the dynamic between the narrative and dialogue portions of the texts. Our analysis shows that these changed dynamics are behind rises in persuasive writing, reflections of psychological processes, and the use of dialogue as an increasingly important driver of the novel as a whole. These results show that computational models that characterize style grammatically, generalizing across time and genre, can lead to literary and methodological insights.

Correspondence:
Grace Muzny, Department of Computer Science 353 Serra Mall, Stanford 94305, CA.
E-mail:
muzny@stanford.edu

1 Introduction

A model of literary style is central to our understanding of the novel. The study of style—the distribution of linguistic elements underlying literary form and the composition of these elements to create stylistic effects—can be a powerful tool for analyzing a novel's genre, meaning, and structure. Style can help us understand how authors innovate and how they respond to their historical and social

doi:10.1093/llc/fqx031

contexts (Allison *et al.*, 2013). Yet it is challenging to develop a model of style on a larger scale that can reveal stylistic patterns or moments of innovation without overfitting a specific genre or tradition.

One important direction in stylistics focuses on the individual words that characterize particular authors, movements, or periods. The close relationship between specific words and the authorial context they were written in and the subjects that they address makes them a powerful, lexical, content feature. Alas, this same property places lexical features at a disadvantage when it comes to making generalizations. This is because words correlate with so many parts of the novel, from specific plot to genre to historical context, that it is impossible to disentangle these effects from those of literary style. Instead we need a characterization of literary style that, unlike lexical features, avoids overfitting to one particular genre or moment in literary history.

We propose to solve this problem by drawing instead on a different linguistic domain—grammar rather than the lexicon—to formulate a flexible notion of literary style that can bring statistical insight across a wide variety of literary contexts. Our new model of literary style draws on Bakhtin (1935), who argued that all texts, all discourses, are fundamentally dialogic. Words, in Bakhtin's argument, partly acquire meaning through the text's dialogic relationship with readers, with contexts, and with other words. From this perspective 'the internal dialogism of the word... the dialogism that penetrates its entire structure' (p. 279) plays a crucial role in how a text works. Bakhtin pointed out that even narrative (non-quotational) regions of texts contain 'hidden, diffused speech', an analogue to spoken dialogue in which 'the speech of another is introduced into the author's discourse (the story) in *concealed form*, that is, without any of the *formal* markers usually accompanying such speech, whether direct or indirect' (p. 303).

Bakhtin's idea of speech 'in concealed form' offers a new direction for analyzing literary style. Taking him at his word, dialogue itself becomes a site of the dialogic encounter: perhaps the critical site of the competing voices that Bakhtin describes. Although it is key to Gérard Genette's dichotomy between narrative time and story time, dialogue itself plays a relatively small role in his narratology (Genette, 1983). Similarly, in her study of the poetics of conversation, Deborah Tannen juxtaposes conversational with literary discourse, eliding the 'literary conversational discourse' that, we argue, animates Bakhtin's theory of the novel (Tannen, 1987). As such, it is his concept of dialogism, rather than its subsequent transformations in current narratology that animates our question. In this study, therefore, we examine the ways in which narrative text and direct quotation both draw on the grammatical properties of spoken dialog. Rather than leaving dialogue out of stylometric analysis because of the fear of authorial noise (Hoover, 2005), we interpret dialogue and the dialogic context of all words as primary features of the author's stylistic landscape.

Our plan for implementing dialogism is to draw on the fact that dialogue is associated with rich grammatical features like tense, modality, deixis, and clause structure. Spoken, or conversational, language tends to be in the present tense, to use short clauses, modals, and to employ 1st- and 2nd-person pronouns. We suggest that these grammatical aspects of dialogue are inescapably intertwined with literary style. The extent to which a part of a text exhibits the grammatical properties of dialogue (is in the present tense, uses modals, refers to 1st and 2nd person, etc.) thus reveals its stylistic underpinnings.

Access to a large corpus of dialogue is essential for finding and distilling the underlying characteristics of dialogue and dialogism. We therefore develop new tools for extracting dialogue—instances of direct quoted speech—that are optimized for the digital analysis of variable quality texts. Using these tools on a corpus of over 1,000 novels spanning three centuries, we create a new corpus of dialogue composed of more than 2 million instances of quoted speech.

This new dialogue corpus allows us to uncover the abstract grammatical features (modality, tense, etc.) that characterize spoken dialogue. The result is a new metric quantifying the dialogic context of any text, regardless of its genre or period, that we call dialogism. We deliberately make our model lexically blind—we do not use features based on specific words. Focusing only on higher-level, abstract grammatical features makes our metric appropriate for evaluating the relative dialogism of a text more

robustly against variation in context, genre, literary movement, and time period.

We use our new quote extraction tools and dialogism metric to identify three historically bounded shifts in stylistic innovation within our corpus. At each historical moment, the relationship between narration and dialogue undergoes a significant transformation which corresponds to an accompanying change in literary style.

The first innovation that we examine occurs at the turn of the 19th century when the emergence of novels of ideas offered a stark contrast to the popular genres of the period. Although the contrast between these genres is typically ascribed to the inclusion of philosophic content in the novel, here we find that underlying these changes is a stylistic innovation in both the narration and dialogue of the text.

In the second historical shift, we discover that modernist experimentation is more expansive than previously thought. Authors of the period thoughtfully engage with dialogue and dialogic text by crafting 'realistic' dialogue and using complex grammatical structures to separate it from narration.

Finally, we add a new level of critical understanding to the function of contemporary literary fiction, and its relationship to the early 20th century. In so doing, we discover a new, stylistically linked category of late 20th-century novels that we call conversation fiction. This category links authors from William S. Burroughs to Stephen King, Philip Roth, and Margaret Atwood. Authors who wrote literary fiction at this time, like David Foster Wallace, innovate by pushing back against the more realist formulation of dialogue that we identified in the earlier period.

Bakhtin's insight that all text is dialogic, taken at face value, leads us to a new kind of stylistic analysis of novels, one that emphasizes the vitally important role that dialogue plays throughout the novel.

2 Data

Our corpus consists of 1,100 canonical novels published from 1782 to 2011. The portions of this corpus published before 1900 come from the Chadwyck-Healey British, American, and Irish corpora. The novels from 1900 to 2011 are from the Stanford 20th-century corpus, which features works drawn from both critical and popular 'Best Novels' lists of the 20th century and represents Anglophone world literature.[1]

Our corpus is therefore large and represents a reasonably long time span, a diverse set of genres, and a wide array of literary styles. Each of these properties is necessary for the goal of understanding literary style. A sufficiently large corpus makes it more likely that any results hold with sufficient validity, and the balance of genres and styles is necessary to ensure that our model is not biased toward subtypes of dialogue associated with only certain kinds of fiction. For example, a model built solely based on entertainment-oriented fiction from the beginning of the 19th century might overfit this genre or period, and fail to generalize to the dialogic elements of post-modern novels.

The later half of the 19th century is slightly overrepresented in our corpus, containing 36% of our 1,100 novels in a 40-year time span, due to the higher relative availability of digital texts from this particular time period. A histogram of the distribution of our texts by decade is shown in Fig. 1. We chose not take a subset of our data because retaining all of the novels helps diversify the data set in terms of genre and style.

In all, 422 authors are represented in our corpus, with nineteen authors contributing ten or more texts and forty-two authors contributing five or more texts. Table 1 shows the twenty most represented authors and the number of novels that they contributed. We do not attempt to categorize our novels based on genre nor do we analyze their narrative point of view, though we recognize that these factors can influence our analysis.

Our corpus thus characterizes the majority of the English canon for the past two centuries, covering a diversity of genres and literary styles, and hence representing most of the variation that occurs in the linguistic representation of dialogue in English fiction.

2.1 Quote extraction and attribution

Creating a robust model of dialogic text hinges on being able to reliably extract dialogue from a novel. Simple regular expression-based algorithms are effective at extracting from clean text with reliably

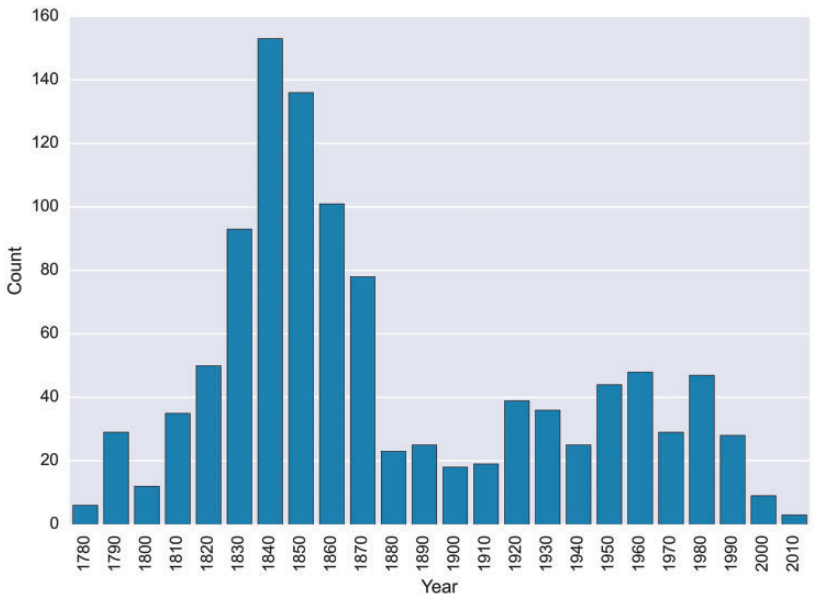

Fig. 1 Number of books per decade in our corpus

Table 1 Most represented authors in the corpus and the number of texts that they contribute

