



Corpus Stylistics and the Modernist Novel: Achievements and Limits of Quantitative Close Reading

Diego Kobayashi^{1*}, Anahit Morin², Shalini Kozlowski³

¹Department of Social Sciences, Pampas Institute of Social Sciences, Iceland

²Department of Social Sciences, Chao Phraya Academy of Social Sciences, Jamaica

³Department of Social Sciences, Ebro Institute for Advanced Social Sciences, Finland

Article Info

Volume 1, Issue 1, 2026

doi: [10.62715/ijesl.2026.v1.i1.pp7-12](https://doi.org/10.62715/ijesl.2026.v1.i1.pp7-12)

Received: 03 May 2026

Accepted: 12 July 2026

Published: 18 September 2026

Abstract

Corpus stylistics applies concordancing and simple statistics to literary texts in order to test and refine claims about style. It developed largely around Dickens and the Victorian novel, and its encounters with modernist fiction have been fewer and more revealing, because modernist prose unsettles the units that the method counts. This narrative review surveys published corpus-stylistic work on the modernist novel in English, from studies of Conrad, Henry James, Woolf and Joyce to the methodological debate that surrounds them. It identifies four achievements: giving long-standing critical impressions a checkable lexical basis, showing whether characters are linguistically distinct, bringing the presentation of consciousness within reach of systematic description, and testing intuitions about connotation against general usage. It then examines three limits: the dependence of counts on editions, tokenisation and tagging that modernist texts resist; the relativity of keyness to a chosen reference corpus; and the unforced step from pattern to interpretation. The relation of the approach to distant reading and to recent critiques of computational literary studies is discussed. The review concludes that quantitative close reading disciplines critical evidence without replacing judgement, and it proposes shared annotation, sensitivity reporting and closer ties with textual scholarship as priorities.

Keywords: *Corpus stylistics, Modernist fiction, Keyword analysis, Free indirect discourse, Stylometry, Distant reading*

© 2026 Diego Kobayashi et al. This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

1. Introduction

Modernist fiction made style its subject. The shifting registers of *Ulysses*, the long interior sentences of *Mrs Dalloway* and the deliberate indistinctness of Conrad's narrators have drawn close reading for a century, and much of what is said about them takes the form of claims about frequency: that a word recurs, that a construction is typical, that one character speaks unlike another. Such claims are quantitative in logic even when no number is given. Corpus stylistics is the branch of stylistics that takes them at their word and checks them against the whole text with concordance software and simple statistics.

* Corresponding author: Diego Kobayashi, Department of Social Sciences, Pampas Institute of Social Sciences, Iceland. E-mail: diego.kobayashi@pampas.edu

© 2026 Diego Kobayashi et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The approach grew up around the Victorian novel, above all Dickens, whose length and repetitiveness reward counting (Mahlberg, 2013). Its dealings with modernism have been more scattered and, in some respects, more instructive, because modernist prose strains the assumptions on which the tools rest. A method that treats the text as a sequence of words separated by spaces meets, in Joyce or late James, writing designed to unsettle the word, the sentence and the stable narrating voice.

This review surveys published corpus-stylistic work on the modernist novel in English and the methodological debate around it. It sets out what the approach has demonstrably contributed, where its limits lie, and how it relates to the larger-scale computational criticism known as distant reading.

2. Scope and Approach

The article is a narrative review of the published literature. It draws on monographs and articles from stylistics, corpus linguistics and digital literary studies that are regularly cited in discussions of the subject, and it makes no claim to have searched the field exhaustively or by a formal protocol. Modernist is used in its ordinary literary-historical sense for experimental fiction in English from roughly the turn of the twentieth century to the Second World War, with Conrad and the later Henry James admitted as threshold figures. Authorship attribution, a neighbouring field with its own extensive literature, is touched on only where its techniques have been turned to interpretive ends.

