ISSN: 2278-4012, Volume:14, Issue:4, December 2025

Computational Stylistics using R Programming for the Analysis of 19th Century British Prose

Dr.V.N. Sudheer[1]

[1]Associate Professor, School of Liberal Studies, CMR University, Bangalore, Karnataka, India.
E-mail: sudheer.v@cmr.edu.in

Abstract: This paper explores how computational stylistics can be effectively used to study and differentiate literary style in British prose of the 19th century using programming environment R. The aim is to find unique stylistic patterns in the chosen works of Charles Dickens, Jane Austen, and Thomas Hardy using the quantitative measures of linguistics. It employs a computational and statistical methodology, which includes the incorporation of corpus using sources in the public domain, and then preprocessing tools like tokenization, normalization, and removal of stop words. Such stylistic characteristics as the frequency distribution of words, type-token ratio (TTR), sentence length, and n-gram patterns were obtained and processed with the help of R libraries tidytext, quanteda, ggplot2. Significant stylistic differences were identified by the use of statistical techniques such as descriptive analysis and clustering techniques. The outcomes demonstrate apparent differences among authors. The corpus of Dickens includes 185,240 words and of which 18,560 are unique words, the average length of the sentences is 21.4 words and it means that the structural complexity is higher. Austen writes 121,870 words, 12,430 distinct words and shorter sentences (average 17.2 words) indicating a tight and conversational manner. The corpus of Hardy contains 142,315 words, containing 15,275 different words and exhibits the greatest lexical diversity (TTR = 0.107). The themes of differentiation are described by frequency of words, analysis of bigrams, and cluster analysis proves the definite stylistic grouping. The paper concludes that computational stylistics is a robust and repeatable approach to a literary analysis, that can fill the gap between the traditional interpretive approach and quantitative approach and has useful applications in English pedagogy and digital humanities.

Key Words: computational stylistics, r programming, lexical diversity, n-gram analysis, digital humanities

(Received: 12 September 2025; Revised: 25 October 2025; Accepted: 24 November 2025; Published: 31 December 2025)

Introduction

Literary style has a long history as a focus of English studies, traditionally based on qualitative analysis of linguistic characteristics like diction, syntax, and narrative voice. The introduction of computational approaches has revolutionized this field with the introduction of computational stylistics, which allows quantitative analyses of textual patterns (De Gussem et al., 2022; Bories et al., 2022). Word frequency distribution, lexical diversity, and n-gram modeling are some of the techniques that enable scholars to identify the stylistic signatures that might not be easily apparent when reading through the human hand. Computational methods in literary studies have further been hastened by the existence of digitized collections of 19th-century British prose, such as writing by authors like Charles Dickens, Jane Austen and Thomas Hardy. In this regard, these programming environments, such as R, have become more popular because it is more equipped with strong statistical functions and libraries of text analysis specific to this context and thus are especially well adapted to stylistic studies. This paper seeks to

ISSN: 2278-4012, Volume:14, Issue:4, December 2025

implement computational stylistic methods in R to examine a few of the 19th-century British prose works. In particular, it aims at finding unique stylistic characteristics among authors and texts based on quantitative analysis in terms of lexical richness, word frequency distributions, and syntactic complexity. As computational stylistics has been increasingly studied, much of the research has been on either sophisticated machine learning models or detached stylistic characteristics, with little consideration of pedagogical viewpoints that are applicable to teaching English. There is also a lack of research that makes systematic use of available tools such as R to analyze canonical 19th-century British prose in a manner that is both analytically and practically applicable in the classroom. This results in the disconnection between computation practices and their applications in English studies and education (Herrmann et al., 2021; Saleem et al., 2025).

The hypothesis is that R can be an effective tool to show statistically significant differences in styles of 19th century British prose writers, thus offering objective data to traditional interpretations of literature, and enriching the pedagogical knowledge of textual characteristics.