Author	Number of novels
Ingraham	33
Cooper	30
Simms	26
Hawthorne	20
Herbert	20
Trollope	16
Dickens	16
Holmes	16
Bennett	16
Cooke	16
Sedgwick	14
Scott	14
Paulding	13
Thackeray	12
Lippard	11
John Grisham	11
Hardy	11
Stowe	11
Thompson	10
Ward	9

delineated quotes, but do not do as well on the novels in our corpus, which sometimes have messy formatting due to Optical Character Recognition (OCR) errors and changing publication standards; Table 2 outlines typical problematic cases.[2]

We introduce a deterministic quote extraction system, QuoteAnnotator,[3] that deals with these common formatting issues by evaluating each candidate quotation mark in context to decide if it is starting or ending a quote, following the set of rules

Table 2 Examples of general types of quotes and formats that frequently cause difficulties for quote extraction systems

Quote	Difficulty
"Miss, "My Lord orders me to acquaint you, that in consequence . . ."	Multi-paragraph quotation
"that it's done by everybody minding their own business!"	Apostrophes in quotes
'Oh, 'tis love, 'tis love, that makes the world go round!'	Quote delineator also used as apostrophe
"'tis our Miss as Mounseer means . . . 'tis a dumb beauty"	Wrong directionality quotation marks

Table 3 Rules for the QuoteAnnotator that help it extract quotes across a wide range of formats

#	Case	Rule
1	Quote delineated by regular ASCII quote marks (")	Ending quote must be followed by white space or punctuation
2	Quote delineated by single quote marks (')	Starting quote must be preceded by white space or punctuation and ending quote must be followed by white space or punctuation
3	Quote delineated by smart quotation marks (" ")	Ending quote must pair with beginning quote and level of embeddedness must match
4	Smart quotes face the wrong direction (" ")	Attempt extraction with quotes converted to ASCII
5	Problematic apostrophes	Attempt extraction disallowing single-quote delineated quotes

Table 4 Precision and recall of the QuoteAnnotator versus bookNLP on a selection of novels from our corpus that have been hand annotated for quote boundaries

Novel	QuoteAnnotator				BookNLP			
	P	R	Partial P	Partial R	P	R	Partial P	Partial R
Emmeline the Orphan, Ch. I–X, Smith	0.42	0.38	0.92	0.35	0.05	0.07	1	0.07
Pride and Prejudice, Austen	1	1	1	1	0.98	0.98	1	0.98
The Spy, Vol. 1, Cooper	0.77	0.76	0.95	0.72	0.69	0.69	0.95	0.65
A Room with a View, Forster	0.97	0.97	1	0.97	0.97	0.97	1	0.97
The Great Gatsby, Fitzgerald	0.96	0.97	1	0.97	0.97	0.98	1	0.98
The Litigators, Grisham	0.98	0.96	0.96	0.93	0.97	0.95	0.96	0.92
Average scores	**0.85**	**0.84**	0.97	**0.82**	0.77	0.77	**0.99**	0.77

Bold values indicate the systems with the highest precision and recall scores.

shown in Table 3. The algorithm thus offers the advantages of regular expressions while still dealing with difficult cases. Our system also incorporates fail safe measures like limiting the length of extracted quotations, allowing it to recover from garden paths caused by bad quotations. In addition, it is able to extract quotes using a variety of settings, such as allowing or disallowing single quotes, and then

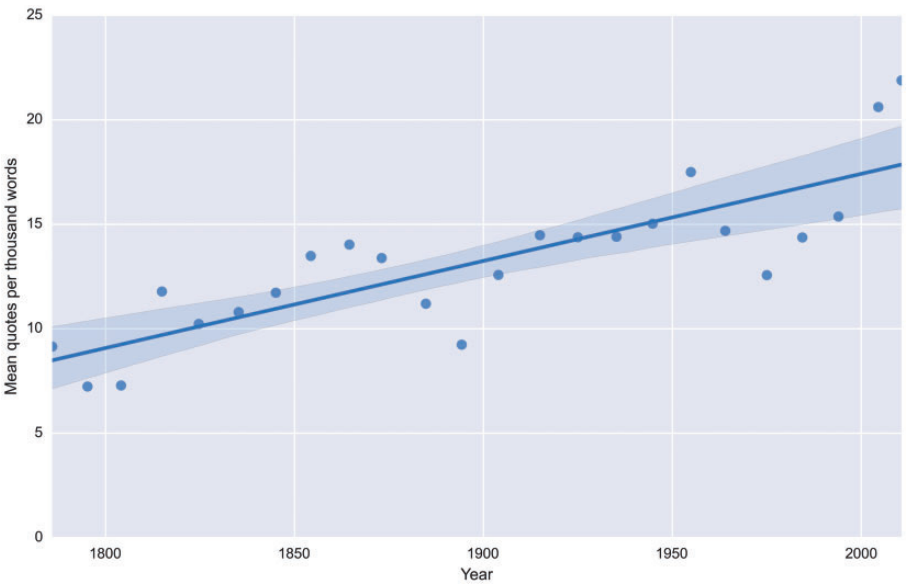

Fig. 2 Number of quotes per thousand words from 1782 to 2011, bucketed per decade, rounded down, so all novels from 1800 to 1809 contribute to the 1800 data point. This approach is used in the all bucketed figures. The line of best fit shows an upward-sloping trend with an r^2 value of 0.65

choose the configuration that results in the largest number of successfully extracted quotations. Of course, while the strategies that we take help tackle problematically delineated quotes, they are not a complete solution.

We built an evaluation test set by choosing six novels that are distributed across time and contain many quotations. Three trained annotators then hand annotated every quote, with an inter-annotator agreement of $\kappa = 0.97$ on the training sample. Following Hollingsworth and Teufel (2005), we then evaluate via both standard precision and recall (exact quote match of every word between system hypothesis and gold labels) and partial precision and recall (measuring the overlap in words between system hypotheses and the gold labels).

In Table 4 we report the results of our system and those of a baseline, bookNLP (Bamman *et al.*, 2014). While bookNLP has a slightly higher partial precision score (almost every token that it claims is inside a quote is actually inside one) overall, our system's high recall, better exact precision, and better performance on texts with lower-quality digitization make it the right tool for our corpus.

Applying the QuoteAnnotator to our corpus resulted in an extracted data set of more than 2 million quotes.

2.2 Dialogic landscape

We first examine the broad temporal dynamics of dialogue across our corpus, asking whether novels have become more dialogue driven over time. We compute the density of quotes per thousand words and compute the slope of the line of best fit. The results confirm that novels incorporate more dialogue over time. Figure 2 shows that authors add roughly one quote per thousand words every 25 years, equivalent to just under one more quote per page every century.

3 Modeling Dialogism

The previous section showed that the use of dialogue is increasing over time. This increase, however, does not tell us about changes in the nature of the dialogue itself, or changes in the relation between dialogue and narration. For that we draw from methods in corpus linguistics (Biber, 1991) to create a new

model that measures the degree to which a span of text is dialogic in nature, a metric called dialogism.

The intuition of our metric is that spoken dialogue exhibits particular grammatical features: it tends to be in the present tense, use short clauses and modal verbs, and refer to 1st- and 2nd-person pronouns. The more dialogic a sentence or other piece of text is, the more it uses these grammatical characteristics of dialogue. In contrast, narrative (non-dialogue) text is more likely to use grammatical characteristics like 3rd-person pronouns and past tense.

Crucially, the dialogism metric does not just apply to literal quoted spoken dialogue. As we discussed above, Bakhtin (1935) pointed out that even narrative (non-quotational) regions of texts contain aspects of spoken dialogue, 'hidden, diffused speech'. Following this idea, we propose to apply our dialogism metric to every sentence. Practically speaking, such a metric would assign a high score to classically conversation-oriented text like the following:

> Really, Dinah ought to have taught you better manners! You ought, Dinah, you know you ought!

and low scores to classically narrative text like the following:

> Celia's face had the shadow of a pouting expression in it, the full presence of the pout being kept back by an habitual awe of Dorothea and principle; two associated facts which might show a mysterious electricity if you touched them incautiously.

We expect our metric to give high scores to spans of text that come from the point of view of the speaker, that have an audience, and that address events in the here and now. We expect lower scores to be given to texts that are more descriptive, theoretical, and that have more complex grammatical structure. Thus while we expect high-scoring spans of text to be more dialogic, this does not mean that such spans necessarily consist of spoken dialogue. Similarly, we expect low-scoring spans to be less likely to be direct quotation, but this does not mean that they are necessarily narration. While there is a relationship between narrative point of view and dialogism, this relationship is complex and as such remains a study for future work.

3.1 Dialogism: A metric

We propose to build a model of dialogism based on grammatical categories like parts of speech, to avoid the genre and era specificity that plagues models based on words. We draw our intuition from the corpus linguistic work of Biber (1991), who looked at the difference between written language and transcriptions of spoken dialogue. Writing tends to contain informational, narrative text that is carefully crafted and highly edited, with a higher degree of abstractness.

