The paradigm of literary work analysis from the perspective of corpus stylistics

Amelia Davis[1, a*], Jack Brown[1, b],

[1]University of Bristol, Tyndall Avenue, Bristol, BS8 1TH, United Kingdom
a.davis_2004@yahoo.com, b.jackbrown_stud81@gmail.com
***Corresponding Author**

Abstract: In recent years, with the development of corpus linguistics, the gradual integration of stylistics with corpus has given rise to a new approach to literary research—Corpus Stylistics. Corpus Stylistics utilizes corpora and retrieval software to conduct empirical quantitative and qualitative analysis of the stylistic features of literary texts. The article starts from the advantages of applying Corpus Stylistics to the analysis of literary texts, discusses the perspectives and approaches of applying Corpus Stylistics to the analysis of literary texts, with the aim of promoting the research and development of Corpus Stylistics in China to a certain extent.

Keywords: Corpus Stylistics, Literary Text Analysis, Paradigm

1. Introduction

With the rapid development of corpus linguistics, the interdisciplinary application prospects of corpus linguistics have been increasingly valued by scholars. In recent years, the gradual integration of corpus linguistics and stylistics has given rise to a new approach to literary research—Corpus Stylistics. Corpus Stylistics, with the help of corpora and search software, uses empirical methods to describe the language, writing characteristics, and ideological expression in discourse. This new research approach can provide a quantitative description and qualitative analysis of the stylistic features of literary texts, offering a fresh perspective for the analysis and appreciation of literary works, and providing a broad prospect for the development of literary criticism. This paper discusses the advantages, approaches, and perspectives of applying Corpus Stylistics to the analysis of literary texts, in the hope of promoting the research and development of Corpus Stylistics in China to a certain extent.

2. Definition and advantages of corpus stylistics

Since the 1980s, scholars have used corpus linguistics methods to study stylistics, incorporating the language and meaning of literary texts into the research scope of corpus linguistics, and the term "corpus stylistics" began to emerge. The book "Corpus Stylistics: Speech, Writing and Thought Presentation in English Written Texts" co-authored by Semino and Short is the first monograph combining corpus linguistics with stylistic analysis. This book initially established the methodology of corpus stylistics research, opening up a new research approach for stylistic studies. Hardy believes that corpus stylistics is the use of corpus linguistics tools and methods to teach and research stylistics [1]. Toolan believes that the combination of corpus tools with literary research is corpus stylistics [2]. Although both definitions are somewhat general and do not define the research object from a micro perspective, they both reflect the characteristics of the combination of literary research and corpus linguistics. The corpus stylistics discussed in this paper is based on this characteristic, taking the language and structure of literary texts as the research object, and using literary work corpora and corpus analysis tools to statistically analyze the plot, vocabulary distribution, stylistic features, rhetoric, and other aspects of literary works, in order to expand the field of literary research with a combination of quantitative and qualitative research methods. Therefore, the corpus stylistics referred to in this study refers to the narrow sense of stylistics, that is, the application of corpus to the study of literary stylistics.

Traditional literary research starts from the reader's experience, conducting qualitative and interpretive analysis of the work. Corpus methods, based on probability, provide quantitative and descriptive statistical