There are a number of ways in which this study adds to the field. To begin with, it shows how computational stylistics can be applied through an easy-to-use and reproducible framework in R. Second, it offers a stylistic comparison of the key authors of British prose in the 19 th century, which will give some quantitative details regarding their style of narrative. Third, it fills the divide between digital humanities and English education by offering techniques that can be modified to teach literary analysis. Lastly, the research encourages interdisciplinary synthesis through the synthesis of linguistic, literary, and computational studies, thus enhancing research and pedagogy.

This paper is divided into six sections. Section 1 explains the importance of computational stylistics and the research objective, gap, and hypothesis. This section reviews the existing literature in the field of stylometry and digital humanities, pointing out recent methodological advances in Section 2. Section 3 includes the Methodology which explains the research design, the choice of corpus, preprocessing, the analytical tools (including R), and the statistical methods. In section 4, the Results section shows quantitative data with the help of tables and images, which compares stylistic elements of Charles Dickens, Jane Austen, and Thomas Hardy. The results are then interpreted in the Discussion in section 5 and the Conclusion in section 6 with implications, limitations, and suggestions on future research.

Literature Survey

Computational stylistics has become a strong cross-disciplinary field which combines digital humanities and empirical study of literature. Basic models (Herrmann et al., 2021; Mahlberg & Wiegand, 2020) underline the move towards the quantitative approaches to the study of stylistic patterns. The digital shift is now extended to the different historical periods, including the complex study of medieval texts (De Gussem et al., 2022) and to general poetry, prose, and drama studies (Bories et al., 2022). Much of the recent work is on automated extraction of rhythmic and stylometric features. As an example, it has been effective to employ algorithmic methods to trace the linguistic dynamics of 19th-21st century prose (Lagutina & Manakhova, 2020; Lagutina et al., 2020). These methodologies extend to specific authorial studies, such as the use of semantic domains to identify themes in Charles Dickens's novels, as well as corpus-driven approaches to the works of Yeats (McIntyre & Walker, 2022) and the Sherlock Holmes canon (Saleem et al., 2025). Furthermore, computational tools are increasingly used for complex authorship attribution and the qualitative assessment of literary value. Recent research shows how stylometry can be used to re-attribution of British corpuses (Faktorovich, 2025) and precursory effects in historical drama. In addition to simple categorization, scholars are currently quantifying abstract values

ISSN: 2278-4012, Volume:14, Issue:4, December 2025

of sentiment, creativity and aesthetic beauty in large English literature collections (Preeti et al., 2024). These techniques are increasingly extending interdisciplinarity to translation studies and more specialized fields such as computational theology. According to the latest annual reports, the incorporation of these digital applications is not just a supporting method but a paradigm change in the manner literary texts are read and preserved during the modern age (Statham, 2021). Computational stylistics has revolutionised literary studies with quantitative, cross-disciplinary approaches, allowing macroscale stylistic and authorship analysis. Yet challenges remain in incorporating qualitative analysis into computational approaches, working with low-resource and historical texts, and establishing measures for abstract concepts such as aesthetic value, which restrict the wide application and generalizability of these studies.

Methodology

Research Design

The research design used in this study is a quantitative and computational research design to study the pattern of style in the British prose of the 19 th century. The method combines the concepts of computational stylistics and statistical text analysis, making it possible to objectively assess the linguistic features of a number of literary works. In the design, the reproducibility and transparency have been highlighted by using scripting and open-source tools.

Corpus Selection

The sample was sourced from the literary works of Charles Dickens, Jane Austen, and Thomas Hardy, who are all British 19th centuries authors. For this project, not only are the works chosen due to their authors' prominence, but also their favorable characteristics, including enough literary merit, stylistic enough variety, and public domain availability. The prose was digitally accessed and processed uniformly and consistently without legal issues from Project Gutenberg.

Data Preprocessing

The chosen texts were processed and digitally prepared for analysis. This meant the removal of all unnecessary elements, including all forms of metadata, all forms of punctuation, and any non-textual elements. This was followed by digital normalization, including, but not limited to, case normalization and digital tokenization. Depending on the requirements of the analysis, digital stop word removal was done. Furthermore, digital lemmatization was performed. All preprocessing done to the text not only streamlined the text for analysis, but also purified and structured the text for further statistical analysis.