Biber showed that these characteristics of speech versus writing could be captured by differences in the use of parts of speech and related grammatical structures, such as nouns, attributive adjectives, past tense verbs, and passive constructions. Our metric therefore uses the distribution of part-of-speech tags to characterize the grammatical properties of dialogue versus non-dialogue.

We first use the QuoteAnnotator to extract all quotation and non-quotation text spans from our novel corpus, resulting in a labeled corpus that includes 2,231,793 dialogue utterances containing 3,968,805 sentences.

We next run the Stanford Part of Speech (POS) tagger (Toutanova *et al.*, 2003) on the dialogue sentences and compute the distribution over tags.[4] Dialogue, for example, is characterized by high quantities of the bare verb tag, or the 1st-person pronoun tag. In contrast, non-dialogue text is characterized by higher frequencies of adjectives and determiners.

We then use multi-dimensional analysis (MDA) to reduce the model to a small number of factors. MDA is a dimensionality reduction technique that takes input data with many features or dimensions and produces output factors that group positively or negatively correlated features together based on the input data. In our application, MDA forms output factors that group POS tags that appear together either unusually frequently or unusually rarely.

It considers both groups of POS tags that are positive indicators of dialogue and those that are negative indicators of dialogue, as shown in Table 5. We use only the eight factors (of the seventeen total

Table 5 The eight factors that reliably distinguish narration from dialogue produced by using MDA on the dialogue in our corpus

Factor	Direction of dialogism (δ)	Correlated features
1	+	Interjections
		1st-person pronouns
		2nd-person pronouns
		Non-3rd-person singular present tense verbs
		Modals
		Bare verbs
2	+	Existential there
		'it' pronouns
		3rd-person singular present tense verbs
3	−	Gerunds
4	−	Coordinating conjunctions
		Prepositions and subordinating conjunctions
5	−	Adjectives
6	−	Determiners
		Nouns
7	−	Past tense verbs
8	−	3rd-person pronouns

Note: The QuoteAnnotator uses a maximum quote length of 30,000 characters.

produced by MDA) with $r > 0.5$ when used to separate narration from dialogue.[5]

To score a text, in our case a contiguous span of narration or dialogue, we compute the POS tags of our target text and count the tags to arrive at a distribution over tags, and then use the model to calculate a dialogism score. The result is a model that is robust across genre, era, and corpus.

Factors have weights δ, positive (+) if the factor is associated with high dialogism, negative (−) if the factor is associated with low dialogism.[6]

The factors in Table 5 with a positive δ are indeed those known to be associated with spoken language. Modal verbs, for example, are strongly associated with spoken language because they are used to indicate desire, obligation, and ability, and bare verbs (forms like 'go' or 'take') often appear after modal verbs. On the other hand, parts of speech like 3rd-person pronouns make sense as indicators of narration because they are related to talking about other people—storytelling, in short.

These findings correspond well with the markers that Biber (1991) found distinguish speech from text, although we find a somewhat smaller number of distinguishing features. This difference is not surprising since, even though written dialogue in novels is an attempt to mimic speech, it nonetheless naturally exhibits some grammatical properties more common to written text than does transcribed speech.

A score for each text is arrived at by combining the scores for each factor according to Equation (1). In this equation, we sum over each feature f in the target factor. For each feature f, we calculate text_f, the percentage of f the target text; μ_f, the mean percentage of f in the whole corpus; σ_f, the standard deviation of f in the whole corpus; and ϕ, the sign of f in the factor; in our model, ϕ is always 1. This means that the score of a single factor for a text is the weighted combination of the degree to which that text is different from the mean for each feature:

$$score_{\text{factor}} = \sum_{f \in \text{factor}} \phi * \frac{text_f - \mu_f}{\sigma_f} \qquad (1)$$

Let us walk through an example. From an input text, the algorithm first computes the POS tag for each token.

(1) "Really, Dinah ought to have taught you better manners!
Ø RB(ALL) Ø Ø MD Ø VB VB(PAST) PRP2ND JJ(ALL) NN(S) Ø
You ought, Dinah, you know you ought! "
PRP2ND MD Ø Ø Ø PRP2ND VBP PRP2ND MD Ø Ø

Next, it computes the presence of each feature in our target factor (e.g. 2nd-person pronouns).

(2) "Really, Dinah ought to have taught you better manners!
∅ RB(ALL) ∅ ∅ MD ∅ VB VB(PAST) **PRP2ND** JJ(ALL) NN(S) ∅
You ought, Dinah, you know you ought! "
PRP2ND MD ∅ ∅ ∅ **PRP2ND** VBP **PRP2ND** MD ∅ ∅

This score, adjusted according to μ_{PRP2ND} and σ_{PRP2ND}, is then combined with the scores for 1st-person pronouns, non-3rd-person singular present tense verbs, interjections, modals, and bare verbs to form $score_{factor_1}$.

Dialogism is then calculated by combining factor scores according to the sign of the t-score, or δ, associated with the distribution for each factor over the whole corpus. Recall that the set of factors contains just the factors that effectively separate narrative and dialogic text. Effectively, this means that the overall score for a text is calculated by adding the score for each factor associated with dialogue and subtracting the score for each factor associated with narration.

$$dialogism = \sum_{factor \in factors} \delta * score_{factor} \qquad (2)$$

This results in an overall result in which each word may contribute positively or negatively to the dialogism score according to their abstract grammatical features. We can visualize our example sentence by color coding words so that positive contributors are shown in bold and negative contributors are shown in italics.

'Really, Dinah **ought** to **have** *taught* **you** *better manners*! **You ought**, Dinah, **you know you ought**!'

In the end, our dialogism metric achieves high overall separation between narration and dialogue (Kolmogorov–Smirnov value = 0.90, which indicates that a classifier based on this metric would have a relatively high f-score), shown in Fig. 3, displaying the distribution of dialogism scores for the narration and dialogue in our corpus. The high separation means that narrative and descriptive texts are well differentiated by the model. The model is also easily transferred to other settings and different

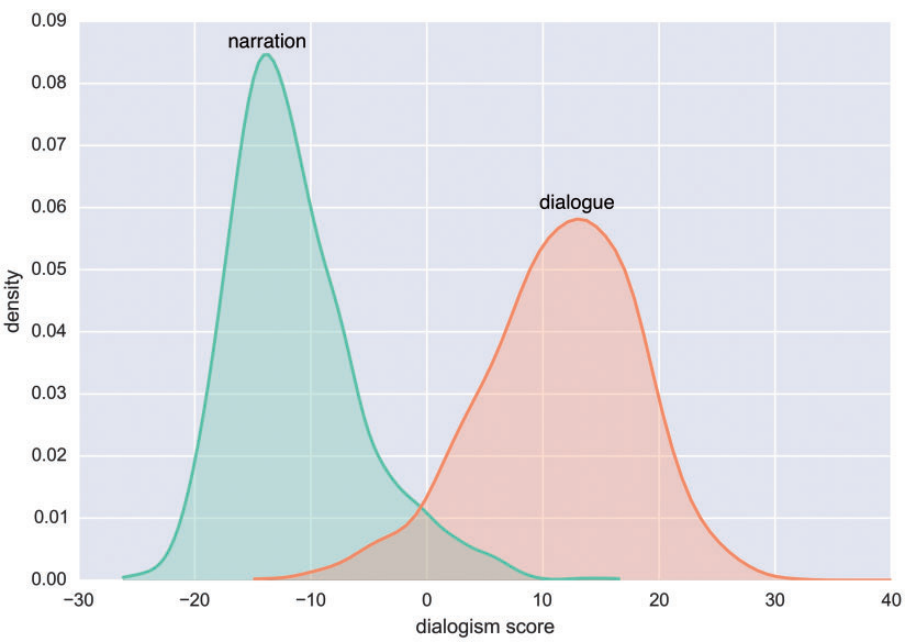

Fig. 3 Dialogism scores for narration and dialogue for all novels in our corpus

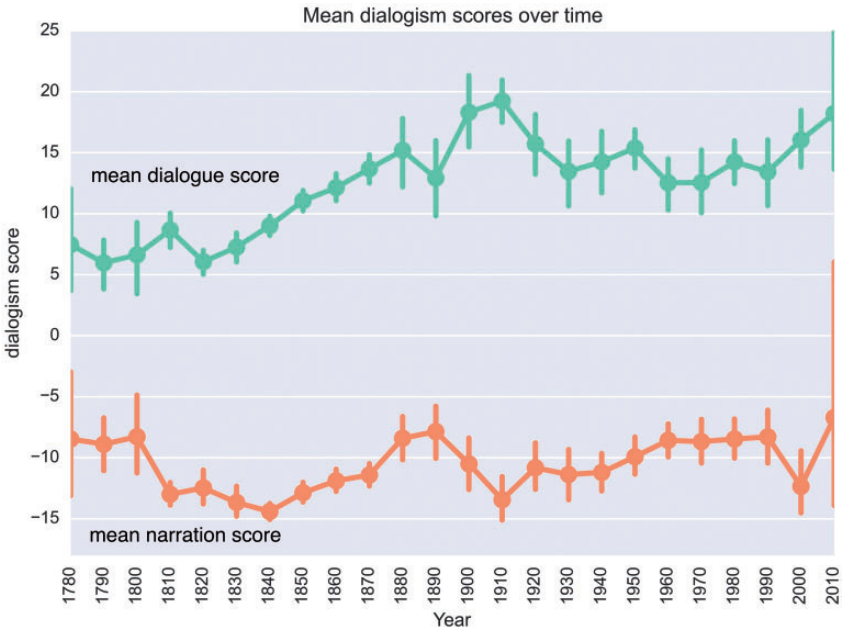

Fig. 4 Dialogism scores for narration and dialogue over time, bucketed per decade, rounded down, for every novel in our corpus. Confidence intervals are shown at 95% confidence

corpora which may exhibit different separations of dialogue and narration.