https://doi.org/10.62852/cahr/2024/59

analysis of texts. The combination of the two can make text analysis more systematic and reliable, not only allowing analysis from hypotheses, concepts to examples but also starting from statistical evidence of the text, constructing hypotheses and theories through feature analysis. As Partington said, corpus technology can present us with unexpected language forms, and corpus analysis can show a large number of examples reflecting language phenomena [3]. Therefore, it can correct, refute, or reinforce the researchers' intuition. Powerful corpus indexing software can not only show the position, frequency, grammatical structure, left and right collocations, and semantic relationships and orientations of a word in the discourse [4], but also highlight the development of the discourse and the content of the topic. Corpus indexing can reveal the implicit structure of the text, stimulate imagination, test the appeal of the text to readers, and has strong objectivity [5]. In analyzing and interpreting the theme of the work, the writer's language characteristics, understanding and questioning or developing the research results of predecessors, the corpus method can provide us with reliable data support, thus avoiding the shortcomings of traditional literary criticism that only focuses on conceptual deduction or rigidly applying theories outside of literature [6]. Wikberg summarized the advantages of applying corpus stylistics to the analysis of literary works: Compared with the traditional qualitative discourse analysis from micro to macro, dynamic, case-by-case, and the interaction of meaning and form, the application of the corpus to the discourse analysis of literary works is the interaction of micro and macro, static, batch, and quantitative research from form to meaning. Therefore, corpus discourse analysis can provide a more detailed, specific, and in-depth description of literary works, providing a new set of effective methods and tools for literary research.

3. Research on corpus stylistics

As early as the 1960s, Milic used corpora to study the style of writers. In the 1980s and 1990s, with the rapid development of corpus linguistics, true corpus stylistics also developed, and the theory, principles, and research methods of corpus stylistics gradually improved, with an ever-expanding scope of research. In 2005 and 2007, the two International Corpus Linguistics Conferences held in the UK both had workshops titled "Corpus Research on Literary Language." In July 2009, at the "Corpus Linguistics and Literature" seminar held at Liverpool University in the UK, scholars fully expounded on the broad prospects for the application of corpus linguistics in literary criticism from different research perspectives. Bettina Fischer-Starcke conducted a corpus study on "Pride and Prejudice," proving that corpus linguistics has great potential for the analysis of literary text styles; Martin Wynne briefly reviewed the methodological insights of corpus linguistics for literary criticism and introduced various corpus stylistics research methods [7]. Michaela Mahlberg used corpus linguistic methods to identify the development of local discourse functions in Dickens' novels as a clue to the word clusters, identifying different degrees of flexible language patterns, and conducted a comparative study of the works of other 19th-century writers and Dickens' works. This study, which takes a new perspective on classic works, has become a typical example of corpus stylistics research. Louw used corpus tools to search for the collocations of the word "utterly" in the British poet Larkin's poem "First Sight," and based on the concept of semantic prosody, analyzed the ironic effect of the word, verifying the intuitive feelings in literary appreciation [8]. Hori compared the Dickens' works corpus with the 19th-century novel corpus (NCFWD) and other 19th-century writers' works corpora, and by systematically exploring the collocation patterns in Dickens' works, showed us the creative use of language in Dickens' works and revealed the stylistic effects of these creative words, thereby proving the significance of collocation research for literary stylistic analysis. Stubbs used corpus methods to study the language style of the novel "Heart of Darkness" [9]. He used part of the written language corpus in the BNC (about 1 million words) as a reference corpus, and used the prose discourse in LOB, Frown, FLOB, and Brown as the observation corpus. Through observation of the word frequency list, vocabulary distribution, and index line analysis, he found that the themes of the novel can be reflected not only by those words that indicate uncertainty and ambiguity but also by some repeatedly appearing phrases and collocations. Mahlberg used the word clusters in Dickens' novels as a clue to the development of local discourse functions, identifying different degrees of flexible language patterns, and found that text building blocks have a special function in the portrayal of characters in novels, providing a novel descriptive tool for literary text analysis. At the same time, she also proposed that text-driven research on word clusters is an important starting point for corpus

https://doi.org/10.62852/cahr/2024/59

stylistics research. This study, which takes a new perspective on classic works, has solidified the theoretical foundation of corpus stylistics and has become a typical example of corpus stylistics research.