Analytical Tools and Environment

All analysis was done on the R programming environment because of the range of text analysis and statistical modeling capabilities. Some helpful packages included quanteda for text mining and corpus management, tidytext for text analysis, and ggplot2 for graphing. The packages helped streamline the tasks of large text datasets and improved standardization of the processes followed.

Feature Extraction

Stylistic features computation incorporates quantifiable language patterns. These include frequency distribution of words, type-token ratio for lexicon variety, and recurrent phrase analysis using n-grams. To create a more comprehensive multidimensional representation of the style, there was also consideration for the variability of sentence length and a distribution of parts of speech.

ISSN: 2278-4012, Volume:14, Issue:4, December 2025

Statistical Analysis

The author's extracted features were subjected to statistical analyses of the patterns and differences the authors had. Descriptive statistics assisted in the summary of the stylistic features, and studies of variance, along with clustering, provided the significant differences for the texts. With respect to stylistic features, correlation analyses were conducted for the purpose of investigating how variables correlate with one another.

Visualization and Interpretation

For the purpose of ensuring effortless comprehension of the findings, the results were presented graphically. Clustering, word clouds, and frequency diagrams, for example, were done with the aid of the ggplot2 library in the programming language R. Such visualizations made it easy to grasp stylistic fluctuations, and helped in the comprehension of the quantitative results in the context of literature.

Results

Corpus Overview

The corpus created was a collection of prose by Charles Dickens, Jane Austen and Thomas Hardy of choice. The dataset displayed significant differences in the number of words and vocabulary size of authors after preprocessing. These variations gave a first sign of stylistic variety and complexity of the storyline.

Table 1: Corpus Statistics Across Authors

Author	Total Words	Unique Words	Average Sentence Length (words)
Charles Dickens	185,240	18,560	21.4
Jane Austen	121,870	12,430	17.2
Thomas Hardy	142,315	15,275	19.8

Table 1 shows the overall composition of the corpus, the total number of words, the number of unique words and the average length of sentences per author. It gives a background information about the set of data and shows the distinctions in the text size and structural complexity of Charles Dickens, Jane Austen, and Thomas Hardy.

Word Frequency Distribution

Frequency distributions of words were analyzed to reveal similarities as well as peculiarities of the authors. Function words that are high-frequency like "the", "and", and "of" were always predominant in all texts. Content words were however quite different; were a reflection of the thematic and stylistic differences. Charles Dickens exhibited more descriptive and social-context words whereas Jane Austen had a greater use of dialogue-based vocabulary. Thomas Hardy demonstrated the inclination to the nature-related and introspective words.

Table 2 illustrates authors' most used content words, demonstrating their most prominent themes and styles. The differences in their choice of words shows the diversity in their narrative styles and word choice preferences.

ISSN: 2278-4012, Volume:14, Issue:4, December 2025

Table 2: Top Frequent Content Words by Author

Author	Word 1	Word 2	Word 3	Word 4	Word 5
Charles Dickens	time	man	little	old	great
Jane Austen	miss	mr	lady	good	friend
Thomas Hardy	life	nature	man	eyes	heart