4 Narration, Dialogue, and Time

In our earlier analysis of dialogue over time, we lacked an understanding of the underlying processes that affect how dialogue driven a novel is. The dialogism metric allows us to measure this effect in terms of dialogism, investigating which processes and stylistic choices create a more dialogic experience as a whole. These choices will show us which authors participate in which literary traditions, thus revealing dialogism-based literary innovation.

Visualizing the amount of dialogism in novels over time confirms that novels are becoming more dialogic over time, driven by increasingly dialogic dialogue,[7] (Fig. 4), consistent with the trend observed in Section 2.2. Having the metric, however, leads to a more nuanced understanding: we see that this trend is dominated by stylistic changes in dialogue complemented by temporal flux in the dialogism of the narrative portions of the novels. Figure 4 shows mean dialogism scores of dialogue and narration over time. In addition, we see that while both dialogue and narration are becoming more dialogic, the rate of change is higher for dialogue, causing the mean distance between the two to grow.

Next, we compute the expected distance between dialogue and narration for each decade in our corpus; the way an author's text deviates from this expected value will form a window onto literary innovation. This model is created by performing a linear regression on the mean difference between novels smoothed by 50-year increments. This model is calculated from the trends shown in Fig. 4. The smoothing has the effect of minimizing the effects of individual authors, who may be overrepresented in a short time frame, on our overall model, resulting in a clearer focus on outlier time periods and movements. The line of best fit ($r^2 = 0.61$) is then used to calculate the mean distance from the predicted difference for each decade to produce the shading shown in Fig. 5. We find that there are specific time periods, shown in Fig. 5, when the expected difference is significantly

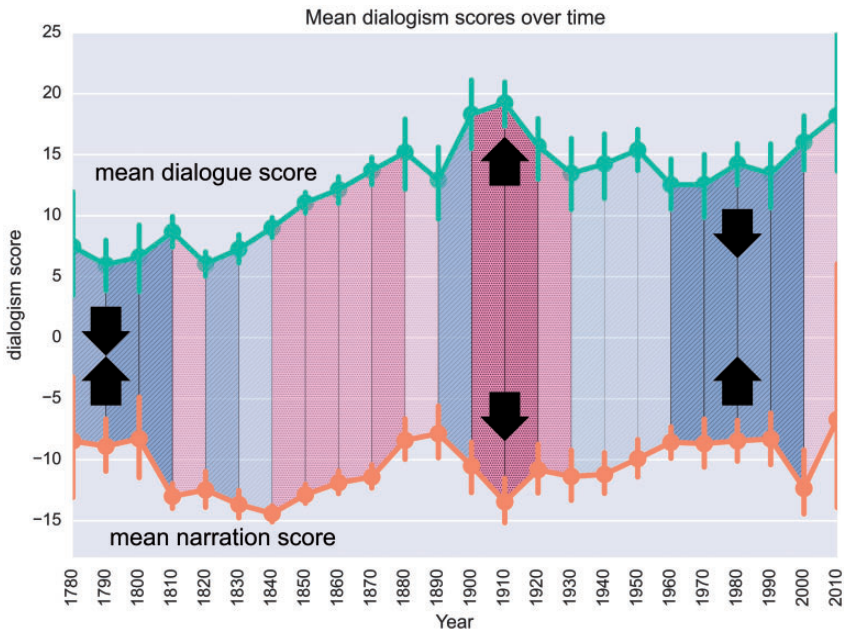

Fig. 5 Dialogism scores for narration and dialogue over time, bucketed per decade, rounded down, for every novel in our corpus. Periods where the difference in dialogism scores is unexpectedly high (shaded with circles) or low (shaded with diagonal lines) are highlighted with shading and arrows. Confidence intervals are shown at 95% confidence

Table 6 Predicted dialogism scores versus actual dialogism scores for the three target periods highlighted in Fig. 5

	Predicted difference	Mean difference	Mean narration	Mean dialogue
1782–1809	19.9	14.9	−8.4	6.3
1900–29	24.6	29.3	−7.1	17.0
1960–99	27.1	21.3	−9.1	15.3
Overall	–	22.9	−11.4	11.4

different from the actual difference between dialogue and narration dialogism scores. Figure 4 also shows the relative stability of the observed trends when 95% confidence intervals are added.

This analysis reveals three particular time periods in which narration and dialogue are closer or farther apart, as measured by dialogism scores, than is predicted for that range of years from the long-term trend. The writing in our corpus from 1782 to 1809 was dialogically more similar than at any other point in these data. From this point on, the early 20th century diverged from 1900 to 1929, and the late 20th century prominently converged from 1960 to 1999. This analysis, consistent with all the figures we have presented, bins novels by decade for two reasons: to avoid sparsity issues and to help identify specific time periods that have unexpected trends. As with any arbitrary cutoff, this methodology is imperfect and makes the time periods that we have identified appear to have more concrete beginning and end dates than likely exist.

Table 6 shows the predicted difference alongside the actual mean dialogism scores for the novels that fall in each time period. Furthermore, we calculate the number of novels that have a difference in dialogism outside of 95% of a normal distribution for each decade as a percentage of novels written in that decade. Our calculations show that on average, 0.5%

of the novels in a given decade have a difference much larger than expected and 3.1% have a difference much smaller than expected. From 1782 to 1809, we find that 7.6% of these novels have a difference much smaller than expected; from 1900 to 1929, 3.6% of the novels have a difference much larger than expected; and from 1960 to 1999, 5.7% of the novels have a difference much smaller than expected.

These specific periods seem to correspond in time to specific literary movements (Romanticism, Modernism, and Post-modernism), whose major authors are represented in our corpus. Therefore, we next identify those novels that have the largest influence during these moments and investigate what they are doing stylistically that causes such variation. This forms the basis of our investigation as to whether literary innovation is revealed by the ways that narration and dialogue interact.

4.1 1782–1809

From 1782 to 1809, the difference in dialogism between narrative and dialogic text is smaller than predicted. Since quotation marks, and therefore, direct or dialogic speech, only came into being in the mid to late 18th century, it is hard to judge if the lines are converging or if they are fluctuating during their establishment. Nevertheless, these novels are, in aggregate, blurring the linguistic line between narration and dialogue by presenting narration as more dialogic and dialogue as less dialogic to a greater degree than at any other time in our corpus.

Not all novels within this time period participate in the blurring of narration and dialogue; in the following subsections, we examine three groups of novels that show how the interplay between narration and dialogue reflects the literary traditions of the authors.

4.1.1 *Entertainment-oriented Gothic novels*

Gothic novels produced for the purposes of entertainment by and large maintain dialogic distance between dialogue and narration. This is an indication that at least some forms of entertainment-oriented writing tend toward a more standard division between dialogue and narration. The formulaic plot that these novels follow, therefore, is driven by stylistically active dialogue and descriptive narration.

A closer look at Ann Radcliffe's *The Italian, or the Confessional of the Black Penitents* shows us dialogue that is action-oriented that centers around the speaker and the addressee:

> 'The passage we are entering opens upon the cliffs, at some distance. I have run hazard enough already, and will waste no more time; so if you do not chuse to go forward, I will leave you, and you may act as you please'.

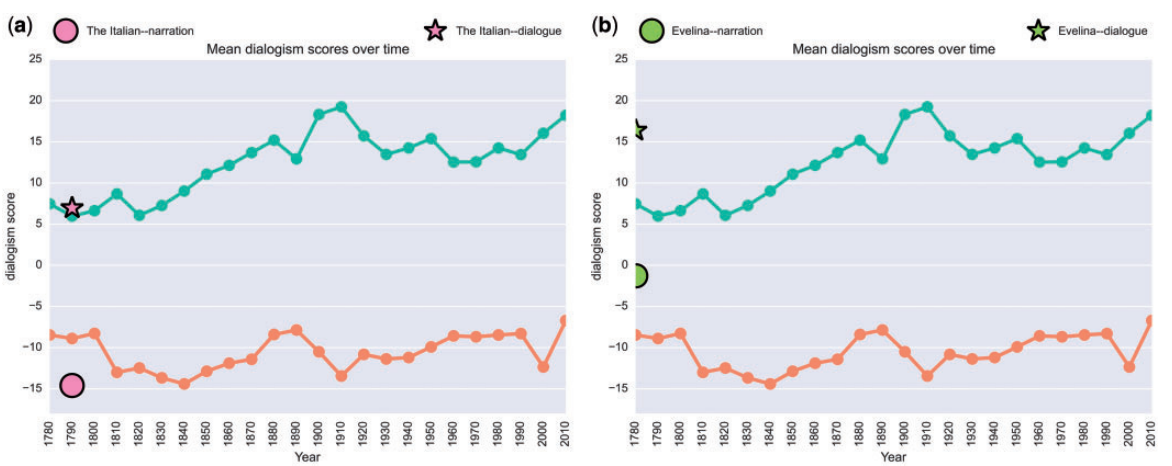

Fig. 6 Dialogism scores for narration and dialogue in Gothic novels

The narration describes concrete people and actions:

> cried Schedoni, in a voice which the strength of his spirit contending with the feebleness of his condition, rendered hollow and terrible.

