

Literary Style Quantitative Analysis Based on NLP: Taking Dickens and Austen as Examples

Bingxian Chen, Jingjing Wu*

Guilin University of Technology, Guilin, 541006, China

*Corresponding author: 2112843513@qq.com

Abstract: *Literary style research has always relied on researchers' subjective perceptions and qualitative descriptions, lacking objective quantitative evidence. In order to make up for this methodological limitation, this study introduced natural language processing technology to conduct a cross-field quantitative comparative analysis of the literary styles of Dickens and Austen, two representative writers. The study constructed a digital corpus of representative works of the two writers, and systematically extracted quantitative features from four dimensions: vocabulary richness, syntactic complexity, emotional distribution and topic model. Through methods such as word frequency statistics, keyword analysis, syntactic analysis, sentiment calculation, and LDA topic modeling, the systematic differences in language use between the two writers are revealed in a data-driven manner. This study aims to use accurate data to verify the core assertions about the styles of the two writers in traditional literary criticism, thereby proving the effectiveness and scientificity of the NLP measurement method as a useful supplement to traditional literary research, and providing an empirical path that can be used as a reference for literary criticism from the perspective of digital humanities.*

Keywords: *natural language processing, quantitative analysis, Dickens, Austen*

1. Introduction

1.1 Research background

The practice in the field of digital humanities can be traced back to humanities computing in 1949. It was first proposed by Roberto Busa of Italy, known as the father of humanities computing. He used a punch card computer to conduct a lexical index of the entire works of Thomas Aquinas. This is regarded as a landmark beginning of digital humanities practice. After forty years of development, humanistic computing has become too technical. As research methods and categories have expanded from simple text analysis to broader aspects such as cultural criticism, data visualization, and digital publishing, the old name of humanistic computing can no longer cover its connotation. Therefore, when planning a new companion series "Digital Humanities Companion" from Blackwell Publishing, scholars such as John Ainsworth, Susan Hawkey and Ray Simon jointly discussed and finally decided on the name digital humanities. This is the first academic book to explicitly use Digital Humanities in its title. It is the first time that digital humanities has been introduced into the academic world as a formal subject name, marking the official entry of this term into the academic world and beginning to systematically replace humanities computing. Digital humanities is not a single discipline but a thriving interdisciplinary field. It aims to fundamentally reshape the way humanities inquiry is conducted, integrating computational technologies with research methods from the humanities to ask and answer new questions that traditional humanities methods can neither solve nor raise. It is not only a methodology for using computing technology to expand the realm of humanities research, but also a reconstruction and reflection of the humanities in the digital age [1].

Quantitative stylistics is a discipline that uses statistical and mathematical methods to quantitatively measure, analyze and model the linguistic style characteristics of literary works. It is a cross-application of quantitative linguistics and computational linguistics in the field of literary research[2]. The origins of metrological stylistics can be traced back to the 19th century. Its development was driven by disputes over the authorship of literary texts. Scholars began to try to go beyond subjective impression criticism and use objective, quantifiable evidence to support their opinions. Polish scholar Wincenty Lutosławski first used the term Stylometry in 1890 while studying the chronology of Plato's works. He established the order of Plato's dialogues by analyzing and

counting syntactic features such as clause proportions, which is regarded as the germination of the discipline consciousness of quantitative stylistics. With the in-depth application of statistical technology and the introduction of computing technology in the mid-20th century, scholars can use more mature statistical methods to process larger-scale statistical data, pushing econometric style to a new height. Today, metrological stylistics has developed from an idea based on simple word length statistics to a mature interdisciplinary subject that integrates statistics, computer science and traditional philology [3].

1.2 Research questions

The research constructed a digital corpus of representative works of Austen and Dickens, and extracted quantitative features from four dimensions: vocabulary richness, syntactic complexity, emotional distribution and topic model. At the level of vocabulary richness and characteristic words, four indicators are used for analysis: vocabulary density, type ratio, high-frequency words and keywords. Lexical density refers to the proportion of content words in the total vocabulary. A high ratio usually indicates a large amount of information and a concise writing style. The type ratio refers to the ratio of the number of unique words to the total number of words. The higher the ratio, the richer the vocabulary. In the high-frequency word list, the 20 most commonly used words of each writer are listed, and stop words such as the, a, and and other words are removed to further analyze the writer's writing style. In keyword analysis, statistical tests are used to identify words that are significantly overused by a writer relative to a large reference corpus and that may explain the writer's writing style.

At the level of syntactic complexity, three indicators were selected for study: average sentence length, clause ratio, and frequency of use of specific sentence patterns. The average sentence length is the most basic indicator. Long sentences often mean complex ideas or gorgeous writing style, while short sentences often mean a more concise and powerful writing style. The clause ratio refers to the ratio of clauses to the total number of sentences in a sentence. A high clause ratio means that the author is inviting readers to engage in contemplative and analytical reading. It is suitable for expressing complex social relationships and philosophical speculations, and is often used in realism and stream-of-consciousness writing styles. The writer's writing style can be seen from the frequency of use of specific sentence patterns, such as passive voice, questions, exclamations and other special sentence patterns. Romantic poets may use more exclamations to express strong emotions.

In sentiment analysis, emotional word density determines the emotional narrative strategy of a work by using a sentiment dictionary to calculate the ratio of positive to negative emotion words in a text. In the emotional trajectory, a curve can be drawn by calculating the emotional value score. From this, we can analyze the emotional changes of the work as well as the deep structure of the narrative, the reader's psychological manipulation and the writer's philosophical rhythm. The emotional volatility index reflects the emotional changes in the narrative process. In text emotional analysis, it can reveal the narrative motivation, deep structure and author's psychology, and ultimately analyze the overall emotional tone of the work. It is also a macroscopic ruler for analyzing the writer's literary style.

In the topic modeling dimension, this study plans to use the LDA topic model to analyze the latent topic structure in the works of Dickens and Austen. By counting the co-occurrence relationships of words in texts, we can identify recurring theme word groups in different works, and further compare the differences in theme distribution and bias between the two writers. In this study, after completing text cleaning, word segmentation, stop word removal and lemmatization, we will use novel chapters or text fragments as the analysis unit to extract several potential themes through the LDA model. Subsequently, the topics were manually named and literary explained based on the keywords with higher weight under each topic and the context of the original text. During the comparison process, we focused on examining whether Dickens's works focus more on themes such as urban space, poor classes, social systems, growth experiences, and moral redemption, and whether Austen's works focus more on themes such as marriage and love relationships, family ethics, property inheritance, social etiquette, and female growth. Through this dimension, this article hopes to reveal the differences in thematic concerns of the two writers from a macro level, and combine the statistical results of thematic models with text close reading in traditional literary criticism, thereby avoiding interpretation bias caused by relying solely on model output.

1.3 Research significance

The significance of this study lies in the following three levels. First of all, the use of data science

methods provides objective, systematic and quantitative evidence for traditional literary qualitative descriptions of Dickens and Austen, partially transforming style criticism from an art that relies on personal knowledge and perception into an evidence-based science, and enhances the persuasiveness and communicability of literary criticism. Secondly, quantitative analysis reveals hidden patterns and structures in the text, in order to produce new knowledge and help the public understand classic works more deeply. Finally, this study is an interdisciplinary paradigm exploration that provides a concrete example of how to effectively apply natural language processing technology to humanities research, aiming to promote literary research towards a new path that combines humanistic insights with scientific rigor.

2. Literature review

2.1 Research basis of digital humanities and quantitative stylistics

The rise of digital humanities provides new methodological perspectives for literary research. Traditional literary criticism mainly relies on researchers' close reading of texts, aesthetic judgment and empirical interpretation, which can deeply reveal the ideological connotation and artistic characteristics of works. However, it has certain limitations in processing large-scale texts, cross-writer comparisons and identification of implicit language patterns. With the development of computer technology, corpus linguistics and natural language processing technology, literary research has gradually shifted from a single qualitative interpretation to a comprehensive research path that combines qualitative and quantitative research.

Zhu Benjun and Nie Hua pointed out that digital humanities originated from humanistic computing, and its research objects have expanded from early electronic texts and vocabulary indexes to multiple levels such as text processing, knowledge discovery, data set construction, and visual presentation [1]. For literary research, digital humanities does not replace traditional close reading with technology, but uses computational tools and large-scale corpora to enable researchers to observe phenomena such as style, theme, emotion, and narrative structure within a larger text range. Research on computational grounded theory by Fan Wenhao and others also emphasized that digital humanities research should adhere to the principle of algorithms being supplemented and researchers being the mainstay, that is, algorithms can improve data processing efficiency and pattern recognition capabilities, but the final interpretation of literary meaning still needs to be completed by researchers based on the text context, historical background and theoretical framework [4].

Quantitative stylistics is an important research path developed in this digital humanities context. Zeng Yiping and Zhu Xiaowen pointed out that after computational methods entered the study of Chinese stylistics, they were initially mainly used for author verification and research on the language style of writers' works, and later gradually expanded to fields such as mathematical model construction, language ontology research, and information science research [3]. Li Yiyun and Niu Yunping believe that quantitative stylistics is a discipline that uses quantifiable linguistic features in the text to study the text style and author's writing habits in a quantitative manner. Its basic premise is that the language form of the text can reflect the language habits of the author or translator to a certain extent [2]. It can be seen that quantitative stylistics can not only provide objective data support for traditional literary criticism, but also reveal text style differences through multi-level indicators such as vocabulary, syntax, emotion and theme.