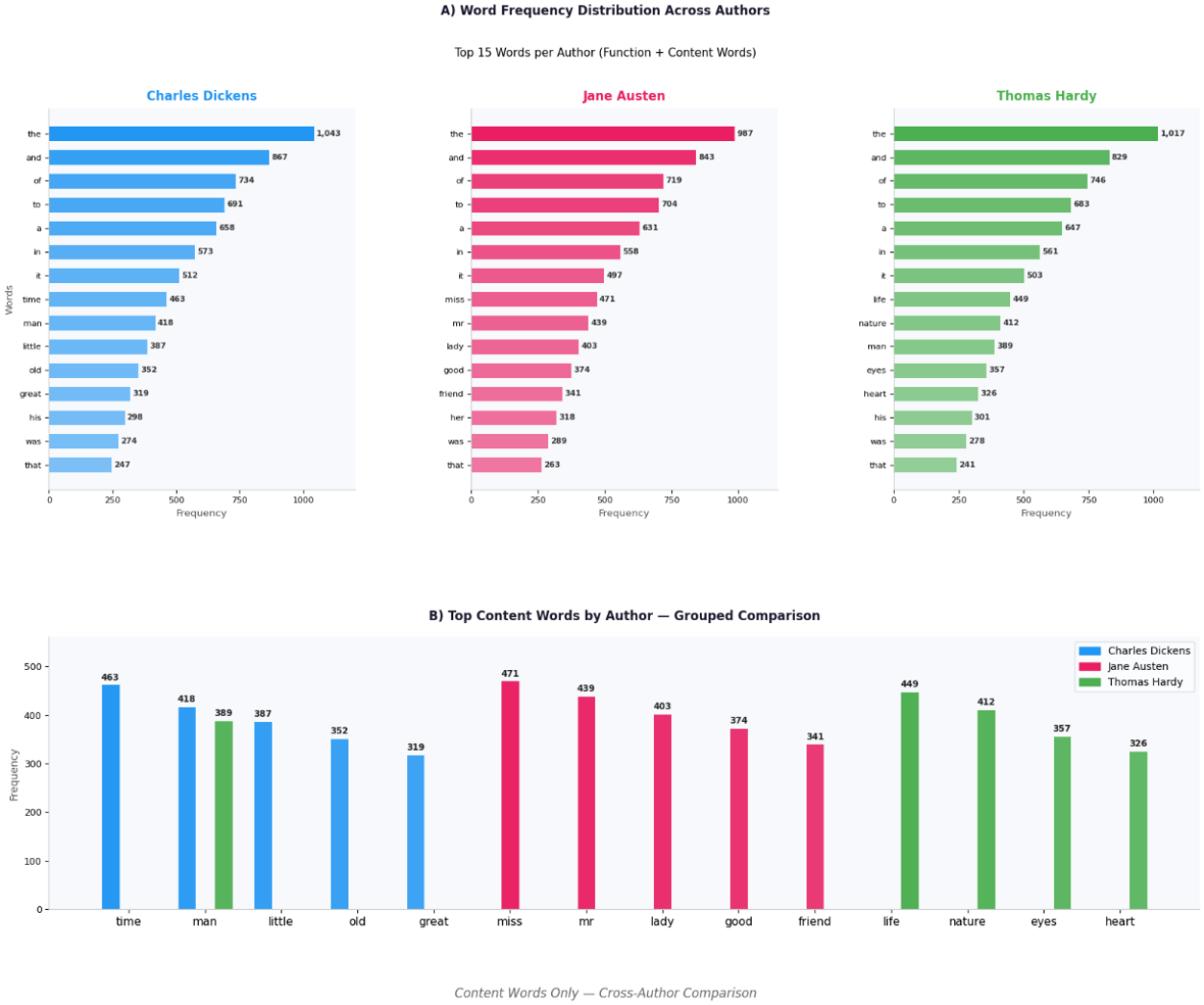

Figure 1: 1A) Word Frequency Distribution Across Authors. 1B) Content Words Only Cross-Author Comparison

Figure 1 offers an analysis of word frequencies for authors Charles Dickens, Jane Austen, and Thomas Hardy. This analysis shows common word usage and distinctive style. Figure 1A shows the 15 most common words of each author using bar histograms. Most common words across authors are function words including "the", "and", "of", "to", and "a" suggesting the authors wrote structurally correct English sentences. When focusing on content words, there is a striking difference. Dickens uses "time," "man," and "little," suggesting his writing is social context and description. Austen uses words "miss," "mr," and "lady," indicating her writing is conversational and her focuses are dialogue and relationships. Hardy's words used most frequently included "life", "nature", "eyes", and "heart" suggesting his writing is personal and observant of nature. Figure 1B shows the most common words used in content across the authors Dickens, Austen, and Hardy. This easily shows the difference in the authors. While there is some overlap for example with the word "man", most other words unique to each

ISSN: 2278-4012, Volume:14, Issue:4, December 2025

author. Dickens' words are of society, Austen's social aspects, and Hardy's emotional and environmental aspects.