Both are reflected in the dialogism scores for *The Italian*, shown in Fig. 6a. These novels are an example of form following function—dialogue is used primarily to represent interaction between characters, while the narration takes a significantly less involved point of view, describing things that have come to pass instead.

4.1.2 *Novels of manners*

Novels of manners offer a different type of entertainment-oriented novels from this era. They portray classically dialogic traits, and they describe a real and concrete world in their narration. The difference in theme between gothic novels and novels of manners is realized in word choice rather than dialogic style. Since our dialogism metric is specifically designed to be robust to differences in word choice, it instead reveals the parallel stylistic structures these groups share.

Examples from Fanny Burney's *Evelina* show dialogue that is action-oriented, but from an emotional perspective:

> 'I don't wonder that you are disconcerted, Madam, it is really very provoking. The best part of the evening will be absolutely lost. He deserves not that you should wait for him'.

while the narration remains concretely descriptive:

> 'I suppose my consciousness betrayed my artifice, for he looked at me as if incredulous; and, instead of being satisfied with my answer. . .[]'

Notice the similarity in distance between dialogue and narration in Fig. 6b, which shows the dialogism scores of *Evelina*, and in Fig. 6a, which shows those of *The Italian*. As a whole, the entertainment-oriented fiction of this era shows conservative adherence to form and function, a stylistic choice that is reflected by their predictable dialogism scores.

4.1.3 *Novels of ideas*

The novels of ideas in our corpus, such as those of Godwin, Rowson, and Brown, pattern together, with more dialogic narration and less dialogic dialogue. A qualitative analysis suggests that this particular group of novels is driven by engagement in philosophical discourse with the goal of persuading the reader. Our metric captures these phenomena stylistically (Fig. 7a–c).

For example, in *Caleb Williams* by William Godwin, the average dialogism score of dialogue is −7.9 (20.7 points less than the corpus average) and the average dialogism score of narration is −8.0 (4.8 points more than the corpus average). On closer inspection, the quotes in this novel are more narrative—they engage in storytelling and relate abstract ideas to the reader, as well as to the interlocutor, both activities that are more narrative in nature, as in the following two samples of dialogue:

> 'Count Malvesi, I feel the utmost pleasure in having thus by peaceful means disarmed your resentment, and effected your happiness. . .'

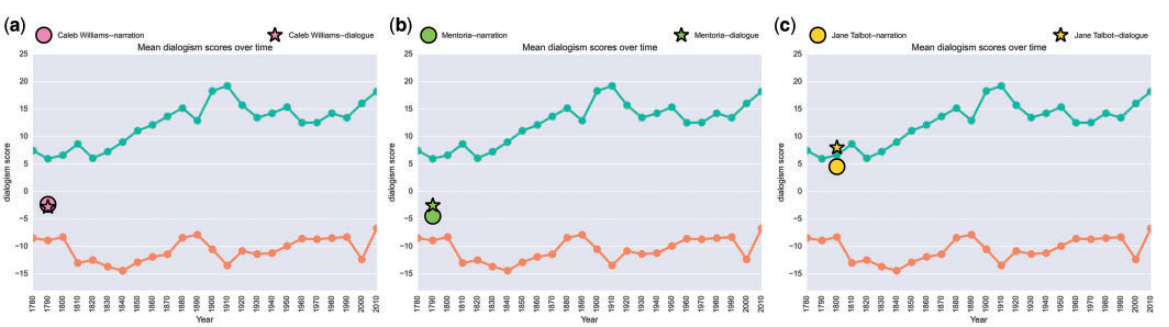

Fig. 7 Dialogism scores for narration and dialogue in novels of ideas

> 'If any one had any thing to say to him, he should know where and how to answer him. . .'

In this novel, and novels of ideas in general, the narrative portions resemble those of dialogue, such as the 1st-person narrator in the following example who is hypothesizing about future events that may come to pass:

> It will also most probably happen, while I am thus employed in collecting together the scattered incidents of my history, that I shall upon some occasions annex to appearances an explanation, which I was far from possessing at the time, and was only suggested to me through the medium of subsequent events.

This mixture of 1st-person narration and abstract thought has the effect of raising the dialogism score of the excerpt and making it appear stylistically more dialogic. This is a feature of novels of ideas as a whole, which are more likely to feature 1st-person narration, abstract ideas and to dramatically enact philosophical discourse.

Susanna Rowson's *Mentoria; Or, the Young Lady's Friend*, contains dialogue that is heavy with conditional modality, hypotheticals, and subordinate clauses, bringing down the overall dialogue score and blurring the line between dialogue and narration:

> 'if, as you think, my father has not behaved to me with the kindness of a parent, it by no means releases me from my duty to him; had he a thousand errors he is my father still; as such I am called upon by nature and religion to do every thing in my power to render his life comfortable. . .'

These examples show that the tendency to engage in philosophic exploration and to explore hypothetical situations might be what drives dialogism scores downward.

Another way that the novels of ideas reduce the gap between narration and dialogue is by writing from an extremely dialogic narrative viewpoint. In cases like these, such as William Brown's *Jane Talbot*, the narration is written from a stream of consciousness perspective. Even though dialogue remains relatively high in terms of dialogism, because of the extreme stylistic movement in the narration, the gap between the two is incredibly small.

> Conscience tells me it is folly, it is guilt to wrap up my existence in one frail mortal; to employ all my thoughts, to lavish all my affections upon one object;. . .

Since the narration contains Jane's inner monologue, it stands to reason that our dialogism metric reflects this and gives it a high score. This stream of consciousness narrative style carries the same elements of persuasion that we saw with Godwin and Rowson. This makes up for the fact that *Jane Talbot* lacks the philosophical discourse of the other novels. It also blurs the line between narration and dialogue by driving the story from a standpoint of personal persuasion, an innovative technique in this era.

4.1.4 *Dialogue and narration from 1782 to 1809*

The dialogue and narration of late 18th century and early 19th century can be characterized as reflective of the overall trends of the period. Authors like Radcliffe and Dacre engage in a more conservative style of novelistic discourse, as reflected in the predictable dialogism scores that these novels exhibit. Authors like Godwin and Rowson were using a more progressive style of novelistic discourse, using a dialogically experimental form to convey progressive ideas (Fig. 12).

4.2 1900–29

The early 1900s are a time when the dialogism scores of narrative and dialogic text diverge more than expected. Narration and dialogue grow apart, linguistically differentiating from one another. This is a time period characterized by two stylistically interesting events: the advent of modernism and the use of free indirect discourse. The dialogism scores of texts during this period suggest a new lens through which to examine the formal aspects of these literary movements. How do authors of the modernist period manipulate the dialogue and narration of the novel toward their experimental ends and to what extent is free indirect discourse complicit in the innovation of the period?

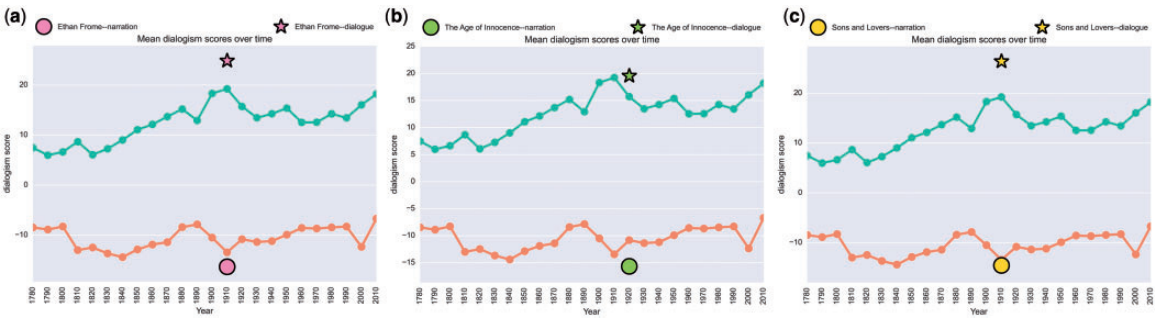

Fig. 8 Dialogism scores for narration and dialogue in modernist novels

We see notable representation of modernist novels among those that separate dialogue from narration at this time. Although the works of D. H. Lawrence and Virginia Woolf are exemplary of this trend, as we would expect from their canonical status among modernist authors, Edith Wharton's novels also belong to this group. This suggests that stylistically her work may be formally closer to the modernists than has traditionally been assumed in literary scholarship. From the high dialogism scores of dialogue, we expect to see dialogue that is highly conversational, is speaker- and audience-oriented, and mimics interlocutional interaction rather than storytelling. Similarly, from the low scores of narration, we expect narration that is descriptive and that contains a high proportion of discursively related abstract thought. Close readings of these portions of text show linguistic support for the dialogism scores that the modernists receive in practice.