3. Origins and Toolkit

Corpus stylistics joins two traditions. From stylistics it inherits the conviction, given its standard form by Leech and Short (2007), that literary effects can be traced to describable linguistic choices. From the corpus linguistics associated with Sinclair it takes the practice of reading many instances of a word in context and the concept of collocation, the tendency of words to keep habitual company (Sinclair, 1991). A separate computational line runs from Burrows (1987), who showed that the commonest words in Jane Austen's novels, the articles, pronouns and prepositions that readers scarcely notice, are distributed differently in the speech of different characters.

The working toolkit is small. A concordance lists every occurrence of a search word with its surroundings. A keyword list ranks the words that are unusually frequent in a text when it is compared with a reference corpus, commonly by means of the log-likelihood statistic that Dunning (1993) introduced to language study. Cluster analysis retrieves repeated sequences of words. Semantic tagging groups vocabulary into fields of meaning so that key domains, and not only key words, can be found (Rayson, 2008). Stylometric measures such as Burrows's Delta compare texts on the frequencies of very common words (Burrows, 2002). Table 1 summarises these techniques, the uses to which each has been put on modernist prose, and the chief limitation of each. McIntyre and Walker (2019) give a full account of the procedures.

4. What the Approach Has Shown

Four kinds of contribution can be distinguished in the published studies. None overturns the critical tradition; each makes a familiar claim more exact or shows where it needs qualifying.

4.1. Checking critical commonplaces

The best-known study is Stubbs's analysis of *Heart of Darkness* (Stubbs, 2005). Critics since Leavis had remarked on Conrad's insistence on the inexpressible and the indistinct. Stubbs showed that this impression has a lexical basis that can be listed: a high incidence of vague nouns and pronouns, of words for darkness and fog, of hedging phrases, and of recurrent multi-word sequences that the reader registers without being able to cite. He claimed two things for the method. It documents systematically what careful readers have noticed in part, and it draws attention to patterns, especially in grammatical words and repeated phrases, that are too diffuse to be noticed unaided. Hoover's work on Henry James makes a comparable point over a career. Stylometric comparison of the novels separates early from late James and indicates that the change of manner critics describe was progressive and not abrupt (Hoover, 2007).

Table 1: Corpus-stylistic techniques, their application to modernist prose and their principal limits

Technique	What it identifies	Use on modernist prose	Principal limit
Frequency and dispersion	Recurrent items and where they fall in the text	Conrad's vocabulary of vagueness (Stubbs, 2005)	Counts forms, not senses
Keyword analysis	Words unusually frequent against a reference corpus	Contrasting the speakers of <i>The Waves</i> (Balossi, 2014)	Result depends on the reference corpus
Key semantic domains	Over-represented fields of meaning (Rayson, 2008)	Thematic profiles of characters	Tagger built for standard usage
Clusters	Repeated word sequences (Mahlberg, 2013)	Refrains and formulae in narration	Misses variation inside a phrase
Collocation and semantic prosody	The habitual company a word keeps (Louw, 1993)	Weighing connotation in Joyce (O'Halloran, 2007)	Norms may come from another period
Function-word stylometry	Signal carried by very common words (Burrows, 2002)	Early and late James (Hoover, 2007)	Detects difference without explaining it
Discourse presentation annotation	Categories of speech and thought (Semino and Short, 2004)	Free indirect discourse in Woolf	Manual, slow and partly subjective

4.2. Distinguishing voices

A second group of studies asks whether characters are linguistically distinct. The question is pointed for Woolf's *The Waves*, whose six speakers deliver formal soliloquies in a uniformly elevated register and have often been said to share one voice. Balossi (2014) compared the six sets of soliloquies for key parts of speech and key semantic domains and concluded that each speaker has a measurable profile consistent with the personality the novel assigns. Ramsay (2011) ran a different calculation on the same novel, a weighting that brings out the words most particular to each speaker, and drew a different lesson. For Ramsay the lists are not evidence that settles the critical debate. They are a transformed version of the text that a critic reads and argues from, much as critics have always rearranged and excerpted.