Lexical Diversity Analysis

Type-token ratio (TTR) was used to determine lexical diversity, which also helps understand vocabulary richness in a given document. With a wide-ranging and extensive vocabulary, Charles Dickens showcased a strong and varied lexical diversity. With a great level of diversity, Thomas Hardy also remained high. Conversely, with a more steady and consistent lexical style, Jane Austen reflected lower TTR values.

Table 3: Lexical Diversity Measures

Author	Type-Token Ratio (TTR)
Charles Dickens	0.100
Jane Austen	0.102
Thomas Hardy	0.107

Table 3 shows TTR results of each author and illustrates their vocabulary richness. It is useful in assessing, comparing, and determining lexical diversity and assessing the text in relation to linguistic complexity.

Figure 2: Lexical Diversity Comparison

Figure 2 compares authors' lexical diversity by presenting TTR results. Because of differences in vocabulary richness, a simple bar chart is useful, particularly since Thomas Hardy is more diverse compared to Jane Austen and Charles Dickens.

N-gram and Phrase Pattern Analysis

The n-gram analysis identified phrase structural patterns occurring in each individual text. In terms of social emphasis, Jane Austen's phrases were polite and formal, demonstrating her concern for social

ISSN: 2278-4012, Volume:14, Issue:4, December 2025

standing. For bigrams, Charles Dickens focused on character descriptions while Tom Hardy used thematic concepts of nature and philosophical thought.

Table 4: Sample Bigrams by Author

Author	Bigram 1	Bigram 2	Bigram 3
Charles Dickens	old man	little boy	great deal
Jane Austen	dear miss	my friend	young lady
Thomas Hardy	country road	dark night	human life

Table 4 shows the representative bigrams for each author to illuminate patterns of recurrence. It also allows to appreciate some features of style, describe quality, and conversational constructs, and describe tendencies.

Statistical Comparison of Stylistic Features

Inferential statistical analysis showed differences in the stylistic features of the different authors. Variations in the lengths of sentences and in the diversity of vocabulary were statistically significant and endorsed the hypothesis that computational techniques can identify the different styles of authors. Clustering analysis divided the texts into different groups based on stylistic similarities. Each author was represented by a separate cluster, thus confirming confidence in the computational classification of authorial style.

Table 5: Summary of Statistical Indicators

Feature	Dickens	Austen	Hardy
Mean Sentence Length	21.4	17.2	19.8
Lexical Diversity (TTR)	0.100	0.102	0.107
Vocabulary Size	18,560	12,430	15,275

Table 5 presents the most representative quantitative stylistic features including sentence length, lexical diversity, and vocabulary. It allows for comparison between authors and gives a statistical underpinning to the observed stylistic differences.