2.2 Main results of research on literary text measurement style

Existing research has shown that literary style can be characterized by various indicators such as vocabulary, syntax, word categories, function words, punctuation, high-frequency words, keywords, text readability, N-grams, cluster analysis and classification models. Liu Ying and Xiao Tianjiu, in their study on the quantitative stylistics of "A Dream of Red Mansions", compared the first eighty chapters with the last forty chapters from the aspects of word length distribution, degree of change in word length, word categories, function words, high-frequency words and unique words, and found that there are obvious differences between the two in terms of one-character words, two-character words, nouns, verbs, adverbs, particles, modal particles and average word length [5]. This study illustrates that stylistic differences in literary works are not only reflected in plot arrangements and characterization, but also in the frequency distribution of microscopic language forms.

In terms of research on martial arts novels, Liu Ying and Xiao Tianjiu conducted a stylistic

comparison between Jin Yong and Gu Long's novels, examining indicators such as average paragraph length, word length dispersion, content words, function words, high-frequency words, parts of speech, and punctuation marks. They pointed out that Jin Yong's novels pay more attention to family and country responsibilities and scene details, while Gu Long's novels are more inclined to individual feelings and atmosphere rendering [6]. Xiao Tianjiu and Liu Ying further introduced methods such as text clustering, text classification, principal component analysis, word and part-of-speech N-gram, punctuation N-gram and other methods, proving that Jin Yong and Gu Long have significant differences in grammatical structure, phrase structure, text rhythm, readability and degree of language change [7]. These studies indicate that writer style comparison should not rely on a single indicator, but should verify each other through multidimensional characteristics.

In addition to Chinese novels, translated texts and foreign literary works have gradually become important research objects of quantitative stylistics. Zhang Jiayi and Li Liangbo took Hemingway's "A Moveable Feast" as the research object and analyzed the style of the work from the dimensions of average word length, type ratio, keywords, single occurrence words, average sentence length and sentence dispersion, revealing its concise language, loose structure and memoir-style narrative characteristics [8]. Li Yiyun and Niu Yunping took the English translation of the Diamond Sutra as the object, analyzed the differences in translation styles from three levels: vocabulary, sentences and discourse, and further explained the causes of differences from the time and space field, religious field, cultural field and reader field [2]. Wu Jingjing and Xu Juanjuan analyzed the attitude, intervention and differential resources in the original poem and the English translation of "Ode to Burial Flowers" based on evaluation theory, explaining that emotion, evaluation and cultural meaning can be revealed through statistics and comparison of discourse resources [9]. Overall, the above research provides an important methodological reference for this article, that is, literary style can be quantified through multi-level linguistic features, but the measurement results still need to be combined with the interpretation of literary meaning.

2.3 Research on the style of Dickens and Austen and the entry point of this article

Dickens and Austen are representative novelists in the history of British literature. There are obvious differences in their creative objects, narrative space and language styles. Traditional literary criticism usually believes that Dickens pays more attention to social reality, bottom life, urban space and moral redemption, and his works have strong realist concerns and dramatic narrative features. Austen pays more attention to family ethics, marriage and love relationships, social etiquette and character psychology. His works are known for their rationality, restraint, irony and delicate psychological description.

Regarding Dickens, Fu Shouxiang pointed out when discussing "A Tale of Two Cities" that Dickens's works embody the strong conscience and benevolent spirit of the times, and his novels reveal social contradictions, class oppression and humanitarian concerns through historical backgrounds such as the French Revolution [10]. This study illustrates the social critical consciousness of Dickens' novels from the perspective of ideological content and value stance, and also provides a literary background for this article to analyze Dickens' works from the perspective of theme models and emotional fluctuations. Regarding Austen, Hui Haifeng used 76 British novels from 1768 to 1830 as corpus under the framework of digital humanities, and used the NRC emotional dictionary to examine the emotional writing of Austen's novels. Research has found that the frequency of negative emotions such as anger, fear, and sadness in Austen's novels is generally low, while dimensions such as joy and expectation are relatively prominent, forming an emotional structure with judgment, reflection, and relationship adjustment as the core [11]. This study transforms traditional criticism's judgments about Austen's emotional restraint into testable data issues, which has direct reference significance for the emotional analysis part of this article.

Overall, existing research has laid a solid theoretical and methodological foundation for this article, but there are still certain shortcomings. First of all, domestic quantitative stylistic research mostly focuses on Chinese classical novels, martial arts novels, and Buddhist scriptures in English translation and translated texts, while relatively little attention is paid to cross-text and cross-style comparisons between British classic novel writers. Secondly, some studies focus on author attribution judgment or classification effect, and there is still room for further deepening the interpretation of the literary meaning behind the data differences. Finally, existing research focuses more on surface language features such as word length, sentence length, word categories, function words, and punctuation, while the comprehensive analysis of higher-level features such as emotional trajectory, topic structure, and narrative focus is insufficient.

Based on this, this article takes the representative novels of Dickens and Austen as the research object, comprehensively uses natural language processing and quantitative stylistic methods to conduct a comparative analysis from the four dimensions of vocabulary, syntax, emotion and theme. On the one hand, this article draws on methods such as word frequency statistics, keyword analysis, sentence length analysis, and text readability analysis from existing research. On the other hand, it further introduces affective computing and LDA topic models to reveal the systematic differences between the two writers in language use, narrative structure, emotional expression, and topic focus. Compared with traditional literary criticism, this article hopes to use quantitative data to provide empirical support for the classic judgment of the two writers' representative works. Compared with purely technical research, this article emphasizes putting statistical results back into the context of literary history and textual reading for interpretation, thereby establishing a more effective connection between data-driven and humanistic interpretation.

3. Research methods

3.1 Research design

This study uses natural language processing and quantitative stylistics as the theoretical and technical basis to conduct a quantitative comparative study on the literary styles of Dickens and Austen, and systematically analyzes the text styles of the two writers from the four dimensions of vocabulary, syntax, emotion and theme. This study attempts to use data-based methods to break down literary style into multiple computable indicators to improve the objectivity, systematicness and verifiability of literary style research. It uses natural language processing technology to discover potential patterns in the text, and then combines literary theory with interpretation.

3.2 Corpus composition

In order to conduct a systematic comparative analysis of the literary styles of Dickens and Austen, this study constructed a corpus. Four core novels were selected from the representative works of the two writers to ensure the breadth and depth of analysis. The Dickens group selected "Great Expectations," "Oliver Twist," "A Tale of Two Cities" and "David Copperfield," while the Austen group selected "Pride and Prejudice," "Emma," "Sense and Sensibility," and "Mansfield Park." The selection of the above works is based on the following core considerations. First of all, the selected works are the most widely known and most academically studied core texts of the two writers, which can best represent their respective typical styles. "Great Expectations" and "Pride and Prejudice" are recognized as the pinnacles of their respective creative careers. Secondly, the selected works cover the core creative concerns of the writer and cover different narrative themes and creative types, which helps to reduce the bias caused by a single subject on the research results. These works span the main creative periods of the two writers and can more comprehensively reflect the literary styles of their mature periods. All text is sourced from the Project Gutenberg database. The platform provides high-quality, open-copyright English literary electronic texts with high reliability and accessibility. In order to ensure text consistency, this study uniformly uses plain text format as the corpus source.

3.3 Text preprocessing

Before formal analysis, the corpus was standardized to improve the accuracy and consistency of subsequent natural language processing tasks.

First, clean up the format of the text, delete the copyright statement, table of contents, publishing instructions, headers and footers and other non-text content that automatically comes with Project Gutenberg, leaving only the main text of the novel. The text will be unified and standardized, all English characters will be uniformly converted to lowercase, special symbols and redundant spaces will be deleted, and the punctuation mark format will be unified. This study uses natural language processing tools to segment the text into words and sentences, providing basic linguistic units for subsequent analysis. In the lexical analysis stage, this study further carried out stop word filtering and lemmatization. Stop words mainly include articles, prepositions, conjunctions and other high-frequency function words, such as the, a, and, etc. The lemmatization principle is used to unify vocabulary with different grammatical forms and improve the consistency of word frequency statistics. This study also uses part-of-speech tagging and dependency syntax analysis to identify grammatical structures in the text to support subsequent syntactic complexity analysis.

3.4 Analysis dimensions and measurement indicators

This study quantifies style from four calculable linguistic dimensions: lexical dimension, syntactic dimension, emotional dimension, and thematic dimension.

3.4.1 Lexical Dimension

The lexical dimension is mainly used to measure the richness and expression tendency of the writer's language use. The following indicators are mainly used. This study uses the lexical density index to calculate the proportion of content words in the total vocabulary, so as to measure the information-carrying capacity and linguistic conciseness of the text. It also calculates the ratio of the number of unique words to the total number of words, namely the type-token ratio, to measure the lexical richness of the text. By counting the high-frequency words in the corpus and outputting the top twenty words, we can analyze the writer's narrative focus. Finally, the log-likelihood test is used to identify significant high-frequency words to discover the writer's identifiable stylistic features.