4.3. Representing consciousness

Modernism's signature technique, the rendering of thought through free indirect discourse and interior monologue, has been harder to capture. The grammar of the form was described by Banfield (1982), and its classic critical account is Auerbach's reading of a passage from *To the Lighthouse*, which stressed the multiplicity of consciousnesses through which the narration passes (Auerbach, 1953). Free indirect discourse has few reliable surface markers, so it cannot simply be searched for. Semino and Short (2004) met the difficulty by hand, annotating a corpus of twentieth-century narrative for categories of speech, writing and thought presentation and so providing a tested scheme and comparative frequencies across genres. Brooke et al. (2017) took a computational route, using automatically derived measures of lexical style to ask how far free indirect passages in Woolf and Joyce differ from narration and from the direct speech of the characters concerned. Work of this kind turns a celebrated critical description into a claim that can be examined passage by passage.

4.4. Weighing connotation

A fourth contribution concerns the single word. Louw (1993) argued that a large reference corpus reveals the semantic prosody of an expression, the favourable or unfavourable colouring it acquires from its usual collocates, and that departures from that prosody in a literary text signal irony or some other marked effect. O'Halloran (2007) applied related reasoning to Joyce's story 'Eveline', using corpus evidence about ordinary usage to support an inference about what the text implies concerning its protagonist's state of mind. The procedure offers an answer to the charge of arbitrariness: the critic's intuition about a phrase is

checked against a large body of uses by other writers.

5. Limits

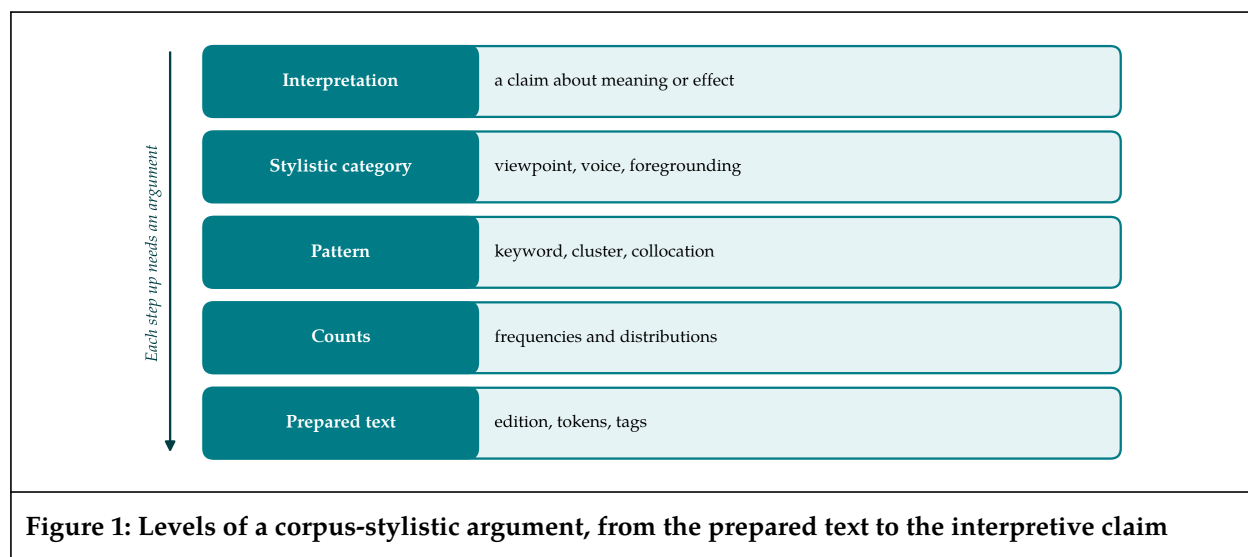
The limits of the approach are best seen by following the path from text to interpretation that every corpus-stylistic argument must travel. Figure 1 sets out the stages. At each one a decision is taken on which the final reading depends, and modernist texts make several of those decisions unusually difficult.

