Corpus-Based Approaches to Style, Register, and Meaning

Quantifying Literary Style: A Corpus-Based Study

Iglika Nikolova-Stoupak, Gaël Lejeune, Eva Schaeffer-Lacroix

Sens Texte Informatique Histoire, Sorbonne Université, Paris, France

iglika.nikolova-stoupak@etu.sorbonne-universite.fr,
{gael.lejeune, eva.lacroix}@sorbonne-universite.fr

Abstract

We propose a highly explainable method for the identification of a literary work's style based on a large number of quantitative characteristics. Drawing inspiration from Biber's method of classifying text (notably, as written versus oral) through multi-dimensional analysis (Biber, 1988), we argue that with the help of contemporary computational methods and an extension of the number of metrics to encompass previously unmeasurable textual aspects such as semantics, a single work's style may be efficiently defined. The analysis explores metrics defined by Biber as well as established readability features and high-level semantic ones (e.g. the use of metaphors). This short paper showcases the process of corpus compilation and related statistical analysis that leads to the definition of several texts' distinctive features as those that demonstrate the highest deviation from the established baseline for the corpus. We also determine the most overrepresented vocabulary items for the selected works. In the framework of our associated wider project, the determined features will then be used to derive adapted versions of the selected texts that seek to retain their essence. The artefacts will be evaluated through qualitative human-based surveys of reader experience. We release the associated corpus and the code used for statistical analysis.

Keywords: literary style, corpus analysis, readability, stylometry, lexical frequency

1 Introduction

The study of a text's style is especially relevant in the current technological era, where it is used, for instance, in the identification of AI-generated text. State-of-the-art classifiers of style typically make use of advanced neural methods. As a result, their function lacks interpretability. What we are proposing herein is a method for style quantification that makes use of a large number of linguistic characteristics, ranging from shallow, length-based ones to ones that make use of neural models and semantics (e.g. metaphor identification). A number of these features have been established as relevant within computer-assisted stylistics. We define a literary work's style as its key deviations in these features against a pre-calculated baseline. The validity of this definition is to be established within the underlying project's next steps, which include human-based evaluation through a survey of reader experience.

The attained conclusions about quantification of the style of a work of literature may be utilised by human authors and Natural Language Processing (NLP) tools to derive literary adaptations that retain the original work's style, thereby addressing the common criticisms that authenticity is lost through the process of adaptation, and that the reading experience associated with an adapted text is strikingly different (Hutcheon, 2006; Thiel, 2013). Our methods' main contributions are: explainability, replicability, inclusion of metatextual characteristics (such as the time and place of writing) as elements of style, and the focus on a single literary text as a stylistic entity.

The associated corpus and the code used for statistical analysis are accessible in the following GitHub repository. We encourage collaboration in the direction of expanding the linguistic feature set and improving its robustness.

2 Background and Related Work

The study of stylistics dates from the 19th century, and it met significant development around the middle of the 20th century. Spitzer (1948) saw a literary author's style as the innovation that distinguishes their texts from other texts associated with the same temporal and geographical framework. He conducted microscopic studies of vocabulary use and

was met by criticism concerning his method's subjectivity and reliance on intuition (Spitzer, 1948: p. 26). Wellek and Warren (1949) challenged the prevalent theory that literary language is merely a reflection of a given language's development, claiming that different genres are associated with different language use and that literature itself often exerts influence on language as a whole. They defined style as "deviations and distortions from normal usage". Wellek and Warren (1949) also suggested that style encompasses characteristics beyond lexical items, such as syntax, sound and rhythm. Hatzfeld (1953) included in the notion of style a variety of linguistic elements, all the way to a text's semantics. Contrary to the common focus on the author, he also viewed an isolated literary work as a main stylistic unit. Toolan (2014)'s claims about the stylistics of literature include that language correlates with meaning, that every encountered linguistic characteristic is meaningful, and that an individual work can never be analysed exhaustively.

Stylometry, also known as "computational stylistics", is the discipline concerned with the quantitative study of style. Early work, such as that of Mendenhall (1901), sought recurring linguistic patterns in corpora through manual analysis, while modern stylometry uses computational methods, thereby achieving greater objectivity and scope. Traditionally, textual features such as word frequencies, parts of speech (POS) and vocabulary richness were discussed, and additional processing typically included Principal Component Analysis (PCA). Eder et al. (2015) define stylometry as the investigation of the relationship between writing style and underlying metadata, such as authorship, textual genre, and date of composition, the main focus being on authorship. The method of defining stylistic specificity through comparison with a reference corpus, utilised in the present project, is well established. For instance, Mahlberg (2013) identifies idiosyncrasy in literary dialogue based on a corpus of non-literary language (Mahlberg, 2013).