3.4.2 Syntactic dimension

The syntactic dimension is mainly used to analyze the complexity of text structure and narrative organization. The following indicators are mainly used. The average sentence length is used to measure the complexity of the text's syntactic structure. The longer the average sentence length, it often indicates that the writer's thoughts are complex or his writing style is more gorgeous; the shorter the average sentence length, it often indicates that the writer's writing style is more concise and powerful. The ratio of clauses to the total number of sentences in a text is counted, that is, the clause ratio, to analyze the writer's writing style. Finally, count the number and proportion of special sentence patterns in the works, including questions, exclamations, passive voice, direct quotations and other sentence patterns to analyze the narrative expression characteristics of different writers.

3.4.3 Emotional dimension

The emotional dimension is mainly used to analyze the emotional expression and narrative rhythm in the text. The VADER sentiment dictionary is mainly used to perform sentiment analysis on texts, and the following indicators are used. Emotional vocabulary density refers to the statistical proportion of positive and negative emotion words used. Then, by calculating the overall emotional score of the text, the overall emotional tendency of the work is analyzed. That is, the overall emotional tone of the work. This study divides the text into multiple segments and calculates the emotional score of each segment to construct an emotional change curve, which represents the emotional trajectory of the work. Through the change range of the sentiment score, the sentiment volatility is calculated to measure the stability and drama of the text's sentiment structure.

3.4.4 Topic Dimension

In order to reveal the differences in narrative content and focus between the two writers, this study uses the LDA topic model for topic modeling analysis. LDA is a typical unsupervised machine learning model. By analyzing word co-occurrence relationships, the model can automatically identify potential topic structures in text. In the study, we first extracted the core themes of the two writers, and then compared the differences in theme distribution between different works, thereby analyzing the different tendencies of the two writers in narrative focus.

3.5 Technical Tools and Processes

This study uses Python as the core programming language and relies on its powerful natural language processing ecosystem to implement analysis.

The third-party libraries used mainly include NLTK and SpaCy, used for word segmentation, part-of-speech tagging, lemmatization and syntactic analysis; Pandas and NumPy, used for data processing and statistical analysis; Gensim, used for LDA topic modeling; VADER, used for sentiment analysis; Matplotlib and Seaborn, used for data visualization.

4. Data analysis and discovery

4.1 Comparison of vocabulary features

After completing text preprocessing, stop word filtering, and named entity cleaning, this article

further conducted a lexical dimension analysis on the Austen and Dickens novel corpus. In order to avoid the interference of character names on the statistical results, on the basis of traditional stop word filtering, SpaCy named entity recognition technology was additionally used to filter the character names and proper nouns in the novel, so that the vocabulary statistical results can more truly reflect the language style and theme expression characteristics of the two writers.

From the perspective of vocabulary size indicators, the total number of words and the number of unique words in the Dickens corpus are significantly higher than that of Austen. The results show that the total effective vocabulary of Dickens's works is close to 300,000 words, while that of Austen is about 180,000 words. Dickens also ranks higher than Austen in terms of the number of unique words. This shows that the overall length of Dickens's works is longer, the narrative scenes are more complex, and the language coverage is wider, reflecting a strong ability to describe social reality and multi-level narrative characteristics. In contrast, the vocabulary of Austen's works is relatively small, reflecting that its text structure is more concentrated, the narrative space is relatively closed, and it prefers to describe family, marriage, and social relationships.

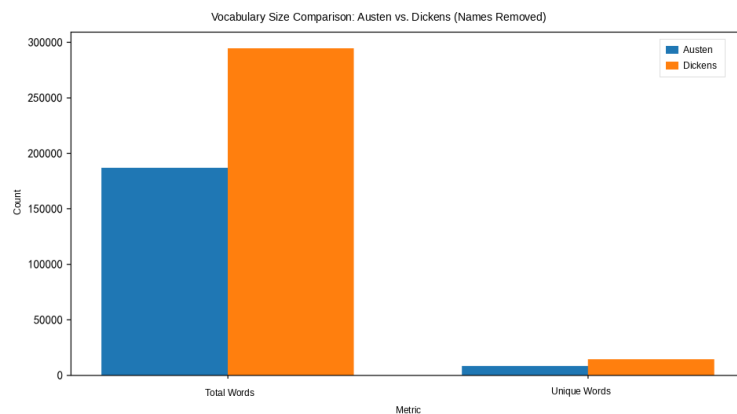


Figure 1 Comparison of vocabulary size between Austen and Dickens

As shown in Figure 1, Dickens's corpus contains more total words and more unique words than Austen's corpus. This result indicates that Dickens's novels cover a broader lexical range and present more extensive narrative scenes, whereas Austen's corpus is relatively compact and concentrated on domestic and social relationships.

In terms of lexical complexity, the typology ratio (TTR) of both writers remains at a low level, but Dickens is slightly higher than Austen, indicating that his lexical diversity is slightly stronger. Since this study has carried out stop word filtering before vocabulary statistics, the vocabulary density indicators are all close to 1, which shows that the retained vocabulary is mainly content words with actual semantic functions. In general, Dickens is more prominent in terms of vocabulary richness and language variation, while Austen presents a more stable, concentrated, and repetitive language pattern.

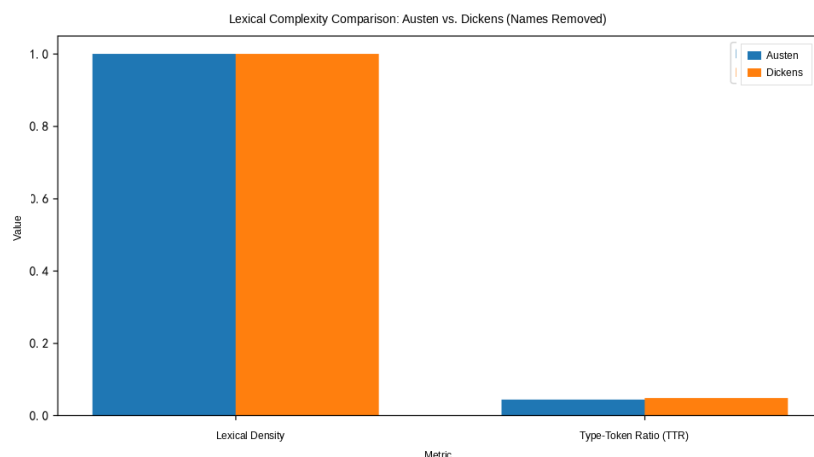


Figure 2 Comparison of vocabulary complexity between Austen and Dickens

Figure 2 further compares lexical density and the type-token ratio (TTR). Both writers show high lexical density after stop-word removal, but Dickens has a slightly higher TTR, suggesting that his

corpus has marginally greater lexical diversity.

High-frequency word statistics further reflect the differences in narrative focus between the two writers. High-frequency words in Austen's works mainly include feel, wish, dear, friend, sister, etc. Most of these words are related to emotional expression, interpersonal relationships and family social interaction, reflecting that Austen's novels pay great attention to the psychological activities of characters, emotional changes and family ethical relationships. Among them, words such as feel and wish appear frequently, which shows that Austen pays special attention to the inner feelings and emotional expression of the characters, and his narrative style has obvious emotional and introspective characteristics. In contrast, the high-frequency words in Dickens's works are more concentrated in specific entity words such as hand, head, face, house, room, and boy. Such words are often closely related to character actions, body descriptions, social spaces and real environments, reflecting that Dickens' novels place more emphasis on visual scene construction and description of social reality. For example, the frequent occurrence of words for body parts such as hand, face, head, etc. shows that Dickens is good at shaping characters through description of their actions and appearance. The spatial words such as house and room reflect the strong environmental narrative characteristics of his works.

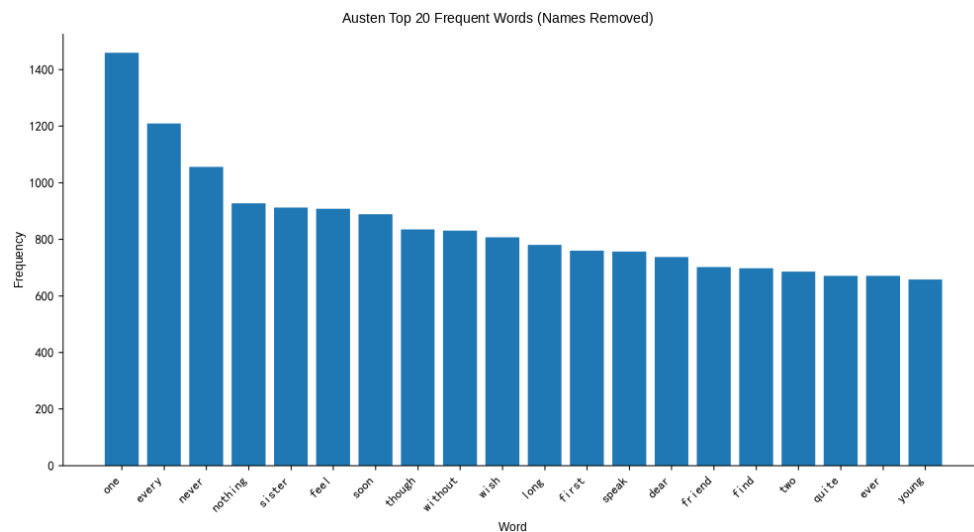


Figure 3 Top 20 high-frequency words in Austen

Figure 3 lists the twenty most frequent words in Austen's corpus after preprocessing. These high-frequency words are mainly associated with feelings, family interaction and social judgment, revealing Austen's concern with character psychology and interpersonal relationships.