5.1. Editions, tokens and tags

Counting presupposes a stable text divided into words. For much of the period in which the methods matured, the major modernists were in copyright, which constrained the making and sharing of reliable digital editions; the works of Joyce and Woolf entered the public domain in the United Kingdom only at the start of 2012. Editorial history adds its own complications, since *Ulysses* exists in competing editions that differ in many readings. The deeper obstacle is linguistic. Compounds, coinages, unpunctuated monologue, foreign phrases and deliberate misspelling confound tokenisers, and part-of-speech and semantic taggers built for standard modern English are less dependable on such prose than on Dickens. A novel like *Ulysses*, whose episodes are written in sharply different styles, also undermines the assumption that a whole-text frequency describes anything at all.

5.2. The reference corpus

Keyness is a relation, not a property. A word is key only against the corpus chosen for comparison, and the choice between a corpus of contemporary fiction, the author's other works or general English of a later date changes the list. Kilgarriff (2005) made the underlying statistical point forcefully: language is not produced at random, so with enough data a significance test will declare almost any difference between two corpora to be real. Significance in this setting is therefore a ranking device and not a finding. The same caution applies to semantic prosody when the reference corpus comes from a different period from the text, since the connotations of words drift.



5.3. From pattern to meaning

The oldest objection is Fish's. Writing about an earlier generation of stylisticians, he argued that the step from a formal description to an interpretation is always arbitrary, because the description does not constrain what it is taken to mean, or else circular, because the features were selected in the light of a reading already held (Fish, 1980). van Peer (1989) added that turning a text into countable units discards much of the sequencing and context on which literary effects depend. Corpus stylisticians reply that the method makes the selection of evidence explicit and repeatable, which unaided close reading does not. The reply is fair, but it concedes the main point: the numbers discipline an interpretation without

producing one. Biber (2011) observes that the quantitative study of literary style is considerably older than the corpus-stylistic label, which suggests that the difficulty is not one that newer software will remove.

6. Corpus Stylistics and Distant Reading

Corpus stylistics is sometimes grouped with the distant reading proposed by Moretti (2013), but the two differ in scale and aim. Distant reading studies hundreds or thousands of books in order to describe literary history, and its recent practitioners use predictive models to trace slow changes in genre and characterisation (Underwood, 2019). Corpus stylistics stays with the individual work and keeps the reference corpus in the background. The distinction matters because the most searching recent critique of computational criticism, by Da (2019), is aimed at the larger enterprise. Da argued that its results tend to be either statistically sound and already known or novel and statistically fragile. Small-scale stylistic work is less exposed to the second charge, since its statistics are simple, and more exposed to the first. Its defence has to be that making the known exact, and locating its exceptions, is a legitimate scholarly gain.

7. Gaps and Future Work

Coverage is uneven. Conrad, Woolf and Joyce account for most of the published studies, and within their work the shorter and more regular texts are favoured. Writers such as Dorothy Richardson, Jean Rhys, Ford Madox Ford and Djuna Barnes appear far less often in this literature, and comparative work across authors is uncommon.

Three developments would strengthen the field. The first is shared, documented annotation of discourse presentation in modernist texts, so that findings about free indirect discourse can be compared and automatic methods evaluated against agreed examples. The second is routine reporting of how results change when the reference corpus, the edition or the tokenisation is varied; a keyword that survives such changes carries more weight than one that does not. The third is closer cooperation with textual scholarship, since the genetic and editorial record of modernist novels is unusually rich and a word count is only as good as the text it is taken from.

8. Conclusion

Applied to the modernist novel, corpus stylistics has achieved something modest and real. It has shown that long-standing impressions of Conrad's vagueness, James's late manner and the voices of *The Waves* correspond to patterns that can be listed and checked, and it has given critics a principled way of testing a hunch about a word against general usage. Its limits are equally clear. The texts resist the tokenising and tagging on which the counts depend, the statistics are relative to a comparison that the analyst chooses, and no count dictates its own interpretation. Where it is practised with open acknowledgement of the choices involved, it sharpens arguments about modernist style without pretending to conclude them.

Conflict of interest

To be declared by the author(s) before publication.