Later stylometric classification methods include machine learning (such as logistic regression and SVM models), pre-trained models to learn authorial representations (notably, BERT) and, ultimately, Large Language Models (LLMs), which require no fine-tuning and can identify authorship based on increasingly small amounts of text (Huang et al., 2024). Common contemporary applications of stylometry include the binary classification of human versus machine-produced text and the attribution of a text to a specific language model. Comparing the performance of different LLMs at authorship identification in Early Modern English drama, Hicke and Mimno (2023) argue that the models' overall high performance is also associated with limitations, such as confident wrong attributions, which may be influenced by an author's overall popularity and prevalence in training data.

The traditional linguistic features utilised in stylometry highly coincide with those associated with readability, the study of texts' difficulty for the purpose of matching them to an appropriate audience (DuBay, 2004). Established formulas of readability, such as Flesch Reading Ease and Dale-Chall, rely on length-based measures (e.g. the number of words in a sentence) and/or the mapping of vocabulary items against frequency lists. A dichotomy of lexical and syntactic features is often utilised, and additional categories, such as "discourse", may be added. Like stylometry, readability progressed from methods based on feature engineering to black box deep learning methods. Neural classifiers are used to map text to gold labels, such as CEFR proficiency levels. Interestingly, there is evidence that state-of-the-art models closely make use of traditional readability features, proving their continuous relevance (Deutsch et al., 2020).

3 Corpus

We compiled a corpus of 100 highly acclaimed books in English. For the purpose, we made use of the ranking provided by TheGreatestBooks.org[1], which in turn is based on multiple online book rankings. To facilitate intertextual comparison, we excluded works such as plays, poetry and essays. Our final list contains mostly novels as well as additional texts that contain long narrative (such as *The Odyssey* and *One Thousand and One Nights*). 65% of the works were originally written in English, and the others include translations from various modern European languages, as well as Ancient Greek, Latin and Arabic. Thirteen authors are represented by more than one work, Dostoevsky and Hemingway coming on top of the list with four works each. The corpus includes texts composed between 700 BCE (*The Odyssey*) and 1987 (*Beloved*). All texts are available online free of charge, and the majority are included in Project Gutenberg. No additional

[1] https://thegreatestbooks.org/

OCR process was required. For the full list of texts, please refer to Appendix A.

It is important to note that broad variables pertaining to the corpus's texts, such as year of writing, original language and literary genre will undoubtedly exert influence in the statistical analysis. We consider this a problem inasmuch as the variables are not equally represented within the corpus, thereby leading to possible biases in the calculated baseline, and we therefore encourage and plan the application of the defined analysis to additional and larger corpora. However, the influence of these variables within a definition of the distinctive style of an individual literary work does not undermine our primary objective. We argue that they contribute to an isolated work's style along with more narrow (e.g. author-specific) features, and that an adapted version of the text that seeks to retain its essence is to keep track of both. Please consult our robustness analysis of the corpus pertaining to original language and year of writing in Appendix B.

With planned microscopic human-based evaluation in mind, we singled out five of the texts in the corpus for the purpose of defining their distinctive characteristics and in extension, within our proposed framework, their style: *Moby Dick*, *The Picture of Dorian Gray*, *Anna Karenina*, *Les Misérables*, and *The Odyssey*. Seeking variety in terms of represented languages and countries as well as years of composition, we selected works that are canonical in the framework of literary adaptation, as the next steps of our larger project involve the application of the attained conclusions to derive adapted versions of the texts.

4 Methods

4.1 Preprocessing

All texts in the corpus underwent a pipeline of initial preprocessing based on hand-crafted rules. All metatextual information, including prefaces (authorial or not), glossaries, notes, page numbers, and text between square brackets, was removed. Titles, including those of volumes, parts, and chapters, were also removed. Text resembling URLs as well as special symbols, typically resulting from a previously conducted OCR process, were removed. Extra spaces and empty lines were removed, and lines separated by line-break hyphens were merged.

4.2 Linguistic Features

We quantified 274 different features within the compiled corpus. Please see Table 1 for representative feature examples. For the full set of features utilised in the statistical analysis, please refer to Appendix C.