Abroad, corpus stylistics research has developed rapidly, but the application of corpus linguistics in literary criticism in China is still in its infancy, with a relatively late start and a lack of systematic research in both breadth and depth. A few researchers, such as Liu Jing, Qian Yufang, L v Chang Hong, Lu Wei, Li Jin, Lang Jianguo, and Zhao Yong gang, have introduced quantitative research methods for style. Empirical researchers like Zhao Qiong and Du Ailing mainly use corpus software such as Wordsmith and Concordance to perform vocabulary searches on a specific foreign literary work, analyzing its thematic ideas or linguistic features. The research subjects are relatively singular, and there is a scarcity of more in-depth and systematic studies on the writing style of a particular author, or the works of authors from a specific era or literary movement. Therefore, exploring the analytical approaches and perspectives of literary research within the scope of corpus stylistics is of positive significance for expanding the avenues of literary research and promoting the development of corpus stylistics in China.

4. Corpus stylistics in literary analysis: perspectives and approaches

Looking at the research on the application of corpus stylistics to literary texts both domestically and internationally, they can be roughly divided into the following seven research areas based on different research perspectives: First, research on the themes and ideological content of works; Second, research on the linguistic expression of works; Third, research on the author's style; Fourth, verification or explanation of the applicability and effectiveness of a certain literary criticism theory; Fifth, horizontal comparative analysis of works by multiple authors or research on the characteristics of literary works from a certain period; Sixth, analysis of genre characteristics and differences in literary texts; Seventh, corpus annotation and analysis aimed at individual linguistic feature stylistic research.

With the help of the powerful functions of indexing software, we can carry out quantitative and qualitative analysis of literary works through the following approaches:

4.1. Text statistics

With the help of corpus software, researchers can perform the following statistics on texts: byte count (bytes), token count (tokens: the total number of words in the text), type count (types: the number of different word forms in the text), type/token ratio and standardized type/token ratio (which can indicate the lexical variation in the text), average word length (average word length) and one-letter words, two-letter words, three-letter words, etc. (which can indicate the length and difficulty of the text's vocabulary), sentence count (sentences), average sentence length and standard deviation of sentence length (which can indicate the length and difficulty of the text's sentences), paragraph count (paragraphs), average paragraph length and standard deviation of paragraph length (which can provide information on the length and difficulty of the text's paragraphs). These data can provide information on text difficulty, lexical variation, and the author's word usage characteristics, and fully prepare for thematic and plot analysis, linguistic feature analysis, rhetorical device retrieval analysis, and character portrayal retrieval and analysis of literary works. For example, Zhao Qiong conducted statistics on the children's literature "The Nightingale and the Rose," where the token count and sentence count indicate that the text is a relatively short literary work. The type/token ratio of 11.23 is significantly lower than the standardized type/token ratio of 33.65, suggesting that the text has little lexical variation and the use of words maintains a relatively stable state. With a total of 84 sentences in 61 paragraphs, it is known that the text mainly consists of dialogues. In addition, the statistics of one to five-letter words show that 83% of the entire vocabulary consists of words with fewer than five letters, indicating that the vocabulary difficulty of the text is relatively small. This information just verifies the characteristics of children's literature, which mainly focuses on the plot, has strong dialogic features, compact paragraphs, and concise and easy-to-understand language.

4.2. Word frequency and keywords

Word frequency is an important type of data that a corpus can provide. Word frequency statistics can offer an overview of the overall vocabulary distribution within a corpus. Frequency data can also reveal differences between multiple corpora. Word frequency helps researchers identify the most basic linguistic

https://doi.org/10.62852/cahr/2024/59

features that contain discourse meaning. By taking a high-frequency word as a node, analyzing its collocation distribution can reveal ideologies hidden in individual words, vocabulary, and grammatical structures, or new insights that traditional literary criticism might not discover.

Keywords are an important term in corpus linguistics. They differ from the traditional sense of "key words" or "important words." In corpus linguistics, keywords represent words that have an exceptionally high frequency of recurrence when compared to a reference corpus. Keyword analysis is significant in text retrieval, theme analysis, style analysis, and critical discourse because keywords can demonstrate ways of expressing information, and it is based on this information that researchers further interpret certain linguistic phenomena using qualitative methods. From the keyword list, one can understand the thematic plot of a work, providing a more scientific basis for story analysis.