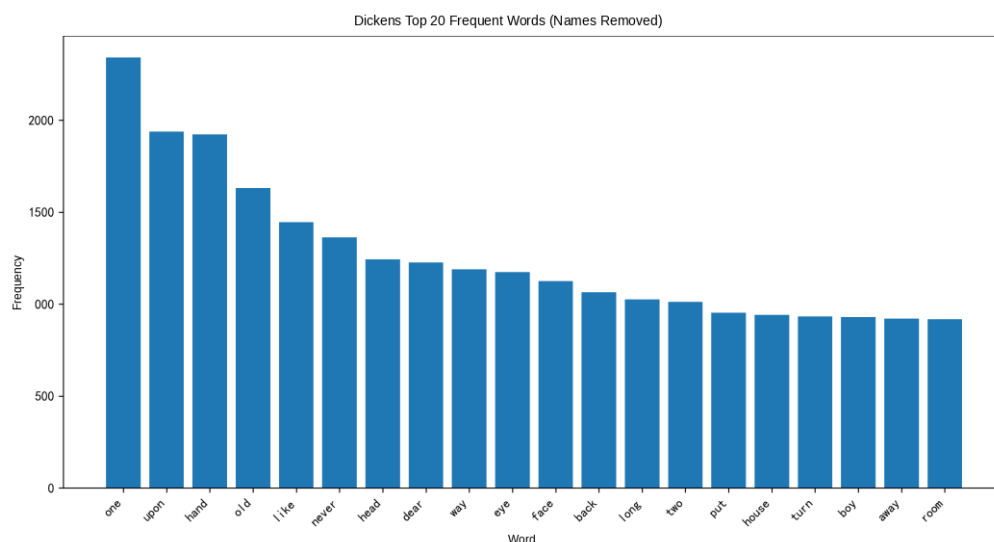


Figure 4 Top 20 high-frequency words in Dickens

Figure 4 shows the twenty most frequent words in Dickens's corpus. The frequent occurrence of

words related to body, space and action indicates that Dickens tends to construct narrative scenes through concrete details and visible social environments.

The linguistic differences between the two writers were further verified in the analysis of significant keywords. Significant keywords in Austen's corpus include feeling, wish, pleasure, dear, etc. These words embody thematic features of emotional expression, interpersonal interaction, etiquette and social interaction, indicating that Austen's works are more inclined to delicate descriptions of emotional relationships and psychological levels. At the same time, the prominent keywords in Dickens's corpus are more concrete words such as hand, face, head, boy, etc., reflecting that his works pay more attention to social reality, character behavior and the living conditions of the bottom, and have strong realist literary characteristics.

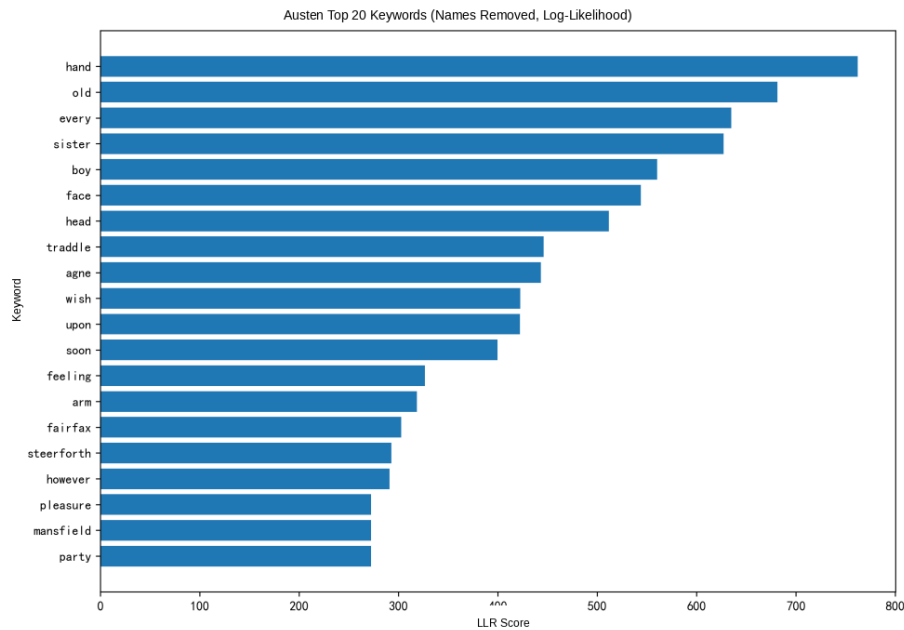


Figure 5 Top 20 significant keywords in Austen

Figure 5 presents Austen's significant keywords identified through log-likelihood analysis. The semantic focus on emotion, pleasure, family and social etiquette supports the view that Austen's style is restrained, introspective and relationship-oriented.

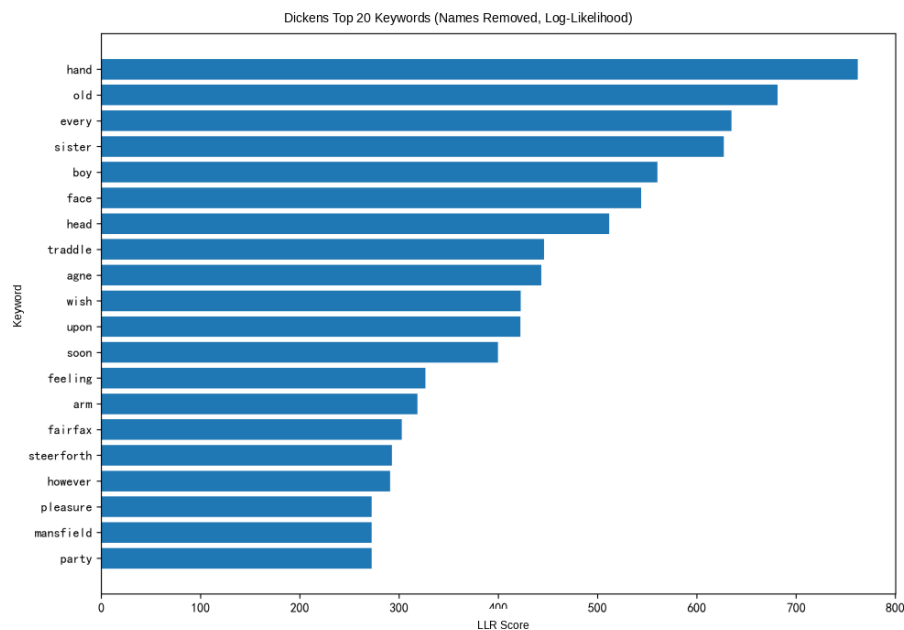


Figure 6 Top 20 significant keywords in Dickens

Figure 6 presents Dickens's significant keywords. The prominence of concrete nouns and social-space words shows that Dickens's language is more closely associated with action description, social reality and dramatic scene construction.

Overall, the lexical dimension analysis shows that there are significant differences in language use between Austen and Dickens. Austen is more inclined to use words related to emotion, family and society, forming a narrative style that is delicate, introspective and emphasizes character relationships. Dickens, on the other hand, pays more attention to the description of realistic environments, character actions and social scenes, and his language style is more concrete, dramatic and social. This difference at the vocabulary level further reflects the fundamental differences between the two writers in their literary creation concepts and narrative focus.

4.2 Comparison of syntactic features

This study further conducts a quantitative analysis of the syntactic complexity of Austen and Dickens corpus. Statistical results show that the Austen corpus contains a total of 25,052 sentences, a total of 564,161 words, and an average sentence length of 22.52 words; the Dickens corpus contains a total of 43,039 sentences, a total of 848,925 words, and an average sentence length of 19.72 words. Overall, Austen's sentence length is significantly longer than that of Dickens, which shows that her narratives tend to use compound sentences with longer structures and more complex levels, while Dickens tends to use relatively short sentence patterns to enhance narrative rhythm and scene advancement.