Broad feature category	Examples of included features
Tense, aspect and verbal morphology	Past tense; Present tense; Perfect aspect; Present participles; Infinitives
Reference and pronominal use	First-person pronouns; Third-person pronouns; Demonstrative pronouns; Demonstrative determiners
Syntax and clause structure	By-passives; Agentless passives; That-complements after verbs; Causal subordinators; Clausal coordination
Modification and phrase structure	Attributive adjectives; Predicative adjectives; Total adverbs; Total prepositional phrases
Lexical and grammatical distribution	Verbs; Nouns; Proper nouns; Punctuation; Content-to-function-word ratio
Lexical richness, rarity and readability	Type-token ratio; Hapax legomena; Out-of-vocabulary words; Mean word length; Words per sentence
Stance, modality and discourse	Hedges; Amplifiers; Conjuncts; Possibility modals; Private cognition verbs; Public reporting verbs
Narrative, semantic and higher-level features	Dialogue; Indirect speech; Death-related words; Mean concreteness; Metaphor use; Humor probability

Table 1: Examples of linguistic features, included in the statistical analysis.

A large number of the utilised features are traditionally used within the fields of readability and stylometry and cover the general categories "length-based", "vocabulary-related", "syntax-related", and "discourse-related" (DuBay, 2004; Collins-Thompson, 2014; Deutsch et al., 2020; Eder et al., 2015). Another large group of features is inspired by Biber's work on multidimensional text classification and includes categories such as "tense and aspects markers", "place and time adverbials", "questions", "passives" and "subordination" (Biber, 1988). The remaining features come from general literary theory, whilst special attention is accorded to elements pertaining to children's literature; the reason being that literary adaptation, which we are planning to apply our conclusions to, typically addresses a younger audience. Beauvais (2023)'s book

on children's literature has been especially useful in the process of feature engineering. Features based on literary theory include, for instance, the length of a text's first sentence as well as the presence of humour and figurative language. Due to their strong semantic implication, the last two features were estimated using transformer-based neural models, fine-tuned for binary classification of the phenomena in question: humor-no-humor[2] and Metaphor-Detection-XLMR(Wachowiak et al., 2022). For our benchmarking of the utilised neural models, please refer to Appendix D.

The assembled list of linguistic features has no claims of being comprehensive; rather, it offers a multi-faceted and consistent quantitative description of literary texts.

4.3 Statistical Analysis

The final set of features was selected not only on theoretical grounds, but also based on the feasibility of their automatic extraction. We privileged the use of automatic annotation tools as readily implementable in Python, such as the spaCy library for lemmatisation, POS tagging and morphological analysis.[3] In some cases, such as for the identification of time and space adverbials, we resorted to hard-coded lexical lists.

Wherever possible, we based our calculations on 100 tokens of text, following Biber's guidelines. Instead of single values per feature, we calculated the following set of aggregators: average value, minimal value, maximal value, and standard deviation within the text. This differentiation has been shown to provide additional insight into readability-based characteristics (Wilkens et al., 2022) and reflects the importance of considering not only feature frequencies but also their position and distribution within a text (Hoey, 1983).

We opted for the inclusion of the five texts preselected for qualitative analysis and human-based evaluation in the corpus-wise calculation of all features, as they take part in the same population of acclaimed narrative-based literary works. To define these texts' distinguishing style, we determined their top five deviating features with regard to the corpus baseline. Deviation was measured using z-scores[4]. To ensure that the identified features represent distinct stylistic dimensions, we filtered out highly correlated features using pairwise Pearson correlation with a threshold of 0.70.

In addition to the described quantitative features, we explored the most overrepresented lexical items within the five investigated texts as compared against a 10k-word frequency list based on Project Gutenberg and provided by Wiktionary[5]. To calculate deviation in word frequency, we used the following formula:

$$\text{Deviation} = \left(\frac{f_{\text{book}}(w)}{f_{\text{ref}}(w)} - 1 \right) \times 100 \qquad (1)$$

where:

- $f_{\text{book}}(w)$ = relative frequency of the word w in the text
- $f_{\text{ref}}(w)$ = relative frequency of the word w in the corpus

For instance, a value of 100% means that a word is used two times more frequently than expected. We disregarded proper nouns as well as word fragments (e.g. "don", "ll").

5 Results

316 out of the 37,401 existing feature pairs were determined to be strongly correlated. As expected, the majority of strong correlations are associated with different aggregators of the same variable as well as other logically related features, such as word-based and lemma-based type-to-token ratio. Less obvious high dependencies include these between "number of verbs (maximal)" and "number of nouns (average)" and between "number of verbs (average)" and "punctuation (maximal)", both with a value of 0.96.

Table 2 presents the top three deviating features per investigated work. Demonstrative pronouns are the most highly deviating feature for two of the texts (*The Picture of Dorian Gray* and *Les Misérables*); a related feature, demonstrative determiners, makes the list for *Moby Dick*. Most of the remaining features are related to POS and grammatical categories (verbs, pronouns, adjectives, subordination). A shallow length-based feature (sentence

[2] https://huggingface.co/mohameddhiab/

[3] We used the `en_core_web_sm` model, which reports approximately 97% POS tagging accuracy on the OntoNotes 5.0 benchmark (AI, 2025). POS tagging forms the basis for subsequent lemmatisation and morphological analysis.