For example, Yang Jian Mei used W concord to count the frequency of the first 244 most frequently occurring words in "The Cop and the Anthem," and from the meaningful content words, one can understand the general plot of the text. Using the Wordlist function in Wordsmith Tools to obtain the word list of "A Story of an Hour," after removing function words and auxiliary verbs, the most frequent content word is 'her' (43 times), followed by 'she' (34 times), and then 'it' (15 times), 'would' (8 times), 'open' (8 times), including 'opened' (2 times), 'door' (6 times), 'free' (5 times), 'days' (4 times), 'life' (4 times), 'Richards' (4 times). It can be seen that the article is narrated from the perspective of the third person, and the whole story unfolds with 'she,' that is, Mrs. Mallard, the female protagonist, as the main line, involving life and freedom.

4.3. Lexical collocations and semantic prosody

There are various definitions given by scholars regarding lexical collocations.

Firth defined collocation as "words which tend to co-occur habitually" [10]. Hunston considers collocation to be "the tendency of words to occur together". Scott uses the term "lexical bundle" to denote the fixed pattern of words that co-occur to the left and right of a particular word. Corpus software can retrieve collocational words within different spans to the left and right of a word and count their frequencies, thus collocations can display the context of a word and its customary structure [11]. Louw once proposed analyzing stylistic effects through collocation studies. Hiurgfori also believes that studying the high-frequency collocation patterns in an author's works can not only reveal the author's collocation preferences but also reflect the individualized language of characters. Quantitative statistical analysis of collocations in literary works can reveal the author's word usage characteristics, identify the styles of different authors or the personalities of different characters in the same novel, and distinguish the similarities and differences between the styles of parodies and the original works. Hori's exploration of the lexical collocation patterns in the Dickens corpus demonstrated Dickens' creative use of language, providing readers with a new and deeper understanding of Dickens' linguistic style.

Collocation analysis can be extended to the study of semantic prosody. Semantic prosody is a special phenomenon of lexical collocation. The study of semantic prosody allows us to link the formal collocation of language with semantics. Generally, only words with certain semantic characteristics can be attracted to a specific node word and co-occur in the same context, creating a special semantic atmosphere throughout the entire context. Semantic prosody can express the speaker's attitude; the study of semantic prosody can reveal the atmosphere of the context set by the author and their attitude towards the characters portrayed. For example, a collocation search for the word "Soapy" in "The Cop and the Anthem" reveals that the adjectives that collocate with it are mostly words with negative semantic prosody, such as "dead," "humble," "callous," "disconsolate," "dreadful," and "degraded." From these words, we can see that the author is trying to portray a character who is lowly in status, poor in economy, miserable in fate, lonely, and degenerate in an indifferent social environment. However, in specific linguistic communication environments, for certain special purposes and to achieve certain linguistic communication effects, people will deliberately defy conventions, violate the principle of collocation harmony, and choose some semantically conflicting words to form unusual or abnormal collocations, creating semantic conflict (prosodic clash). The use of semantic conflict can create special communication effects and produce rhetorical figures such as irony, metaphor, and hyperbole in rhetoric. Therefore, the study of semantic prosody and semantic conflict is conducive to understanding the author's emotional attitudes and the use of rhetorical techniques, thereby gaining a deeper understanding of the ideological content of the work. Taking the collocation words of "Soapy" in "The Cop

https://doi.org/10.62852/cahr/2024/59

and the Anthem" as an example, in addition to most of the collocation words being negatively prosodic, the author also used non-negative words such as "cognizant," " hibernation," "desirable," and "benign." Further observation of the context reveals that the author is using irony, intending to show that Soapy is very aware that he should find a suitable place to spend the winter, and the place he longs for is prison, because in his mind, prison is better than a charity institution. It can be seen that in the negative context created by the author, the use of these non-negative words creates the satirical effect of the novel.