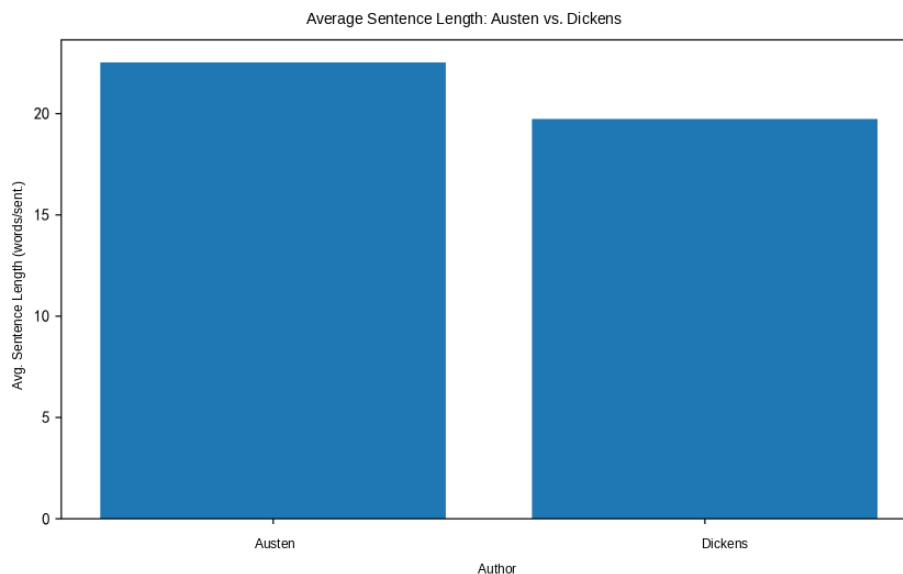


Figure 7 Comparison of average sentence length between Austen and Dickens

As shown in Figure 7, Austen's average sentence length is higher than Dickens's. This suggests that Austen more often uses longer and more layered syntactic structures to express psychological judgment and social evaluation, while Dickens tends to use shorter sentences to strengthen narrative rhythm and plot movement.

In terms of clause complexity, the sentences containing clause structures accounted for 68.27% of the total sentences in the Austen corpus, and each sentence contained an average of 1.8142 clauses; the sentences containing clauses in the Dickens corpus accounted for 65.62%, and the average sentence contained 1.5995 clauses. The results show that Austen has a higher degree of nestedness and complexity in syntactic structure, and its extensive use of object clauses, adverbial clauses and attributive clauses makes the text more suitable for presenting characters' psychological activities, emotional judgments and delicate social relationship analysis. In contrast, although Dickens also has a high usage rate of clauses, his overall syntactic structure is more biased towards dynamic narrative and event advancement, and his syntactic organization is relatively more direct and action-oriented.

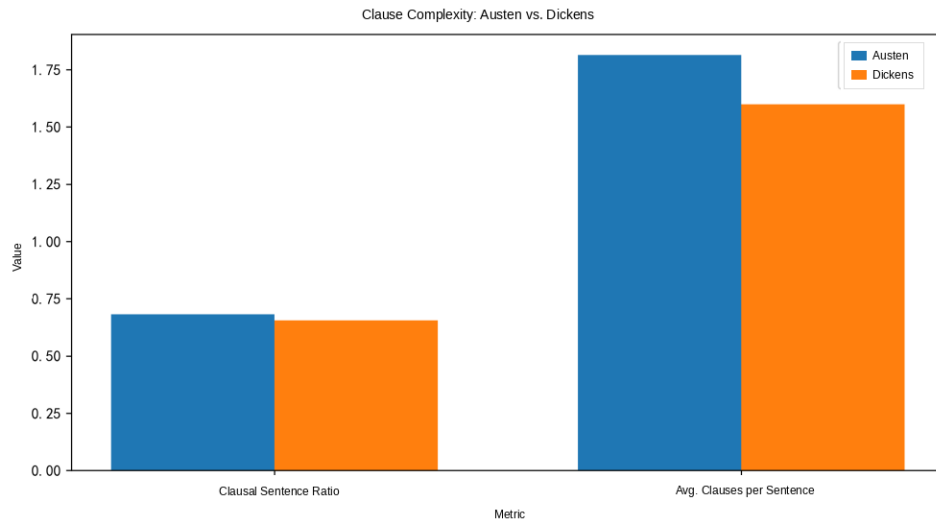


Figure 8 Comparison of clause complexity between Austen and Dickens

Figure 8 compares the clause ratio and the average number of clauses per sentence. Austen scores higher on both indicators, indicating that her writing relies more heavily on embedded clauses and complex sentence structures, which helps present subtle mental activities and social relations.

The two writers also show obvious differences in the use of special sentence patterns. In terms of interrogative sentences, the proportion of interrogative sentences in Austen's corpus is 3.41%, slightly higher than Dickens's 2.76%, which shows that there are more conversational expressions around etiquette, marriage and love, social evaluation and character psychology in Austen's works. In terms of exclamatory sentences, the proportions of both writers are relatively low, with Austen at 5.84% and Dickens at 5.29%, showing that both rely less on explosive expressions of strong emotions and prefer to indirectly present emotions through narrative and character interaction. In terms of the use of passive voice, the proportion of passive sentences in Austen's corpus reaches 20.61%, which is significantly higher than Dickens's 13.63%. To a certain extent, this difference reflects the more polite, indirect and social expression tendency in Austen's texts, and also reflects that her narratives pay more attention to the social relationships and etiquette norms between characters. In contrast, Dickens prefers to use active structures to enhance action advancement and realistic scene descriptions. In terms of direct quotations, the proportion of direct quotations in Dickens's corpus reaches 54.41%, which is significantly higher than Austen's 44.57%. This result shows that Dickens more frequently uses character dialogue to promote plot development, shape social scenes, and enhance dramatic effects. His works have strong colloquial and staged characteristics. Austen, on the other hand, uses more of a combination of narrative and psychological description to develop character portrayals, and her narrative is more biased toward inner psychology and social analysis.

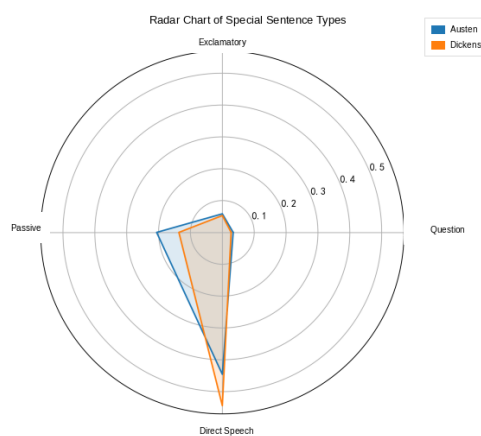


Figure 9 Radar chart of special sentence patterns of Austen and Dickens

Figure 9 provides a radar-chart comparison of special sentence patterns. The figure makes the contrast among interrogative sentences, exclamatory sentences, passive voice and direct speech more visible, showing that the two writers differ not only in sentence length but also in sentence function and

narrative expression.

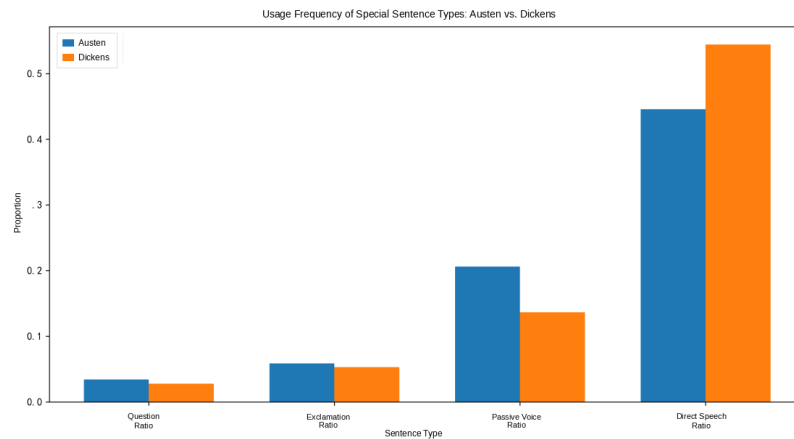


Figure 10 Comparison of the frequency of use of special sentence patterns in Austen and Dickens

Figure 10 further displays the frequency of special sentence patterns in bar-chart form. Austen uses a higher proportion of passive structures, reflecting a more indirect and socially mediated style, whereas Dickens uses more direct speech, which strengthens dialogue, dramatic effect and scene presentation.

Overall, the quantitative results of the syntactic dimension further confirm the classic view in traditional literary criticism, that is, Austen prefers delicate, analytical and rational psychological and social narratives, and his language structure has higher complexity and clause nesting. Dickens, on the other hand, puts more emphasis on scenario-based, dramatic and realistic dynamic narratives, and his syntactic structure is more direct, full of action and dramatic tension.

4.3 Comparison of emotional characteristics

Quantitative analysis of the emotional dimension further reveals the significant differences in the narrative emotional structure between Austen and Dickens. From the overall emotional tone, the average emotional score of Austen's corpus is about 0.97, while Dickens's is about 0.55, indicating that Austen's works as a whole show a more stable, positive and gentle emotional tendency. Austen's novels mainly revolve around marriage and love relationships, family ethics and middle-class social interaction. Their narratives usually end with emotional reconciliation, personality growth and the restoration of social order, so the overall emotional tone is relatively stable and optimistic. In contrast, although Dickens' works also have positive emotional expressions, because they deal with realist themes such as poverty, social oppression, orphan growth, and life at the bottom of society, the mood of the text contains more negative emotions and dramatic conflicts, so the overall emotional score is significantly lower than that of Austen.