4.2.1 *Modernist novels*

It is immediately apparent that the dialogue in novels such as Edith Wharton's *Ethan Frome* is structurally different from dialogue in the early 19th-century novels (Fig. 8a). The underlying stylistic effects of the dialogue in these novels are revealed by their unpredictably high dialogism scores.

> 'We can fetch it; I know we can fetch it –'

The usage of generally short utterances that feature repetition and self-interruption exemplifies a kind dialogic naturalism that we see emerging in these novels. Such features are not explicitly used when calculating dialogism. These are dialogues that reflect conversational realities. In contrast with the philosophic dialogue in novels of ideas and the action-oriented discourse in fiction centered on entertainment, these utterances portray spoken language as rife with errors and corrections. Linguistic studies of conversational speech find that strategies like repetition serve to ease production, boost comprehension, and support the overall dynamics of the interaction (Tannen, 2007). The following exchange from Wharton's *The Age of Innocence* shows just this.

> 'Well, then –?'
> 'Well, then; she bolted with his secretary'.

The second notable trait of this kind of dialogue is that it concentrates on psychological processes. In these novels, the internal conflicts of the characters are foisted onto the dialogic stage rather than implicitly existing in the background of the text. This is to say, the reader must explicitly acknowledge the internal conflicts that the characters undergo when they resolve their own internal conflicts through dialogue. The overall effect of using more realistic utterances that are oriented toward psychological processes is that dialogue shifts from performing a purely literary function to performing a psychological one.

This example of narration from *The Age of Innocence* (Fig. 8b) shows that while dialogue is becoming more realistic, narration is becoming more literary:

> Now, by some queer process of association, that golden light became for him the pervading illumination in which she lived.

With subordinate clauses often fronted, the complex syntactic structure of these sentences drives dialogism scores down.

D. H. Lawrence's *Sons and Lovers* (Fig. 8c) shows us that the drive toward dialogic realism is a stylistic innovation central to the modernist literary movement. Once again, we see dialogue that features short sentences heavy with repetition that pay special attention to the speaker's psychological state.

> 'What have I to do with all this? Even the child I am going to have! It doesn't seem as if I were taken into account'.

The following examples show how Lawrence embeds long descriptive subordinate clauses into the middle of his narrative sentences. Though *Sons and Lovers* does not front subordinate clauses, they are still notable in their complexity and the frequency with which they are used.

> His dark brown hair and fresh coloring, and his exquisite dark blue eyes shaded with long lashes, together with his generous manner and fiery temper, made him a favorite.

These novels simultaneously exhibit realistic dialogue and linguistically complex narration, creating a stylistic gap between these different textual elements.

4.2.2 *Free indirect discourse*

Free indirect discourse represents speech using lexically similar but grammatically different markers of speech than does direct dialogue (Banfield, 1982). Because our model of dialogism relies on abstract grammatical categories rather than lexical content to determine degrees of dialogism, novels heavy in free indirect discourse (like Virginia Woolf's *To the Lighthouse* and *Mrs. Dalloway*, shown in Fig. 9a) receive low dialogism scores for their narration (Fig. 9b).

Our discussion of free indirect discourse focuses on Woolf, and the modernist period overall, as the technique is most visible, at the level of the period, among the authors of the early 20th century. While Jane Austen's narrative dialogism scores are similar to Woolf's, indicating the presence of free indirect discourse in her novels, as an early adopter of the method, her relatively modern narrative technique makes her an outlier during the Romantic period. We leave a historical analysis of the interaction of dialogism with free indirect discourse for future work.

Through the lens of dialogism, in other words on a syntactic level, free indirect discourse operates as a narrative distillation of dialogue. Dialogism reveals that free indirect discourse is in fact less dialogic than traditional narration. This is in line with Vološinov's position that indirect discourse loses

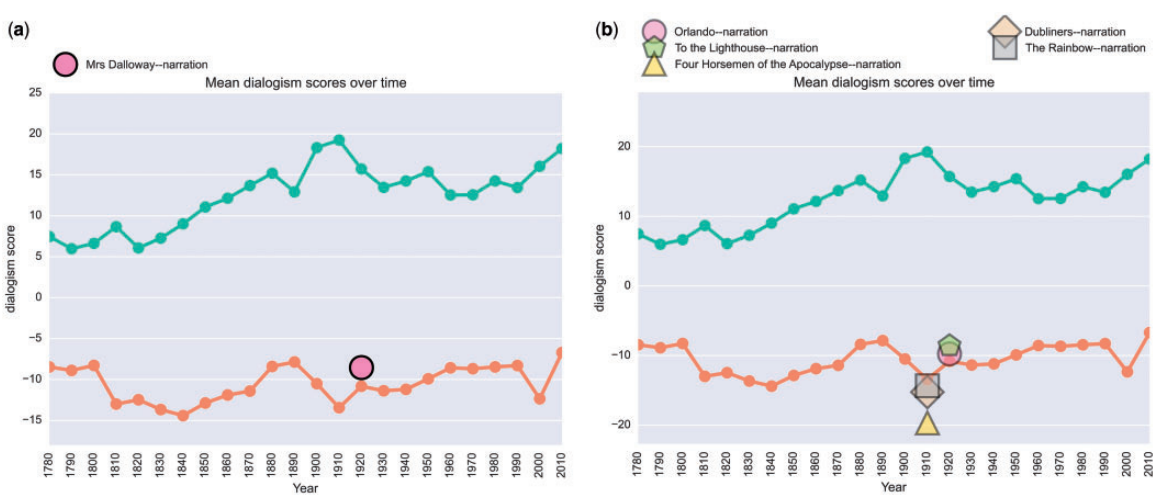

Fig. 9 Dialogism scores for narration in novels with free indirect discourse

its emotive–affective form while maintaining the underlying content (Vološinov, 1986).

We examine free indirect discourse in the novels of Virginia Woolf to see the linguistic causes of the low dialogism scores given to free indirect discourse. Sentences with free indirect discourse undergo a kind of 'grammatical translation'. In the examples below, 1st-person pronouns are transposed to 3rd-person pronouns, verbs tend to become more passive, and the boundary between quote and narration disappears. There are no stylistic markings between the free indirect discourse portions of the sentences and the quotatives ('said she', 'thought Clarissa Dalloway') with which they are associated. The following examples are highlighted to show that free indirect discourse contains mostly words that decrease the dialogism score (shown in italics):

> Mrs. Dalloway *said she* **would buy** *the flowers herself*.
> *And* then, *thought* Clarissa Dalloway, what *a morning—fresh as if issued to children on a beach*.
> Her *only gift was knowing people* almost *by instinct*, *she thought*, *walking on*.

Since in our model dialogue is characterized by 1st- and 2nd-person pronouns, present tense verbs, and modals, exactly the elements that are systematically translated in free indirect discourse, these sentences receive dialogism scores more typical of narrative. Together, the systematic translation of these sentences to a more narrative style and the removal of the dialogue/narration boundary cancel out the dialogic effects of the underlying dialogue that is being represented. This analysis shows that along with genre and narrative point of view, literary techniques like the usage of free indirect discourse affect dialogism in general.

Free indirect discourse is thus a literary style that, rather than creating more dialogic narration, creates narrative dialogue. From this perspective, the innovation that authors like Woolf are engaging in is the removal of linguistically dialogic elements from lexically dialogic text.

4.2.3 *Dialogue and narration from 1900 to 1929*

The early 20th century saw the rise of modernism. This literary style carries important dialogic markers that are revealed through careful analysis. In the case of the modernists, we see a movement toward dialogic naturalism, characterized by the explicit communication of psychological processes in dialogue. In parallel and closely intertwined, our attention is drawn to the use of free indirect discourse, which demonstrates the process for distilling dialogue to its narrative elements. Authors such as Virginia Woolf use this technique to embed dialogue in narration. This in turn preserves the typically less dialogic elements of the narration. These innovative influences are summarized in Fig. 12. The stylistic innovation that occurs during this time period is once again revealed via dialogic interaction and careful inspection of specific literary movements.

4.3 1960–99

Our final target period spans post-war literature and is roughly contemporary with the post-modernist novel. This period is, like the period from 1782 to 1809, a time when the difference between narration and dialogue is smaller than predicted. Once again two stylistically interesting groups of novels rise out of this era: those novels that become more dialogic as a whole, and those in which the relative dialogism of dialogue drops while that of narration rises. Even though the overall pattern of convergence is similar to that from 1782 to 1809, the underlying causes are different because the literary and historical context has shifted dramatically. Since literary innovation is couched in the traditions that it pushes back against, we expect innovation from this time period to exhibit different traits than those found in either period we have already examined.

4.3.1 *Conversation fiction*

Our analysis reveals a new group of novels that we call 'Conversation Fiction'. This group contains novels like *Portnoy's Complaint*, *Waiting for the Barbarians*, *Ender's Game*, *The Handmaid's Tale*, and *Mission Earth*. This group of novels seems at first glance to be completely unrelated. However, they all rely on highly dialogic narration and highly dialogic dialogue, giving them a common, conversational, stylistic underpinning.