[4] Defined as: $z = \frac{x-\mu}{\sigma}$, where x is the book value, μ the corpus mean, and σ the corpus standard deviation.

[5] https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists/PG/2006/04/1-10000

Book	Top deviating features
The Picture of Dorian Gray	Demonstrative pronouns (sd) ↑; Past tense (sd) ↑; Predictive modals (max) ↑
Moby Dick	Seem/appear (max) ↑; Demonstrative determiners (avg) ↑; By-passives (max) ↑
Anna Karenina	Private cognition verbs (max) ↑; Metaphor use (max) ↑; Present participles (avg) ↑
Les Misérables	Demonstrative pronouns (max) ↑; Nouns (avg) ↑; Time adverbials (max) ↑
The Odyssey	Words per sentence (min) ↑; Conjuncts (avg) ↑; Other subordination (avg) ↑

Table 2: Top three most significantly deviating features for the five investigated texts. The arrows denote the direction of deviation.

length) comes first in the case of *The Odyssey*. All top-three deviations are upward compared to the corpus baseline. For a more detailed table of results that includes five features per text as well as the associated values within the text in question and the whole corpus, please refer to Appendix E.

See Appendix F for the top 10 most overrepresented vocabulary items in the five investigated texts. The words associated with *Moby Dick*, a novel whose narrative unfolds within the self-contained microcosm of a whaling ship, demonstrate the highest values of overuse. The word "whale", most overused globally, is narrowly linked to the text's topic. The majority of all identified items are content words (notably, nouns) that are intuitive given the respective work's content (e.g. "painter" for *The Picture of Dorian Gray*, "countess" for *Anna Karenina*). Occasionally, different forms based on the same lemma are present in the lists (e.g. "peasants" and "peasant" for *Anna Karenina*). For the purpose of this study, we decided to keep such words as separate entries, as is the case in the utilised frequency list. In *Les Misérables*, the original French language's influence is clearly observable through the words "sous", "reverie" and "sombre", all of French origin. It is important to note that this specificity pertains to the specific utilised English translation of *Les Misérables* rather than the original work.

6 Conclusion and Future Directions

We compiled a corpus of 100 acclaimed literary works with strong narrative elements that span across genres, time periods and countries of origin. We went on to define a large number of quantitative textual characteristics and to use them in the establishment of a baseline pertaining to the corpus. Subsequently, we proposed a method for establishing specific texts' distinctive features as those exhibiting the highest deviations from the corpus baseline. Planned future work will focus on the following steps: 1) use of the identified stylistic features to derive (through hybrid manual and automatic methods) literary adaptations that seek to retain the original works' distinctive characteristics 2) evaluation of the resulting adaptations against neutral baselines via LLM-based as well as human evaluation. The former will consider the number of tokens necessary for recognition of the texts, and the latter will be based on a survey consisting of comprehension questions and questions describing the overall reading experience.

7 Limitations

Firstly, the current study is limited to English text and its cross-lingual relevance remains to be determined. Also, as explained, broad factors such as a work's original language of writing, genre and period of composition are all considered to play a role in its style. An alternative study seeking to control these variables would require the crafting of a more uniform and/or larger corpus. We also note that, while traditional stylistics tends to pursue a qualitative analysis of literary features, including their intended or observed effects on the reader, our project follows the tradition of stylometry in that it does not attempt a meaning-based analysis.

Although OCR-related errors or typographical inconsistencies are minimised through the processes of corpus composition and preprocessing, they have not been eliminated completely and, therefore, may influence the achieved statistics. Also, the multiple utilised linguistic features are not perfectly crafted. In particular, the ones that include hard-coded lists (such as the list of non-child-friendly words) may not be exhaustive, and the ones that make use of automatic tools (such as POS tagging) are not error-free. In particular, while the neural models used to annotate humour and metaphoric language offer a proxy of the phenomena that is consistent between texts, their annotations do not necessarily match human opinion.

References

Explosion AI. 2025. spacy: Facts & figures. https://spacy.io/usage/facts-figures. Accessed July 2026.

Clémentine Beauvais. 2023. *Écrire comme une abeille : la littérature jeunesse, de la lecture à l'écriture*. Gallimard Jeunesse, Paris.

Douglas Biber. 1988. *Variation Across Speech and Writing*. Cambridge University Press, Cambridge.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *International Journal of Applied Linguistics*, 165(2):97–135.