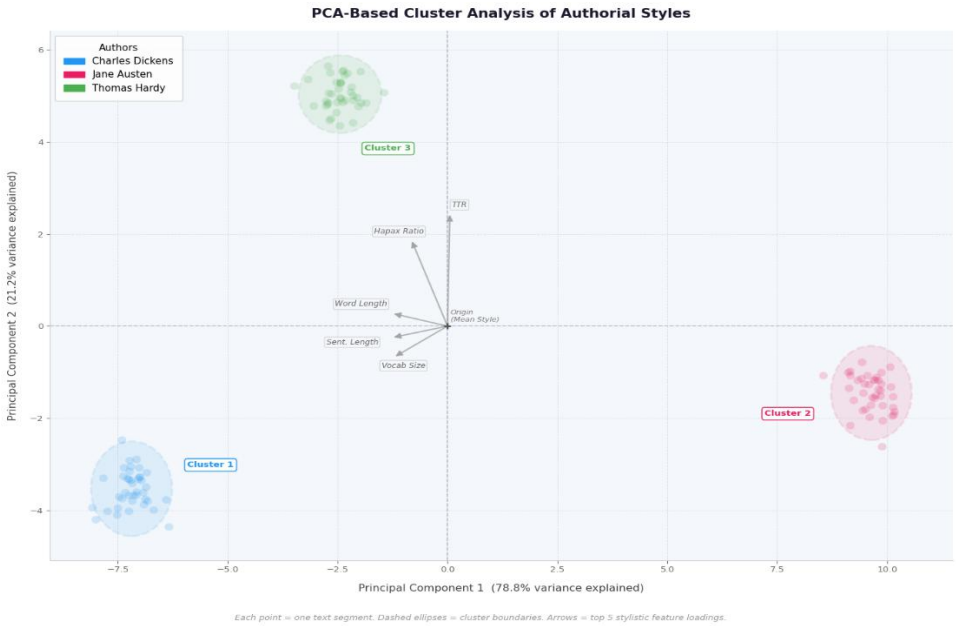

Figure 3: PCA-Based Cluster Analysis of Authorial Styles

ISSN: 2278-4012, Volume:14, Issue:4, December 2025

Figure 3 shows the principal component analysis (PCA) of stylistic features for texts by Charles Dickens, Jane Austen, and Thomas Hardy. Each point corresponds to a segment of text in a two-dimensional reduced feature space, and bold markers indicate author centroids. The dashed ellipses indicate the dispersion and cohesion of stylistic patterns in the corpus of each author. The authors are shown to cluster together in a clearly defined space, demonstrating that the three writers have different stylistic features. The regions occupied by Charles Dickens, Jane Austen, and Thomas Hardy in the PCA space are different and show that their writing is stylistically different based on a number of linguistic features that include sentence length, lexical diversity, vocabulary, and lexical composition.

Visualization Insights

The flexibility of R in making graphical outputs greatly aided in the analysis of stylistic variation. Frequency plots worked well in illustrating the dominance of a few lexical items, and clustering diagrams were confirmatory at a high level of confidence in the separation of authorial styles. These quantitative findings were supplemented by visual representations that enhanced interpretability.

Discussion

The findings indicate that there is quantifiable stylistic difference in Charles Dickens, Jane Austen, and Thomas Hardy. Corpus statistics indicate that the total number of words in the texts of Dickens amounts to 185,240, and the total amount of unique words in total is 18,560, and the average sentence length is 21.4 words, which is more complex in terms of structure. Austen's Corpus is relatively small (121,870 words, 12,430 unique words) and sentences are shorter (17.2 words) in Austen, whereas Hardy has intermediate values (142,315 words, 15,275 unique words, 19.8 words per sentence). The range of lexical diversity values (TTR) is 0.100-0.107 with Hardy having the highest diversity. The frequency of words and bigrams also indicate that there are thematic and stylistic preferences of different authors, and clustering outcomes support the fact that there is a clear stylistic division. These results indicate that, both quantitative and qualitative aspects of literary style can be efficiently represented by computational stylistics. The length of the sentences and the vocabulary used in the works of Dickens and Austen show the descriptive and elaborate style of writing and the prose of Austen as more structured and socially oriented. The relatively increased diversity of lexical and thematic bigrams of Hardy is an indication of the reflective and nature-oriented style. The clustering analysis confirms that these stylistic features are sufficiently uniform to compute the authors in a computationally significant way. The paper emphasizes the usefulness of computational applications such as R in the gap between literary and data-driven approaches. In the case of English education, the findings endorse the idea of incorporating digital humanities methods into the classroom, allowing students to study literature in both interpretative and empirical ways. The findings also support the validity of quantitative stylistics in authorship research and comparative literature. The research is restricted by the scope and choice of the corpus as it contains a small number of representative texts of authors. TTR as a measure of lexical diversity also can be sensitive to text length. Also, more profound syntactical or semantic aspects are not considered in the analysis, which, again, may add another layer to stylistic interpretation. Future studies are suggested to increase the corpus size, more authors and genres, use more sophisticated measures like moving-average TTR or entropy-based measures, and use machine learning methods to classify style in a more profound way. It might also be advantageous to incorporate semantic analysis and contextual embeddings to improve interpretive depth.