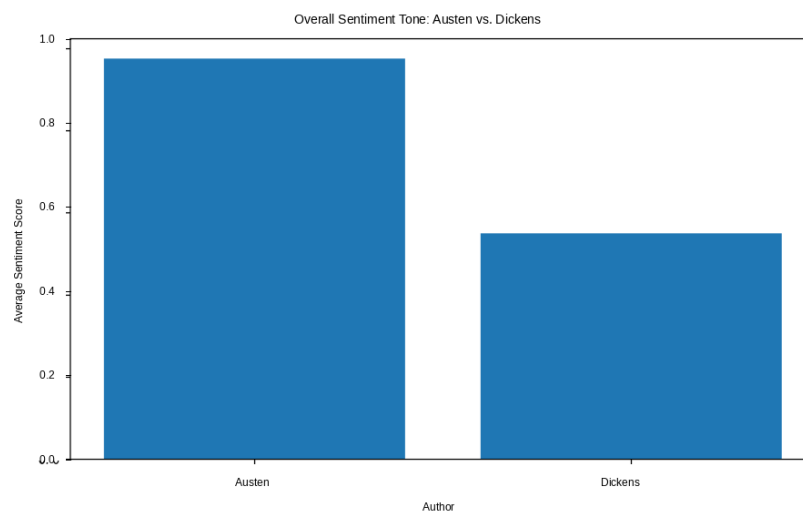


Figure 11 Comparison of the overall emotional tone of Austen and Dickens

Figure 11 shows the overall sentiment scores of the two corpora. Austen's higher average score indicates a more stable and positive emotional tone, while Dickens's lower score reflects the stronger presence of hardship, conflict and social pressure in his realist narratives.

It can be seen from the emotional trajectory diagram that there are obvious differences in the emotional change patterns of the two writers. The emotional curve in the Austen corpus is relatively smooth overall. Most of the text fragments maintain a high positive emotion range, with only a short decline in a few plot conflict parts. This shows that Austen is more inclined to adopt a stable, rational and restrained way of emotional expression, with a more balanced narrative rhythm and relatively mild changes in character emotions. The emotional trajectory of Dickens's corpus shows more violent fluctuations, and a large number of fragments rapidly switch between positive and negative emotions, showing stronger dramatic tension and emotional conflict. This high frequency of emotional ups and downs is closely related to the large number of social crises, character suffering and plot reversals that appear in Dickens's novels, and also reflects the strong dramatic characteristics of his realist narratives.

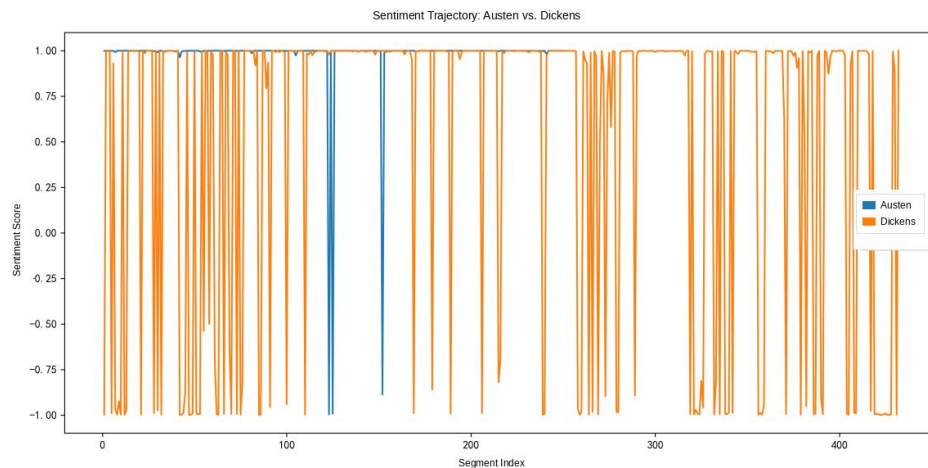


Figure 12 Comparison of the emotional trajectories of Austen and Dickens

Figure 12 presents the emotional trajectory of the two writers' corpora across text segments. Austen's curve is relatively smooth, suggesting restrained emotional development, whereas Dickens's curve fluctuates more sharply, indicating stronger dramatic tension and more frequent emotional reversals.

The five-stage narrative emotional structure analysis further reflects the differences in the narrative organization methods of the two writers. The average emotional score of Austen's works in the five stages has always been maintained at a high level. The emotional value in the beginning, development and ending stages is close to 1.0, indicating that his novels maintain a relatively stable positive emotional tone as a whole. Even in the conflict advancement stage, the emotional decline is relatively limited. This structure reflects the strong sense of order and rational narrative characteristics in Austen's works. Character conflicts can usually be alleviated and resolved through communication, cognitive growth, and social norms. In contrast, the five stages of emotional changes in Dickens's works are more obvious. Its emotional score is relatively high in the development and advancement stages, but drops significantly in the climax and ending stages, especially the ending stage where the emotional value is only about 0.35. This shows that although Dickens's works have a certain tendency of warmth and moral idealism, the core of his narrative is still deeply affected by social reality problems. The fate of characters and social conflicts often burst out in a concentrated manner in the later period, resulting in more violent emotional fluctuations.

Figure 13 divides the narrative process into five stages and compares the average emotional score in each stage. Austen maintains a relatively high and stable emotional level throughout the narrative, while Dickens shows greater stage-by-stage variation, especially in later narrative phases where conflict becomes more intense.

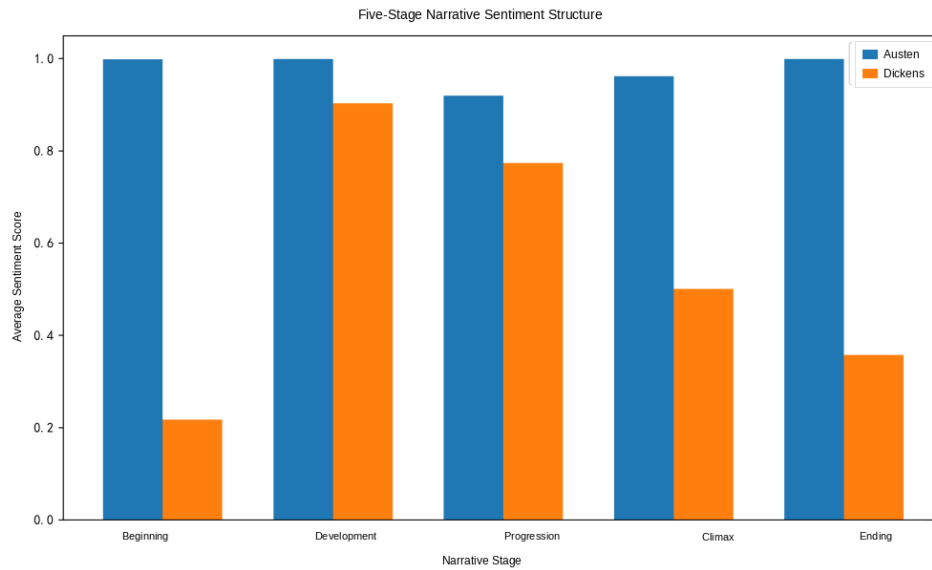


Figure 13 Comparison of five stages of narrative emotional structure

Sentiment volatility analysis further verified this difference. The emotional standard deviation of Dickens' corpus is about 0.81, which is significantly higher than Austen's 0.21. The average change amplitude of adjacent fragments also reaches 0.47, while Austen's is only about 0.05. This shows that the emotional changes in Dickens's works are more frequent and larger, and his texts have stronger dramatic conflicts and emotional tension. In contrast, the emotional structure of Austen's works shows higher stability and continuity, reflecting his rational, restrained and delicate emotional expression.

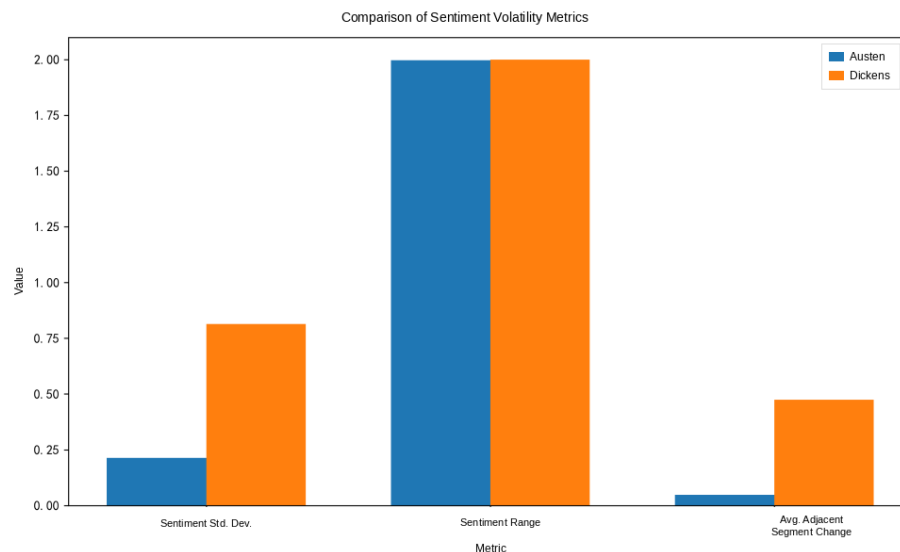


Figure 14 Comparison of sentiment volatility indicators

Figure 14 compares sentiment volatility through the standard deviation of sentiment scores and the average change between adjacent segments. Dickens's higher values indicate more frequent emotional shifts and stronger narrative tension, whereas Austen's lower values show a more continuous and balanced emotional structure.