Tovly Deutsch, Masoud Jasbi, and Stuart M. Shieber. 2020. Linguistic features for readability assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics.

William H. DuBay. 2004. *The Principles of Readability*. Impact Information, Costa Mesa, CA.

Maciej Eder, Jan Rybicki, and Mike Kestemont. 2015. Stylometry with r: A package for computational text analysis. *The R Journal*, 8:107–121. Https://doi.org/10.32614/RJ-2016-007.

Helmut Anthony Hatzfeld. 1953. *A Critical Bibliography of the New Stylistics Applied to the Romance Literatures, 1900–1952*. Number 5 in University of North Carolina Studies in Comparative Literature. University of North Carolina Press, Chapel Hill.

Rebecca M. M. Hicke and David Mimno. 2023. T5 meets tybalt: Author attribution in early modern english drama using large language models.

Michael Hoey. 1983. *On the Surface of Discourse*. George Allen & Unwin, London.

Baixiang Huang et al. 2024. Authorship attribution in the era of large language models. In *Proceedings of the ACM Conference*.

Linda Hutcheon. 2006. *A Theory of Adaptation*. Routledge, New York.

Michaela Mahlberg. 2013. *Corpus Stylistics and Dickens's Fiction*. Routledge, London.

Thomas Corwin Mendenhall. 1901. A mechanical solution of a literary problem. *Popular Science Monthly*, 60:97–105. Accessed January 2019.

Leo Spitzer. 1948. *Linguistics and Literary History: Essays in Stylistics*. Princeton University Press, Princeton.

Elizabeth Thiel. 2013. Downsizing dickens: Adaptations of oliver twist for the child reader. In Anja Müller, editor, *Adapting Canonical Texts in Children's Literature*, pages 143–162. Bloomsbury Academic, London.

Michael Toolan. 2014. The theory and philosophy of stylistics. In Peter Stockwell and Sara Whiteley, editors, *The Cambridge Handbook of Stylistics*, pages 13–31. Cambridge University Press, Cambridge.

Lennart Wachowiak, Enrica Troiano, and Roman Klinger. 2022. Drum up support: Systematic analysis of image-schematic conceptual metaphors. In *Proceedings of the 3rd Workshop on Figurative Language Processing*, Abu Dhabi. Association for Computational Linguistics.

René Wellek and Austin Warren. 1949. *Theory of Literature*. Harcourt, Brace and Company, New York.

Rodrigo Wilkens, Thomas François, Anaïs Tack, and Cédrick Fairon. 2022. Fabra: French aggregator-based readability assessment toolkit. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1217–1233.

Appendix A Corpus "100 Novels"

#	Title	Author	Year	Original language	Translator
1	Ulysses	James Joyce	1922	English	–
2	In Search of Lost Time	Marcel Proust	1913	French	Charles Kenneth Scott-Moncrieff
3	The Great Gatsby	F. Scott Fitzgerald	1925	English	–
4	The Catcher in the Rye	J. D. Salinger	1951	English	–
5	One Hundred Years of Solitude	Gabriel García Márquez	1967	Spanish	Gregory Rabassa
6	Nineteen Eighty Four	George Orwell	1949	English	–
7	Moby Dick	Herman Melville	1851	English	–
8	Don Quixote	Miguel de Cervantes	1605	Spanish	John Ormsby
9	The Sound and the Fury	William Faulkner	1929	English	–
10	Anna Karenina	Leo Tolstoy	1877	Russian	Constance Garnett
11	Crime and Punishment	Fyodor Dostoevsky	1866	Russian	Constance Garnett
12	Pride and Prejudice	Jane Austen	1813	English	–
13	Lolita	Vladimir Nabokov	1955	English	–
14	War and Peace	Leo Tolstoy	1869	Russian	Aylmer & Louise Maude
15	Wuthering Heights	Emily Brontë	1847	English	–
16	To Kill a Mockingbird	Harper Lee	1960	English	–
17	The Lord Of The Rings	J. R. R. Tolkien	1959	English	–
18	The Brothers Karamazov	Fyodor Dostoevsky	1880	Russian	Constance Garnett
19	The Trial	Franz Kafka	1925	Czech	David Wyllie
20	The Adventures of Huckleberry Finn	Mark Twain	1884	English	–
21	Madame Bovary	Gustave Flaubert	1857	French	Eleanor Marx Aveling
22	The Stranger	Albert Camus	1942	French	Stuart Gilbert
23	The Odyssey	Homer	$\sim$ -700	Ancient Greek	Samuel Butler
24	The Grapes of Wrath	John Steinbeck	1939	English	–
25	The Divine Comedy	Dante Alighieri	1308	Italian	Henry Francis Cary
26	The Magic Mountain	Thomas Mann	1924	German	H. T. Lowe-Porter
27	To the Lighthouse	Virginia Woolf	1927	English	–
28	Jane Eyre	Charlotte Brontë	1847	English	–
29	Middlemarch	George Eliot	1871	English	–
30	Heart of Darkness	Joseph Conrad	1899	English	–
31	Alice's Adventures in Wonderland	Lewis Carroll	1865	English	–
32	Mrs. Dalloway	Virginia Woolf	1925	English	–
33	Catch-22	Joseph Heller	1961	English	–
34	The Master and Margarita	Mikhail Bulgakov	1967	Russian	Michael Glenny
35	Invisible Man	Ralph Ellison	1952	English	–
36	The Iliad	Homer	$\sim$ -700	Ancient Greek	Alexander Pope
37	Les Misérables	Victor Hugo	1862	French	Isabel Florence Hapgood
38	Great Expectations	Charles Dickens	1860	English	–
39	The Red and the Black	Stendhal	1830	French	Horace Barnett Samuel
40	Frankenstein	Mary Shelley	1818	English	–
41	On the Road	Jack Kerouac	1957	English	–
42	The Little Prince	Antoine de Saint-Exupéry	1943	French	Katherine Woods
43	Absalom, Absalom!	William Faulkner	1936	English	–
44	One Thousand and One Nights	unknown	$\sim$ 800	Arabic	Edward William Lane
45	Journey to the End of The Night	Louis-Ferdinand Céline	1932	French	Ralph Manheim
46	Lord of the Flies	William Golding	1954	English	–
47	Brave New World	Aldous Huxley	1932	English	–