In Philip Roth's *Portnoy's Complaint* (Fig. 10a), dialogue is very informal and conversational in

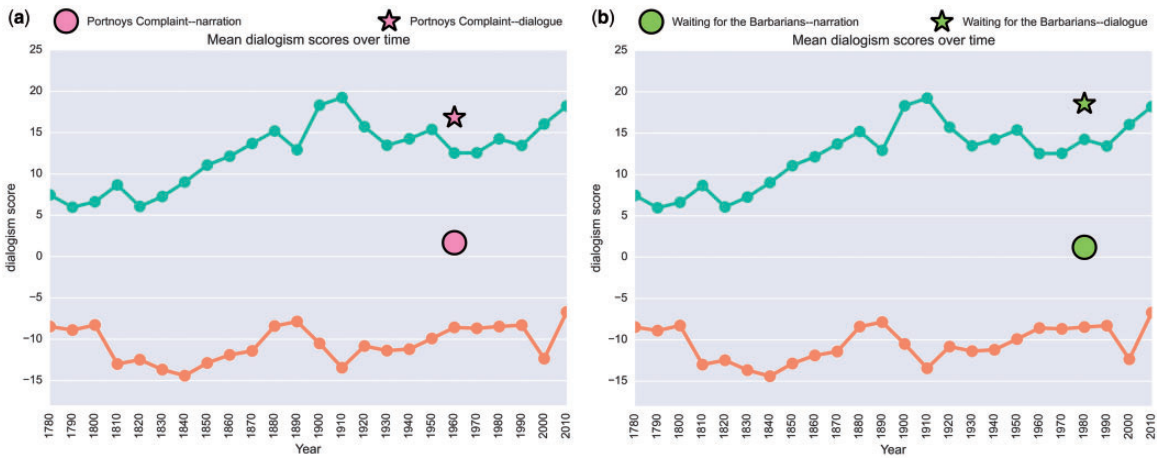

Fig. 10 Dialogism scores for narration and dialogue in conversation fiction

nature, leading to elevated dialogism scores for dialogue overall.

> 'Loving? You? Too-ey on you! Self-loving, boychick, that 's how I spell it! With a capital self! . . .'

At the same time, narrative text is also driven to higher dialogism scores. One contributing factor is the use of short, exclamatory narrative sentences that act like interjections:

> Who cares!
> There!

Even when they appear in more typical narrative form, Roth's sentences defy traditional narrative style by using the present tense:

> Tacked above the Girardi sink is a picture of Jesus Christ floating up to Heaven in a pink nightgown.

J. M. Coetzee's *Waiting for the Barbarians* (Fig. 10b) shows elevated dialogism scores for both dialogue and narration, reflective of a similarly active style:

> I sweat and strain, there is a stab of pain in my back, but the bar does not budge.
>
> 'We hope for three thousand bushels from the communal land this year. We plant only once. The weather has been very kind to us'.

Since both 1st-person perspectives and the use of the present tense drive dialogism scores up, the narrative passages maintain high scores. As we see below, this stands in contrast to the novels of the contemporaneous post-modernists.

More generally, dialogism reveals that novels are becoming more dialogic, not just in dialogue, but in narration as well, particularly at this time.[8] However, even though these novels feature dialogic narration, they maintain greater linguistic distance between narration and dialogue, than, for instance, novels of ideas. They move together primarily because the dialogic change in narration is so high, not because dialogue becomes more narrative.

4.3.2 *Non-conversational fiction*

What we characterize as non-conversational fiction in our corpus has unpredictably low dialogism scores associated with its dialogue, suggesting that dialogue in these novels shares stylistic traits with narrative texts. Examples of non-conversational fiction like David Foster Wallace's *Infinite Jest* and the novels of William S. Burroughs, shown in Fig. 11a and b, once again blur the line between narration and dialogue.

The blurring between narration and dialogue at this time period is rooted in post-modernism. The dialogue has a more stilted, formal tone, with sentences rife with subordinate clauses and taking a

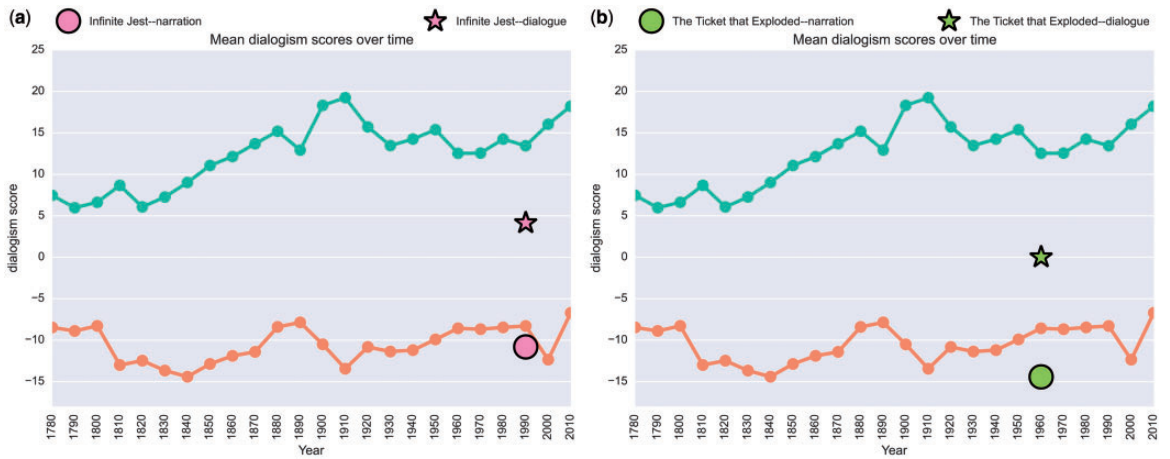

Fig. 11 Dialogism scores for narration and dialogue in non-conversational fiction

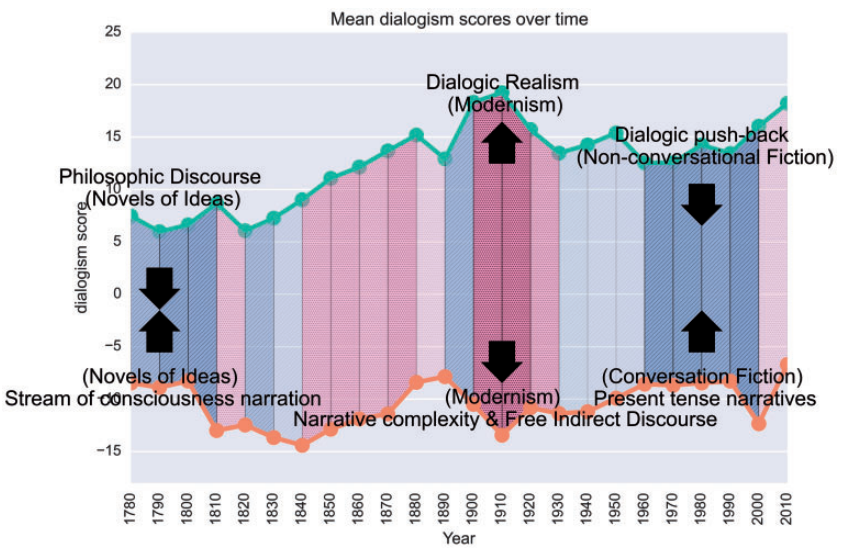

Fig. 12 Dialogism scores for narration and dialogue over time, bucketed per decade, rounded down, with target periods highlighted. Innovations revealed in our analyses and their effects on dialogism are identified and summarized by their corresponding arrows

more descriptive perspective. The dialogue in *Infinite Jest* has the tone of a narrator speaking rather than realistic conversational utterances, although remaining more dialogic than the narrative portions of the text:

> 'Just a bit of a let's call it maybe a facial tic, slightly, at all the adrenaline of being here on your impressive campus, justifying his seed so far without dropping a set, receiving that official written offer of not only waivers but a living allowance from Coach White here, on Pac 10 letterhead, being ready in all probability to sign a National Letter of Intent right here and now this very day, he 's indicated to me'.

The dialogue pushes back against the dialogic realism of the modernists. One of the ways that this

push back manifests itself is through a certain type of alienation that is typically associated with post-modernism. The way that they use dialogue contributes to the overall effect of alienation because they remove an element of conversational relatability from the discourse. This in turn reveals a stylistic technique that this kind of literary fiction uses to separate itself from the traditions that preceded it, especially that of modernism.

Novels like Burroughs' *The Ticket that Exploded* show another way that literary fiction engaged in dialogic push back. Rather than being extremely narrative, as in the examples from *Infinite Jest*, these examples show disjointed and chaotic speech that seems to be fragmented narrative with dialogic interruptions.

> "This is a novel presented in a series of oblique references . . shave? . . did he? . . an amputation . . three young burglars one wearing a black overcoat stopped on the stairs by two English detectives . . One of the thieves is nicknamed Genial . . "

These examples use strikingly disconnected phrases that are anchored by noun phrases, short descriptions, and sentences lacking a clear subject. Even when the dialogue exhibits more dialogic traits, the use of run-on sentences, which are indirectly related to POS tag distribution, in which the speaker describes tangential stories has the overall effect of depressing the dialogism score. This style of dialogic pushback, though very different than the extremely narrative dialogue in *Infinite Jest*, is another technique that breaks from styles that use more realistic dialogue. In this case, Burroughs uses the chaotic state of his characters to produce dialogue that is not correspondent with familiar states of mind.