Overall, the quantitative results of the emotional dimension further confirm the classic view in traditional literary research, that is, Austen is better at constructing stable, rational emotional narratives with character relationships as the core, while Dickens puts more emphasis on emotional conflicts and dramatic tension in social reality, and his works are more dynamic and volatile in emotional expression. This difference not only reflects the different literary focuses of the two writers, but also reflects their unique realistic writing methods.

4.4 Topic model analysis

After completing text cleaning, lemmatization, stop word filtering and character name removal, this article uses the improved LDA topic model to conduct topic mining on the Austen and Dickens novel corpus, divides the novel into multiple text fragments, and uses the fragments as LDA input documents to improve the stability and semantic validity of topic identification. The consistency score results show that when the number of topics is 3 and 6 respectively, the model obtains a higher consistency score (about 0.425 and 0.423), indicating that the topic division within this interval has good semantic interpretation ability, and 6 topics were finally selected for analysis.

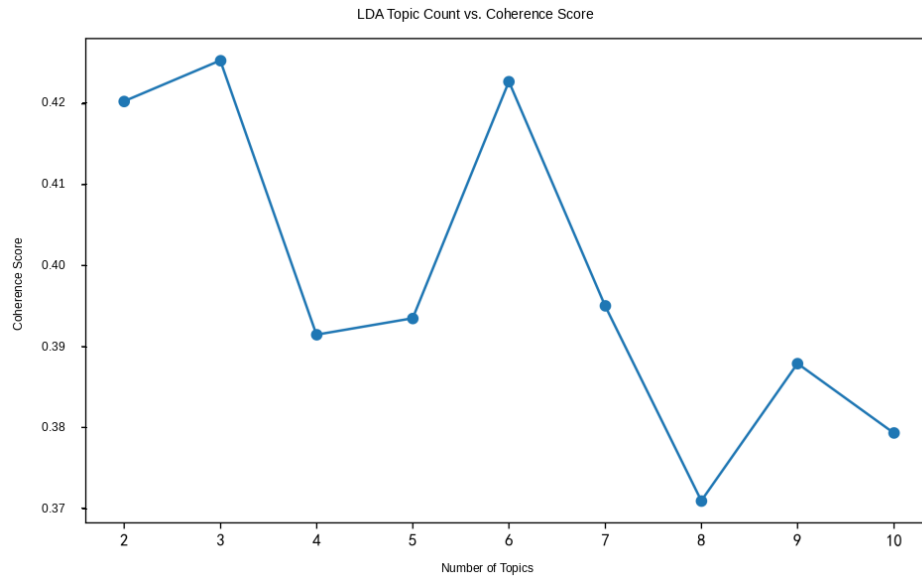


Figure 15 LDA topic number and consistency score changes

Figure 15 shows how the LDA model's coherence score changes with the number of topics. The relatively high coherence scores around three and six topics indicate that the model has good semantic interpretability in this range; therefore, six topics are selected for subsequent topic interpretation and comparison.

Table 1 LDA topic categories and core semantic features

Topic number	Topic name	keywords	core semantic features
Topic 1	The bottom class of society and survival dilemma	poverty, orphan, work, hunger, street	Reflecting poverty, orphans, survival pressure and the living conditions of marginalized groups in society
Topic 2	Marriage Ethics and Family Relations	marriage, family, love, sister, feeling	Reflects marital relationships, family structure, female emotions and social ethics
Topic 3	Growth experience and identity	growth, gentleman, education, youth, identity	Focus on personal growth, self-awareness and social identity transformation
Topic 4	Social Conflict and Historical Space	city, revolution, prison, society, class	Emphasis on urban space, class contradictions, social order and historical conflicts
Topic 5	Individual fate and emotional trauma	memory, suffering, emotion, life, loneliness	Describe the character's fate, emotional repression and psychological change process
Topic 6	Daily life and narrative connection	house, room, conversation, time, daily	Used to connect daily scenes, character interactions and narrative advancement

Table 1 summarizes the six latent topics extracted by the LDA model, including the topic labels, representative keywords and core semantic features. This table provides the interpretive basis for the following comparison of topic distribution between Austen and Dickens, and it connects the statistical

model output with literary thematic analysis.

Judging from the author's overall theme distribution results, Austen's novels have a high theme concentration, with theme 2 occupying an absolute advantage (about 97%) in the text, showing that Austen's works have strong consistency in narrative structure, emotional expression, and description of social relationships; while the theme distribution in Dickens's works is relatively scattered, with themes 4 and 5 accounting for the main proportion, while themes 1 and 3 also have a certain weight, indicating that Dickens's novels show more obvious complexity and diversity in the description of social space, class conflict, and character fate.

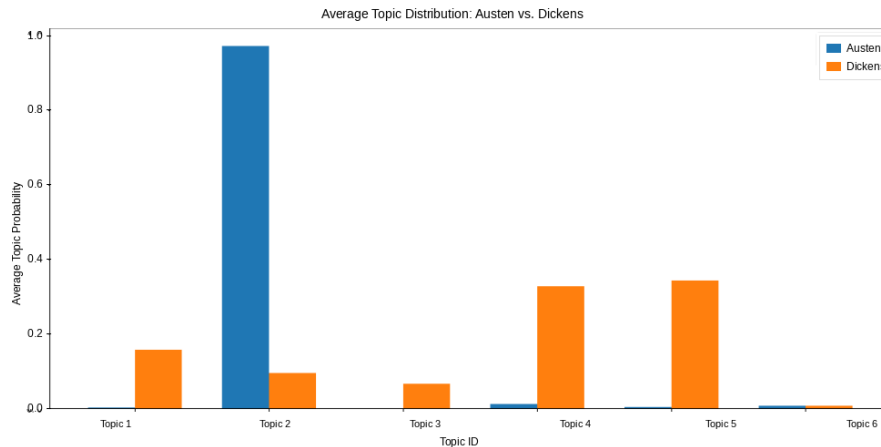


Figure 16 Comparison of average topic distribution between Austen and Dickens

Figure 16 compares the average distribution of the six topics in the Austen and Dickens corpora. Austen's topic distribution is highly concentrated on marriage ethics and family relations, whereas Dickens's distribution is more dispersed across social conflict, individual fate, growth experience and lower-class survival.

Further observation of the theme heat map at the work level shows that there are obvious differences in the theme composition of different novels. Austen's works such as "Pride and Prejudice", "Sense and Sensibility" and "Emma" are highly concentrated on theme 2, reflecting their core concerns on marriage ethics, family relationships and emotional order; while "A Tale of Two Cities" mainly corresponds to theme 4, and "David Copperfield" mainly corresponds to theme 5. "Great Expectations" is influenced by theme 3, theme 4 and theme 5 at the same time, showing a more complex growth narrative and social experience structure. "Oliver Twist" shows an obvious intersection between theme 1 and theme 4, reflecting the connection between poverty, survival and the description of social reality.

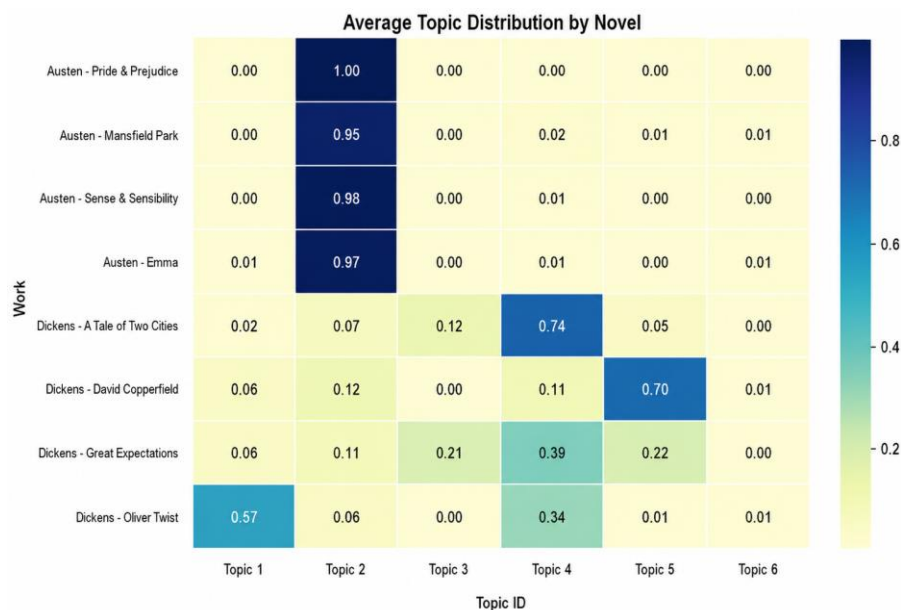


Figure 17 Heat map of the average theme distribution of each novel

Figure 17 further presents topic distribution at the level of individual novels. The heat map shows that Austen's novels are consistently concentrated on the topic of marriage ethics and family relations, while Dickens's novels differ more clearly from one another, reflecting the broader thematic diversity of his realist writing.