Funding

To be declared by the author(s) before publication.

References

- Auerbach, E. (1953). *Mimesis: The Representation of Reality in Western Literature* (trans. W.R. Trask). Princeton University Press, Princeton, NJ.
- Balossi, G. (2014). *A Corpus Linguistic Approach to Literary Language and Characterization: Virginia Woolf's The Waves*. John Benjamins, Amsterdam.
- Banfield, A. (1982). *Unspeakable Sentences: Narration and Representation in the Language of Fiction*.

Routledge & Kegan Paul, Boston.

- Biber, D. (2011). *Corpus linguistics and the study of literature: Back to the future? Scientific Study of Literature*, 1(1): 15-23.
- Brooke, J., Hammond, A. and Hirst, G. (2017). *Using models of lexical style to quantify free indirect discourse in modernist fiction. Digital Scholarship in the Humanities*, 32(2): 234-250.
- Burrows, J. (2002). 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3): 267-287. <https://doi.org/10.1093/lc/17.3.267>
- Burrows, J.F. (1987). *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method*. Clarendon Press, Oxford.
- Da, N.Z. (2019). *The computational case against computational literary studies. Critical Inquiry*, 45(3): 601-639. <https://doi.org/10.1086/702594>
- Dunning, T. (1993). *Accurate methods for the statistics of surprise and coincidence. Computational Linguistics*, 19(1): 61-74.
- Fish, S. (1980). *Is There a Text in This Class? The Authority of Interpretive Communities*. Harvard University Press, Cambridge, MA.
- Hoover, D.L. (2007). *Corpus stylistics, stylometry, and the styles of Henry James. Style*, 41(2): 174-203.
- Kilgariff, A. (2005). *Language is never, ever, ever, random. Corpus Linguistics and Linguistic Theory*, 1(2): 263-276. <https://doi.org/10.1515/cllt.2005.1.2.263>
- Leech, G. and Short, M. (2007). *Style in Fiction: A Linguistic Introduction to English Fictional Prose, 2nd edn*. Pearson Longman, Harlow.
- Louw, B. (1993). *Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In: Baker, M., Francis, G. and Tognini-Bonelli, E. (eds) Text and Technology: In Honour of John Sinclair. John Benjamins, Amsterdam*, 157-176.
- Mahlberg, M. (2013). *Corpus Stylistics and Dickens's Fiction*. Routledge, New York.
- McIntyre, D. and Walker, B. (2019). *Corpus Stylistics: Theory and Practice*. Edinburgh University Press, Edinburgh.
- Moretti, F. (2013). *Distant Reading*. Verso, London.
- O'Halloran, K. (2007). *The subconscious in James Joyce's 'Eveline': A corpus stylistic analysis that chews on the 'Fish hook'. Language and Literature*, 16(3): 227-244.
- Ramsay, S. (2011). *Reading Machines: Toward an Algorithmic Criticism*. University of Illinois Press, Urbana.
- Rayson, P. (2008). *From key words to key semantic domains. International Journal of Corpus Linguistics*, 13(4): 519-549. <https://doi.org/10.1075/ijcl.13.4.06ray>
- Semino, E. and Short, M. (2004). *Corpus Stylistics: Speech, Writing and Thought Presentation in a Corpus of English Writing*. Routledge, London.
- Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford University Press, Oxford.
- Stubbs, M. (2005). *Conrad in the computer: Examining Heart of Darkness with quantitative methods. Language and Literature*, 14(1): 5-24. <https://doi.org/10.1177/0963947005048873>
- Underwood, T. (2019). *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press, Chicago.
- van Peer, W. (1989). *Quantitative studies of literature: A critique and an outlook. Computers and the Humanities*, 23(4-5): 301-307.

Cite this article as: Kobayashi D., Morin A. and Kozłowski S. (2026). *Corpus Stylistics and the Modernist Novel: Achievements and Limits of Quantitative Close Reading. International Journal of English Studies and Literature*. 1(1), 7-12. doi: 10.62715/ijesl.2026.v1.i1.pp7-12