#	Title	Author	Year	Original language	Translator
48	David Copperfield	Charles Dickens	1849	English	–
49	The Old Man and the Sea	Ernest Hemingway	1952	English	–
50	Beloved	Toni Morrison	1987	English	–
51	Gone With the Wind	Margaret Mitchell	1936	English	–
52	Animal Farm	George Orwell	1945	English	–
53	The Life And Opinions Of Tristram Shandy	Laurence Sterne	1759	English	–
54	The Idiot	Fyodor Dostoevsky	1869	Russian	Eva Martin
55	Dracula	Bram Stoker	1897	English	–
56	Under the Volcano	Malcolm Lowry	1947	English	–
57	The Leopard	Giuseppe Tomasi di Lampedusa	1958	Italian	Archibald Colquhoun
58	The Sun Also Rises	Ernest Hemingway	1926	English	–
59	The Golden Notebook	Doris May Lessing	1962	English	–
60	Things Fall Apart	Chinua Achebe	1958	English	–
61	Gulliver's Travels	Jonathan Swift	1726	English	–
62	Rebecca	Daphne du Maurier	1938	English	–
63	The Aeneid	Virgil	∼ -19	Latin	John Dryden
64	Robinson Crusoe	Daniel Defoe	1719	English	–
65	Midnight's Children	Salman Rushdie	1981	English	–
66	The Scarlet Letter	Nathaniel Hawthorne	1850	English	–
67	Their Eyes Were Watching God	Zora Neale Hurston	1937	English	–
68	A Farewell to Arms	Ernest Hemingway	1929	English	–
69	The Metamorphosis	Franz Kafka	1915	Czech	David Wyllie
70	The Castle	Franz Kafka	1926	Czech	Anthea Bell
71	A Passage to India	E. M. Forster	1924	English	–
72	The Plague	Albert Camus	1947	French	Robin Buss
73	The Portrait of a Lady	Henry James	1881	English	–
74	Candide	Voltaire	1759	French	William F. Fleming
75	Slaughterhouse-Five	Kurt Vonnegut	1969	English	–
76	As I Lay Dying	William Faulkner	1930	English	–
77	A Portrait of the Artist as a Young Man	James Joyce	1916	English	–
78	All Quiet on the Western Front	Erich Maria Remarque	1928	German	A. W. Wheen
79	The Count of Monte Cristo	Alexandre Dumas	1844	French	unknown
80	The Man Without Qualities	Robert Musil	1930	German	Burton Pike
81	Little Women	Louisa May Alcott	1868	English	–
82	The Good Soldier	Ford Madox Ford	1915	English	–
83	The Picture of Dorian Gray	Oscar Wilde	1890	English	–
84	Emma	Jane Austen	1815	English	–
85	Buddenbrooks	Thomas Mann	1901	German	H. T. Lowe-Porter
86	For Whom the Bell Tolls	Ernest Hemingway	1940	English	–
87	Demons	Fyodor Dostoevsky	1872	Russian	Richard Pevear & Larissa Volokhonsky
88	The Age of Innocence	Edith Wharton	1920	English	–
89	Orlando	Virginia Woolf	1928	English	–
90	Dune	Frank Herbert	1965	English	–
91	Native Son	Richard Wright	1940	English	–
92	The Unbearable Lightness of Being	Milan Kundera	1984	Czech	
93	Dead Souls	Nikolai Gogol	1842	Russian	D. J. Hogarth
94	Doctor Zhivago	Boris Pasternak	1957	Russian	Max Hayward & Manya Harari
95	The Bell Jar	Sylvia Plath	1963	English	–
96	Charlotte's Web	E. B. White	1952	English	–
97	In Cold Blood	Truman Capote	1966	English	–
98	Treasure Island	Robert Louis Stevenson	1883	English	–