Through LDA topic model analysis, it can be found that although Austen and Dickens both belong to the British realist literary tradition in the nineteenth century, their topic structures show obvious differences. The theme distribution of Austen's novels is more concentrated, showing continuous attention to marriage ethics, family relationships and emotional order, while Dickens's novels have a more complex thematic structure, and their texts simultaneously involve multiple dimensions such as social conflict, growth experience, urban space and bottom survival. This difference not only reflects the different literary styles of the two writers, but also reflects the structural difference between the female domestic narrative and the critical narrative of Victorian society. Overall, Austen's novels show stronger theme concentration and stability, while Dickens's novels show higher theme diversity and social narrative complexity.

5. Summary and outlook

Based on natural language processing and quantitative stylistic methods, this study conducted a systematic quantitative comparison of the literary styles of Dickens and Austen from the four dimensions of vocabulary, syntax, emotion and theme. Combining various data results, the writing styles of the two writers can be clearly depicted, and the classic conclusions in traditional literary criticism can be further verified.

First, in the lexical dimension, Dickens and Austen present obviously different ways of organizing language. The total number of words, the number of unique words, and the type ratio in Dickens's works are all higher than those of Austen, indicating that his language system has higher vocabulary richness and expression variation. This difference is closely related to the wide range of urban spaces, social classes, occupational groups and complex narrative scenes in Dickens's works. The large number of concrete words appearing in his high-frequency words and prominent keywords shows that Dickens paid special attention to character movements, body descriptions and the construction of social environments, showing strong realist visual characteristics. In contrast, the high-frequency words and keywords in Austen's works are more focused on words related to emotion, family, and social relationships, reflecting that her works pay more attention to the psychological activities of characters, marriage ethics, and middle-class social order. This result confirms the classic judgments about the two in traditional literary criticism at the data level.

Secondly, in the syntactic dimension, there are also significant differences in the narrative structures of the two writers. Austen's average sentence length and clause complexity are significantly higher than Dickens's, and her text makes extensive use of nested clauses, passive structures, and analytical syntactic organization. This complex syntactic structure makes the narrative more suitable for presenting the subtle emotional relationships, ritual interactions and psychological judgments between characters. It also reflects the rational, restrained and analytical narrative characteristics of Austen's works. In contrast, although Dickens also has high syntactic complexity, its sentence structure is more biased towards action advancement and scene transitions, and the proportion of direct quotations is significantly higher than that of Austen, indicating that Dickens prefers to promote narrative through character dialogue and dramatic scenes. This difference in language structure also further reflects the different realistic expression methods of the two writers.

Thirdly, in terms of the emotional dimension, the quantitative results further revealed the different narrative emotional structures of the two writers. The overall emotional score of Austen's works is relatively high, and the emotional trajectory fluctuates less, indicating that the overall narrative presents a stable, gentle and restrained emotional structure. Even in conflict plots, the range of emotional changes is relatively limited, and character relationships can usually be repaired through cognitive growth and social order. This stability reflects the strong sense of rationalism and order in Austen's works. In contrast, the emotional fluctuations in Dickens's works are significantly stronger, and his texts frequently switch between positive and negative emotions, showing stronger dramatic tension and realistic conflicts. Especially in plots involving poverty, orphans, social oppression and class conflicts, emotional scores tend to drop significantly. This shows that Dickens not only pays attention to social reality issues, but also tends to strengthen readers' moral resonance and social critical awareness through strong emotional ups and downs.

Finally, in the topic model analysis, LDA topic modeling further revealed the structural differences in the narrative focus of the two writers from a macro level. The theme distribution of Austen's works is highly concentrated, and most of the texts focus on the themes of marriage ethics and family relations, which shows that her creations have long been centered on family order, marriage and love relationships, and female growth, and the theme structure has strong stability and consistency. Dickens's works show higher thematic complexity, and his texts simultaneously involve multiple thematic dimensions such as the bottom of society and survival dilemmas, social conflicts, identity, and individual destiny. This shows that Dickens's works not only have a broader social space, but also reflect a more complex realist narrative structure. Compared with traditional literary criticism, the results of the theme model not only verified the creative tendencies of the two writers, but also further revealed the organization and distribution patterns within their thematic structures.

Overall, the quantitative results of this study form a strong mutual corroboration relationship with traditional literary criticism. For a long time, literary studies have generally believed that Austen was more concerned with marriage ethics, female growth and family order, while Dickens placed more emphasis on social reality, class conflict and urban poverty. This study provides more systematic quantitative support for these traditional judgments through multi-dimensional data analysis such as vocabulary, syntax, emotion and theme. At the same time, data analysis also revealed some structural patterns. For example, Austen's works show a high degree of stability at the emotional level, while Dickens has significant emotional fluctuations. Austen's thematic structure is highly concentrated, while Dickens's thematic structure is more diffuse and complex. These findings further enrich traditional literary criticism's understanding of the stylistic differences between the two writers.

However, the method of this study also has certain limitations. First of all, natural language processing methods still mainly rely on vocabulary frequency and statistical structure in nature, so their ability to identify literary texts is still limited. Secondly, the topic naming and literary interpretation of the LDA topic model still rely on the researcher's manual judgment, and there is a certain degree of subjectivity. In addition, topic models are sensitive to text segmentation methods, and different text segment divisions may affect the topic distribution results. Sentiment analysis also has the problem of dictionary dependence. Sentiment dictionaries such as VADER are mainly built for modern English, and there are still certain errors for complex narrative contexts. Future research can further introduce deep learning and semantic representation models, such as Word2Vec, GloVe, BERT and other word embedding models, to conduct a more in-depth analysis of the underlying conceptual structure in the writer's language. Compared with traditional word frequency statistics, word vector models can reveal more complex semantic relationships between words. Finally, the size of the corpus of this study is still limited. It only selected representative novels of two writers, which cannot fully cover all their creative styles and language changes at different stages.

In general, natural language processing and quantitative stylistic methods can effectively supplement traditional literary research and reveal the underlying laws and structural characteristics of literary style through large-scale text data analysis. But at the same time, data analysis cannot replace literary close reading and humanistic interpretation, but should serve as an important auxiliary tool for traditional literary research. Only by putting quantitative analysis back into the context of literary history, cultural background and close reading of texts can digital humanities methods truly exert their explanatory value in the humanities.

Acknowledgements

This research is supported by Innovation Project of Guangxi Graduate Education (JGY2025184).

References

- [1] Zhu Benjun, Nie Hua. *Cross-border and integration: Digital Humanities from a Global Perspective—A Review of the First Peking University "Digital Humanities Forum" Conference* [J]. *Journal of University Libraries*, 2016, 34(5): 16-21.
- [2] Li Yiyun, Niu Yunping. *Exploring the reasons for differences in the English translation of "The Diamond Sutra" from the perspective of metrological stylistics* [J]. *Foreign Language Teaching*, 2024, 45(2): 91-98.
- [3] Zeng Yiping, Zhu Xiaowen. *Application of computational methods in the study of Chinese stylistics*[J]. *Journal of Fujian Normal University (Philosophy and Social Sciences Edition)*, 2006(1):

14-17.

[4] Fan Wenhao, Zhao Yuxiang, Li Jinhao, et al. Method evolution of computational grounded theory and its application in digital humanities research [J]. *Library Forum*, 2026, 46(1): 11-23.

[5] Liu Ying, Xiao Tianjiu. Research on metrological style of "Dream of Red Mansions" [J]. *Journal of Dream of Red Mansions*, 2014(4): 260-281.

[6] Liu Ying, Xiao Tianjiu. Research on the quantitative style of Jin Yong and Gu Long's novels [J]. *Journal of Tsinghua University (Philosophy and Social Sciences Edition)*, 2014, 29(5): 135-147.

[7] Xiao Tianjiu, Liu Ying. Analysis of Jin Yong and Gu Long's novel style based on clustering and classification [J]. *Journal of Chinese Information Science*, 2015, 29(5): 167-177.

[8] Zhang Jiayi, Li Liangbo. Research on the style of "A Moveable Feast" from the perspective of metrological stylistics [J]. *Xinchu Culture*, 2025(27): 27-29.

[9] Wu Jingjing, Xu Juanjuan. Research on the evaluation of the English translation of "The Song of Burial Flowers" [J]. *Foreign Language Studies*, 2021, 7(3): 78-87.

[10] Fu Shouxiang. "A Tale of Two Cities": Conscience and the Spirit of Benevolence of the Times [N]. *Chinese Science Journal*, 2019-07-03(003).

[11] Hui Haifeng. The emotional style of Austen's novels: Research based on the corpus framework of British novels from 1768 to 1830 [J]. *Journal of Guangdong University of Foreign Studies*, 2026(1): 149-160.