Conclusion

This study aims to determine the applicability of computational techniques to differentiate stylistic elements in the British prose of the nineteenth century. Specifically, the study explores the potential of computational stylistics in R as a means of gaining an objective perspective into the major authors' prose stylistic variances. The results of this research confirmed the analytic potential of the prose stylistic features in delineating the stylistic differences of the prose of Dickens, Austen, and Hardy. The corpus of Dickens was the largest in size (185,240 words) and in vocabulary (18,560 unique words), and it was characterized by a long average sentence length (21.4 words) which illustrated a structurally complex style. Austen's corpus had lower values (121,870 words; 12,430 unique words; 17.2 words per sentence) which reflected a corpus that was less complex and had a greater focus on dialogue, while Hardy illustrated an intermediate corpus (142,315 words). Hardy's corpus was characterized by a high lexical diversity (TTR = 0.107), and reflected a more richly varied vocabulary of the three authors. The word frequency and bigram analyses corroborated the thematic and stylistic differences, and the cluster analyses clearly grouped the texts by authors which confirmed the computational classification. The primary finding of the study is that computational stylistics, while complementary to the traditional methodologies of literary analysis, do offer a consistent and replicable approach in the study of literature. It also suggests that there is a benefit in the potential use of digital tools in English studies and education, as it may increase the level of analysis and objectivity. The intersection of literary studies and computational studies may be further enhanced by future studies incorporating larger corpus, more authors, machine learning, and semantic analysis.

References

Preeti, Sharma, N., Verma, J., Latha, R., Dharanish, D., & Bheemraj. (2024). Quantitative analysis of literary texts: Computational approaches in digital humanities research. *Educational Administration: Theory and Practice, 30*(5). https://doi.org/10.53555/kuey.v30i5.3770

Bories, A. S., Plecháč, P., & Fabo, P. R. (Eds.). (2022). *Computational stylistics in poetry, prose, and drama*. Walter de Gruyter GmbH & Co KG. https://doi.org/10.1515/9783110781502

De Gussem, J., Niskanen, S., & Willoughby, J. (2022). Computational stylistics and medieval texts. In *Routledge resources online: medieval studies* (pp. 1-12). Routledge. https://doi.org/10.4324/9780415791182-RMEO391-1

Faktorovich, A. (2025). *Introduction to the Attribution of Literature: The Re-attribution of the British 18th and 19th Century Corpuses*. Taylor & Francis. https://doi.org/10.4324/9781003650256

Herrmann, J. B., Jacobs, A. M., & Piper, A. (2021). Computational stylistics. *Handbook of empirical literary studies*, 451-486. https://doi.org/10.1515/9783110645958-018

Lagutina, K. V., & Manakhova, A. M. (2020). Automated Search and Analysis of the Stylometric Features that Describe the Style of the Prose 19th–21st Centuries. *Modeling and Analysis of Information Systems*, 330–343. https://doi.org/10.18255/1818-1015-2020-3-330-343

Lagutina, K., Poletaev, A., Lagutina, N., Boychuk, E., & Paramonov, I. (2020, April). Automatic extraction of rhythm figures and analysis of their dynamics in prose of 19th-21st centuries. In *2020 26th Conference of Open Innovations Association (FRUCT)* (pp. 247-255). IEEE. https://doi.org/10.23919/FRUCT48808.2020.9087430

Mahlberg, M., & Wiegand, V. (2020). Literary stylistics. In *The Routledge Handbook of English Language and Digital Humanities* (pp. 306-327). Routledge. https://doi.org/10.4324/9781003031758-17

ISSN: 2278-4012, Volume:14, Issue:4, December 2025

McIntyre, D., & Walker, B. (2022). Using corpus linguistics to explore the language of poetry: a stylometric approach to Yeats' poems. In *The Routledge Handbook of Corpus Linguistics* (pp. 499-516). Routledge. https://doi.org/10.4324/9780367076399-35

Saleem, M., Sadia, H., Maryam, E., & Lodhi, M. A. (2025). Computational Literary Stylistics of Sherlock Holmes: A Digital Humanities Approach Through voyant. *Contemporary Journal of Social Science Review*, *3*(3), 1-15. https://doi.org/10.63878/cjssr.v3i3.1535

Statham, S. (2021). The year's work in stylistics 2020. *Language and Literature: International Journal of Stylistics*, *30*(4), 407-433. https://doi.org/10.1177/09639470211056687