#	Title	Author	Year	Original language	Translator
99	Vanity Fair	William Make-peace Thackeray	1847	English	–
100	The Heart Is A Lonely Hunter	Carson McCullers	1940	English	–

Appendix B Corpus Robustness Analysis

We conducted a robustness analysis in order to measure the influence of the variables of time and original language of writing in relation to our corpus. For this purpose, we removed the following from the original corpus: all works whose original language of writing was not English (35 in number), and all works composed outside the years 1860–1970 (14 in number; the time span was selected based on the general timeline of literary modernism while ensuring that the resulting corpus remained approximately half the size of the original corpus). The resulting corpus consists of 49 works. For each linguistic characteristic used in the project, we compared its mean value in the full and restricted corpora. We report the following information: the percentage change between the two means and the Hedges' g (a standardised measure of the magnitude of the difference).

Statistic	Absolute % change	Absolute Hedges' g
Mean	6.4576	0.1196
Median	3.9748	0.0656
Standard deviation	8.4231	0.1427
Minimum	0.0000	0.0000
Maximum	65.9864	0.4450

Table 4: Summary of the absolute changes in feature-wise statistics between the full and restricted corpora.

Overall, the restricted corpus shows only limited value differences compared to its full counterpart. The mean and median absolute Hedges' g values (0.120 and 0.066, respectively) are both below the conventional threshold for a small effect ($g = 0.20$). Even the largest observed effect ($g = 0.445$) is below the 'medium' threshold. The largest differences are associated with the following features: humour, adverbials of place and time and length of first sentence.

Appendix C Quantitative Linguistic Features

Verb tense and aspect
Past tense
Present tense
Perfect aspect
Present participles
Past participles

Adverbials and questions
Place adverbials
Time adverbials
Direct wh-questions

Pronouns and reference
Pronouns, total
First-person pronouns
Second-person pronouns
Third-person pronouns
Demonstrative pronouns
Indefinite pronouns
Demonstrative determiners

Nominal and verbal form
Nominal suffixes
Gerundive nominals in *-ing*
Words differing from lemma
Infinitives

Voice and predication
By-passives
Agentless passives
Copular constructions
Existential constructions

Complementation and relatives
That-complements after verbs
That-complements after adjectives
Wh-free relatives
Sentence relatives

Subordination
Causal subordinators
Concessive subordinators
Conditional subordinators
Other subordinators

Modification and phrase structure
Total prepositional phrases
Attributive adjectives
Predicative adjectives
Total adverbs
Phrasal coordination
Clausal coordination

Lexical and grammatical categories
Punctuation
Verbs
Nouns
Proper nouns
Content-to-function-word ratio

Lexical richness and rarity
Type-token ratio (words)
Type-token ratio (lemmas)
Hapax legomena (lemmas)
Out-of-vocabulary words

Readability and length
Mean word length
Words per sentence
Syllables per sentence
Letters per word

Syllables per word
Number of tokens*
Number of sentences*
Words in first sentence*
Characters in first sentence*

Stance and discourse markers
Conjuncts
Downtoners
Hedges
Amplifiers
Emphatics
Discourse particles

Modality and reporting
Possibility modals
Necessity modals
Predictive modals
Public reporting verbs
Private cognition verbs
Suasive/directive verbs
Seem/appear

Negation
Analytic negation
Synthetic negation

Narrative and discourse structure
First-person narration
Dialogue
Indirect speech
Fragments without verbs
Non-finite fragments

Semantic and lexical content
Death-related words
Sex-related words
Total flagged words
Concreteness coverage
Mean concreteness
Concreteness sum

Other higher-level features
Metaphor use (%)
Mean metaphor probability
Metaphorical words total
Humor probability
Non-humor probability
Humor label

Figure 1: The quantitative linguistic features measured in the context of our corpus-based analysis. All features except those marked with * were measured per 100 textual tokens and are represented by average, minimum, maximum and standard deviation values within the given text.