4.3.3 *Dialogue and narration from 1960 to 1999*

The later 20th century can be characterized by an overall shift toward a more dialogic style. Conversation fiction in particular exhibits highly dialogic features. These stories are carried by both dialogue that is conversationally realistic and narration that uses a dialogic writing style to diminish the gap between narration and dialogue. Together, these traits work to establish a new, stylistically linked, type of fiction. Standing in contrast to these stories are those of non-conversational fiction, in which dialogue becomes less dialogic. These influences are summarized in Fig. 12. The analysis of these novels in the context of the Modernist traditions that these works formulate themselves against allows us to analyze this stylistic decision as a specific dialogic pushback against the dialogic realism of the modernists. Non-conversational fiction uses unrealistic dialogue, be it unrealistically formal and narrative or unrealistically fragmented and chaotic, to foment a grammatical shift away from the modernists. In this time period, the relationship between dialogue and narration shows us the evolving standard of novelistic discourse in general fiction as well as one way that the post-modernists engaged in stylistic rejection of modernist methods.

5 Conclusion

The goals of our work are two-fold: to explore dialogue using a rigorous and insightful quantitative metric, and to see the implications of this metric historically at key moments of the novel's evolution.

Our first contribution toward this goal is the analysis of a new, large, corpus of extracted dialogue and the release of new tools and algorithms for extracting quotes. From this corpus, we fill out the picture of dialogue in the novel from the 18th through the 20th centuries, finding that novels have become more dialogue driven both in terms of quote density and in terms of dialogic text.

Our second contribution is a new, transferable metric of dialogism that measures the extent to which any span of text exhibits the grammatical features characteristic of spoken dialogue. We use this metric to evaluate dialogism over time, revealing that novels have become more grammatically dialogic over the past 220 years.

We then observe three distinct moments when certain groups of authors reject the dialogic expectations of their era, showing us how literary innovation is embedded both in dialogue and in the relationship between dialogue and narration in the novel.

Novels of ideas, like *Caleb Williams*, worked not only in the ideas that they promoted but also in the

way they were communicated. These authors use dialogue that relies on a style of pontification and persuasion that is couched in a more narrative style. At the same time, the narration in these works is more dialogic, written from a more personal perspective.

When modernists like D. H. Lawrence experiment with the genre, we find direct grammatical indications that this experimentation emphasized dialogic realism and that, surprisingly, Edith Wharton shares more dialogic traits with these authors than we might expect given her distance from the core modernists in literary criticism. This realism reflected a new authorial perspective of the relationship between dialogue and psychological processes. The modernists contrasted this realist dialogue with narration that was thick with complex grammatical structures and subordinate clauses, helping to amplify the realist effect of the dialogue. Furthermore, we find that free indirect discourse is a narrative distillation of dialogue when viewed through the lens of higher-level abstract grammatical features.

Finally, in the late 20th century, we uncover a new grouping, conversation fiction, defined by works such as those of Coetzee, Roth, and Atwood. These novels work by maintaining the grammatical distance between dialogue and narration while increasing overall dialogism. At the same time, the contemporary non-conversational fiction of Wallace and Burroughs works by making dialogue more narrative and pushing back against the dialogic realism of the modernists. Together, these moments of innovation show that different kinds of literary innovation reveal themselves through the essential and complex interaction of dialogue and narration.

The methodological implications of our work demonstrate the usefulness of grammatically sensitive (but lexically blind) stylistic analysis to Digital Humanities research, when combined with close reading. Furthermore, our work shows that dialogue is an active participant in the stylistic landscape of the novel. A literal implementation of Bakhtin's framing of dialogue offers insight into literary innovation because dialogic context tracks stylistic innovation and methods of meaning making in the novel. The result is a more nuanced understanding both of how novels work, and of how innovation works, rooted in the interplay between dialogue and narration.

Funding

This work was supported by the National Science Foundation (grant number DGE-114747). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. NSF graduate fellowship to the first author.

References

Algee-Hewitt, M. and McGurl, M. (2015). Between canon and corpus: six perspectives on 20th-century novels. In *Pamphlets of the Literary Lab 8*.

Allison, S., Genmam, M., Heuser, R., Moretti, F., Tevel, A., and Yamboliev, I. (2013). Style at the scale of the sentence. In *Pamphlets of the Literary Lab 5*.

Bakhtin, M. M. (1935). Discourse in the novel. In *The Novel: An Anthology of Criticism and Theory 1900–2000*, pp. 481–510.

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

Banfield, A. (1982). Unspeakable sentences. In *Narration and Representation in the Language of Fiction*. Boston and London: Routledge & Kegan Paul.

Biber, D. (1991). *Variation Across Speech and Writing*. Cambridge, UK: Cambridge University Press.

Gemma, M., Glorieux, F. and Ganascia, J. -G. (2015). Operationalizing the colloquial style: repetition in 19th-century American fiction. *Digital Scholarship in the Humanities*, **32**, 312–335.

Genette, G. (1983). *Narrative Discourse: An Essay in Method*. Ithaca, New York, USA: Cornell University Press.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J. and Smith, N. A. (2011). Part-of-speech tagging for twitter: Annotation, features, and experiments. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers, vol. 2. Association for Computational Linguistics, pp. 42–7.

Hollingsworth, B. and Teufel, S. (2005). Human annotation of lexical chains: Coverage and agreement measures. In *ELECTRA Workshop on Methodologies and Evaluation of Lexical Cohesion Techniques in Real-world Applications (Beyond Bag of* Words), p. 26.

Hoover, D. (2005). Word frequency, statistical stylistics, and authorship attribution. In *Advanced ICT Methods Guide to Linguistics*. AHRC ICT Methods Network, Centre for Computing in the Humanities, Kay House, 7 Arundel Street, London, WC2R 3DX.

Jørgensen, A., Hovy, D. and Søgaard, A. (2016). Learning a POS tagger for AAVE-like language. In *Proceedings of NAACL-HLT*. San Diego, CA, USA, pp. 1115–20.

Tannen, D. (1987). Repetition in conversation: toward a poetics of talk. *Language*, 574–605. Washington, DC, USA: Linguistic Society of America.

Tannen, D. (2007). *Talking Voices: Repetition, Dialogue, and Imagery in Conversational Discourse*, vol. **26**. Cambridge, UK: Cambridge University Press.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, vol. 1. Association for Computational Linguistics. Edmonton, Alberta, Canada, pp. 173–80.

Vološinov, V. N. (1986). *Marxism and the Philosophy of Language*. Cambridge, MA, USA: Harvard University Press.

Notes

1 These lists are (1) Modern Library Board's List of 100 Best Novels of the 20th Century, (2) Modern Library Reader's List of 100 Best Novels of the 20th Century, (3) Radcliffe's Rival List of the 100 Best Novels of the 20th Century, (4) Larry McCaffery's List of the 100 Best Novels of the 20th Century, and (5) The yearly best-selling works of the 20th Century. A complete description can be found in the Literary Lab's pamphlet #8 (Algee-Hewitt and McGurl, 2015). A full list of titles and derivative data are available at https://nlp.stanford.edu/∼muzny/dialogism.html.

2 For example, bookNLP (compared against in Table 3) has problems with these cases because it relies on the underlying tokenizer to correctly identify forward 'and backward quotes' (Bamman *et al.*, 2014). Once this system detects a forward quote, 'it assumes that it is in a quote until it finds backward quotes'. If bookNLP encounters a forward quote when it is already in a quote, it resets the beginning of the quote to the most recent set of forward quotations. We believe these are the reasons that our system outperforms bookNLP, especially for *Emmeline the Orphan* and *The Spy*.

3 Released as part of the Stanford CoreNLP Java toolkit, documentation available at https://stanfordnlp.github.io/CoreNLP/quote.html.

4 While this tagger achieves close to 97% accuracy on *Wall Street Journal* data (Toutanova *et al.*, 2003), Gimpel *et al.* (2011) show that it achieves an accuracy of 85.85% on Twitter data, while Jørgensen *et al.* (2016) show that it achieves 74.3% accuracy on a variety of texts that use African American Vernacular English. While we do not expect 97% accuracy from the POS tagger on our data set, we hypothesize that it will remain relatively high for novels with standard sentence structure and that it will remain acceptably high even for novels with dialect-heavy dialogue. The study of the effects of dialect on dialogue is complex (Gemma *et al.*, 2015) and deserves more attention than the scope of this article allows.

5 In practice, the factors that we isolated contain only groups of features that are positively correlated with one another (though this is not a requirement of MDA).

6 Although MDA can produce multiple factors that are influenced by the same feature (e.g. modal verbs could contribute to both factors 1 and 2), we follow Biber's procedure which only include each feature once, in the factor for which it has a larger weight produced by MDA.

7 Calculating a line of best fit for the extracted dialogue shows a slope of 0.041 at an r^2 value of 0.86.

8 While not included in this close reading, novels such the genre and popular fiction of this era follow similar patterns.