4.4. Index line analysis

Index lines can provide frequency information of search terms within a corpus, and can display collocational words as well as the context of the term, which is the most basic function of index lines. Whether it's keywords or lexical collocations, what they provide is only quantitative linguistic patterns, and researchers must further interpret to answer specific research questions. Since frequency itself is just a numerical value and cannot explain the reasons for high or low frequencies, and lexical collocations can only provide limited context information, researchers can only fully and effectively utilize keywords and collocations by accurately and effectively interpreting the index lines, conducting qualitative analysis of quantitative linguistic patterns, and thus seeing through the surface language phenomena to the deep essence of discourse.

Corpus Stylistics offers us a fresh perspective for understanding and appreciating literary works. Currently, the use of corpus tools for the study of literary texts in China is still in its infancy. Making full use of powerful corpus software to conduct multi-angle analysis on a large number of complete literary texts, and to carry out literary criticism that is more authentic, rational, and objective, represents the future development trend of Corpus Stylistics. Therefore, the further research and development of large-scale, stylistically comprehensive, and specialized literary corpora that can code various elements such as parts of speech, syntax, semantics, and discourse in as much detail as possible, and the creation of indexing software suitable for multi-level, multi-angle retrieval and analysis of literary language, are the challenges that researchers currently face.

5. Conclusion

This paper has explored the rise and development of Corpus Stylistics, an emerging field that combines methods of corpus linguistics with stylistic theory to provide new perspectives and methods for the analysis of literary texts. Compared with traditional qualitative analysis, Corpus Stylistics uses computer-aided quantitative analysis methods to systematically and scientifically interpret the linguistic characteristics of literary works. The article reviews the definition, advantages, research progress, and specific application approaches of Corpus Stylistics in literary analysis, including text statistics, word frequency and keyword analysis, lexical collocation and semantic prosody analysis, and index line analysis. Through these methods, researchers can reveal the deep linguistic structures and stylistic features of literary works, enhancing the objectivity and depth of literary criticism. Although Corpus Stylistics research in China is still in its early stages, its potential to promote innovation in literary research methods and theoretical development is enormous. Future research needs to further develop and improve specialized literary corpora and analysis tools to promote in-depth development in this field.

6. References

[1] Hardy, Donald. Corpus Stylistics as a Discovery Procedure [C]//Greg Watson, Sonia Zyngier. Literature and Stylistics for Language Learners: Theory and Practice. New York: Pal- grave Macmillan,2007:79-89.

[2] Toolan, Michael. Narrative Progression in the Short Story: First Steps in a Corpus Stylistic Approach [J]. Narrative, 2008(2):105-120

[3] Partington, Alan. The Linguistics of Political Argument [M]. Routledge,2003:12.

[4] Hoey, Michael. Patterns of Lexis in text [M]. Shang Hai: Shanghai Foreign Language exhalation press,2000:138.

[5] Yang Jian Mei. Corpus Index Analysis of "The Cop and the Anthem" [J]. Journal of Sichuan International Studies University, 2002(3): 56-59.

https://doi.org/10.62852/cahr/2024/59

[6] [6] Yang Huizhong. Introduction to Corpus Linguistics [M]. Shanghai: Shanghai Foreign Language Education Press, 2002: 242-243.
[7] Wynne, Martin. Stylistics: Corpus Approaches [J].0xford: Elsevier,2006(12):223-226.
[8] Louw, Bill. Irony in the text or insincerity in the writer? - The diagnostic potential of semantic prosodies [A]//Baker, Mona. G. Francis, E. Tognin Bonelli. Text and Technology: In honour of John Sinclair. Amsterdam: John Benjamins Publishing Company,1993:157-176.
[9] Stubbs, Michael. Conrad in the computer: examples of quantitative stylistic methods [J]. Language and Literature,2005, 14(1):5-24.
[10] [10] Qian Yufang. Corpus and Critical Discourse Analysis [J]. Foreign Language Teaching and Research, 2010, 42(3): 198-202.
[11] Hunston, Susan. Corpora in Applied Linguistics [M]. Cambridge: CUP,2002:68.

https://doi.org/10.62852/cahr/2024/59