Appendix D Neural Models Benchmarking

We performed manual benchmarking in relation to metaphor and humour detection as neural-based features that are of particularly subjective nature and therefore challenging to quantify. Our main goals were to test the models' consistency and their performance on literary text. For the metaphor detection model, we manually annotated a short (380-word) extract and compared it with the model's output. Our gold standard consisted of 14 metaphors, whereas the model identified 23, 6 of which coincided with our annotations. We determined that the model's threshold is rather low in that it marks a large number of commonly used fixed expressions as metaphors (e.g. "plenty of time", "to make out", "to have nothing to do with"). For the humour detection model, we manually annotated 20 random 100-token extracts. We identified 7 humorous extracts, while the model identified 13, with 7 false positives and 1 false negative. In its clear bias toward the positive label, it tended to mark dialogue, casual and adult language as humorous. Ultimately, we concluded that despite the models' biases, they provide valuable linguistic information for our project's purposes due to their highly consistent identification tendencies.

Appendix E Representative Characteristics

Book	Feature	Value	Mean	SD	Z
The Picture of Dorian Gray	Demonstrative pronouns (sd)	1.16	0.85	0.13	2.38
	Past tense (sd)	4.91	3.88	0.46	2.26
	Predictive modals (max)	8.00	5.24	1.29	2.14
	Copular constructions (max)	11.00	7.77	1.61	2.00
	Present tense (sd)	4.23	3.24	0.51	1.96
Moby Dick	Seem/appear (max)	10.00	4.51	1.27	4.33
	Demonstrative determiners (avg)	1.11	0.57	0.19	2.87
	By-passives (max)	4.00	2.35	0.63	2.64
	Attributive adjectives (avg)	5.98	3.83	0.90	2.40
	Non-finite fragments (max)	26.00	10.55	7.05	2.19
Anna Karenina	Private cognition verbs (max)	12.00	6.43	1.74	3.19
	Metaphor use (%) (max)	36.84	27.49	4.07	2.30
	Present participles (avg)	0.91	0.52	0.18	2.18
	Verbs (avg)	148.10	59.10	42.42	2.10
	Causal subordinators (max)	5.00	3.77	0.65	1.90
Les Misérables	Demonstrative pronouns (max)	20.00	8.51	1.97	5.82
	Nouns (avg)	305.94	70.93	52.48	4.48
	Time adverbials (max)	12.00	5.64	1.54	4.13
	Amplifiers (max)	12.00	5.84	1.55	3.97
	Out-of-vocabulary words (max)	84.00	35.20	12.94	3.77
The Odyssey	Words per sentence (min)	12.50	5.75	1.72	3.92
	Conjuncts (avg)	0.41	0.14	0.10	2.73
	Other subordinators (avg)	0.95	0.68	0.12	2.20
	Phrasal coordination (avg)	4.16	2.84	0.61	2.18
	Downtoners (max)	1.00	2.59	0.75	-2.13

Table 5: Most strongly deviating features for the five investigated texts. The columns denote the features' values for the text in question, their mean values and standard deviation for the corpus and the calculated z-values. Numbers are rounded to two decimal points.

Appendix F Lexical Analysis

Work	#	Word	Overuse (%)
Moby Dick	1	whale	71434
	2	aye	8214
	3	mast	6744
	4	sir	4438
	5	mates	3287
	6	rigging	2848
	7	carpenter	2743
	8	jaw	2274
	9	aloft	2150
	10	aft	2113
The Picture of Dorian Gray	1	duchess	7025
	2	sir	3971
	3	painter	2278
	4	portrait	2148
	5	coloured	2113
	6	gilt	1981
	7	studio	1980
	8	hideous	1844
	9	monstrous	1724
	10	valet	1681
Anna Karenina	1	countess	4268
	2	princess	2737
	3	marsh	2323
	4	divorce	2300
	5	everyone	2164
	6	peasants	1883
	7	yes	1665
	8	coachman	1487
	9	peasant	1458
	10	onto	1428
Les Mis-érables	1	sous	3841
	2	convict	3031
	3	mayor	2423
	4	bishop	2034
	5	grating	1854
	6	sir	1753

Work	#	Word	Overuse (%)
	7	a	1686
	8	reverie	1495
	9	indescribable	1323
	10	sombre	1321
The Odyssey	1	raft	3862
	2	meats	3754
	3	maids	3152
	4	groaning	2769
	5	mast	2680
	6	goddess	2651
	7	armour	2546
	8	offerings	2446
	9	gods	2213
	10	barley	2105

Table 6: Top 10 most overused words in the five investigated works as compared against the Project Gutenberg 10k frequency list. All values are rounded to whole numbers.