Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

The Applicability of Numerical Methods in the Quantitative Study of Syntactic Structure of Chinese Language and Literature

Huayang Shi[1,2,✉]

[1] *Department of General Education, Henan Vocational University of Science and Technology, Zhoukou, Henan, 466000, China*

[2] *Doctor of Education, School of Graduate Studies, Central Philippine University, Iloilo, 5000, Philippines*

ABSTRACT

Under the dual background of the construction of the "new liberal arts" and the digital wave, the interdisciplinary practice of combining humanities and technology continues to develop. Taking a number of Chinese language and literature works as examples, this paper selects language features from the vocabulary and sentence levels, analyzes the syntactic structure of the selected Chinese language and literature works with the help of natural language processing technology and numerical measurement method of language features improved TF-IDF method, and realizes the discussion of the lexical categories of literary works, such as word length, word frequency, word class distribution and word density, as well as the study of sentence categories such as average sentence length, sentence dispersion and sentence class distribution. It is found that most of the utterances of the selected literary works are monosyllabic words and polysyllabic words, the cumulative proportion of both of them is more than 90%, the highest frequency of occurrence is nouns and verbs, both of them are more than 22%, the average sentence length and sentence dispersion do not differ much, and the overall readability of the selected literary works is better, with a free change of syntactic structure and a stronger narrative of the text.

Keywords: natural language processing, TF-IDF, metric analysis, syntactic structure, Chinese language literature

1. Introduction

Chinese Language and Literature is the most culturally distinctive literature in China, which is an embodiment of China's cultural heritage. Chinese Language and Literature focuses on the study of Chinese phonology, vocabulary, grammar, rhetoric, culture and Chinese literary works [1]. Its goal is

✉ Corresponding author.
E-mail address: HuayangShi1989@163.com (H. Shi).
Received 12 January 2024; Revised 05 March 2024; Accepted 25 December 2024; Published Online 15 April 2025.
DOI: 10.61091/jcmcc127a-237

to cultivate students' knowledge and research ability of Chinese cultural traditions, modern Chinese language, literature and culture, and to form specialists with certain writing ability of literary theories and basic tools needed for criticism and research [2].

The most important literary manifestation of Chinese language and literature is the syntax of Chinese language and literature, which is a foundation of Chinese language and literature. Words and word combinations constitute syntactic structure. Syntactic structure can be a phrase or a sentence, a subject-predicate phrase or a separate sentence [3]. From the point of view of internal combination, the basic types of syntactic structures are subject-predicate, verb-object, bias, complementary, and joint. These five types embody the basic grammatical relations of the Chinese language, and we call them the basic syntactic structures, which are called subject-predicate, verb-object, bias, complementary, and joint, respectively [4-5]. And quantitative research is the method of obtaining quantitative research data by using measurement, experiment, survey and other methods, and analyzing the data by using mathematics, statistics and other methods to obtain research conclusions [6]. It is applied to the field of social sciences, emphasizing the establishment of hypotheses in advance, going through a rigorous experimental design, and finally expressing the results of the study with numbers or scales to make judgments on the validity and reliability of the hypotheses [7-9]. The complexity of Chinese language and literature utterances is often a teaching difficulty, and quantitative research on them can reveal their deep-seated laws and help linguists understand and apply them.

This paper cuts through the perspective of digital humanities and analyzes its application in Chinese language literature. In view of its way of measuring style research, a number of representative Chinese language novels are selected to construct a corpus. After pre-processing the corpus, the linguistic feature indicators used for syntactic structure analysis are selected from the vocabulary and sentence levels, and their computational methods are introduced. Then the natural language processing technology is introduced to optimize the TF-IDF algorithm, which is used in the syntactic structure analysis of literary works. Finally, numerical measures are applied to analyze the lexical features of the selected Chinese language literary works in terms of four dimensions: word length, word frequency, word class distribution and lexical density, and to explore their sentence features in terms of three dimensions: average sentence length, sentence dispersion and sentence class distribution, so as to realize quantitative research on the syntactic structure of Chinese language literature.

2. Overview

The Chinese language has evolved and evolved over thousands of years, and its evolution has accelerated the complexity of syntax [10]. The syntax of Chinese sentences needs a theory that pays more attention to the order between words and the application of real and imaginary words in the study of Chinese language and literature, and the pattern syntax theory proposed in the literature [11] can solve this problem. However, it is also a pattern theory with few styles, and the application of Chinese language literature syntax involves ancient Chinese language and modern language, whose patterns are difficult to analyze syntactically accurately. And quantitative research can help to analyze more inter-syntactic patterns and features, literature [12] research shows that the use of restriction in interpersonal interaction in different countries shows opposite patterns. Different language structures in different countries may have created this phenomenon, which shows the complexity of language syntactic structure. Literature [13] quantitatively analyzed the relationship between word polysemy and compound word productivity of monosyllabic verbs and showed that the average number of senses of monosyllabic verbs is positively correlated with compound word productivity. This study suggests that the diversity of compounds of monosyllabic verbs becomes more with the increase of the sensory effect of the applicant, and that compounds of them with other words can be formed into a syntactic structure, leading to syntactic diversification of sentences. And

numerical methods are used to reveal the intrinsic connection or law between data by quantitatively analyzing the data [14]. But there is no research to apply it to the syntactic structure of Chinese language literature, theoretically it can quantitatively analyze the sentence components, syntactic structure and so on, so as to provide more references for the teaching of Chinese language literature.

3. Digital Humanities in Chinese Language and Literature

Over the past few decades, society has undergone a sea change in the creation, recording, analysis and dissemination of information, with digitization overwhelmingly transforming human life. On the one hand, the digital has become a rich resource, the most prominent manifestation of which is the downward spread of digital technologies and the digitization of vast amounts of information. On the other hand, the embarrassing situation that the vast amount of available resources cannot be digested by traditional methods has triggered a change in methodology. Digital humanities has emerged from the irreversible wave of digitization, which is not only a response of scholars to the changes in humanities research, but also a proactive choice to try to expand the boundaries of humanities research.

Digital humanities is firstly "digital", including digital technology, algorithms and data. The second is "humanities", that is, the study of traditional humanities methods, problems and theories. When digital penetrates into various fields of humanities, the study of digital humanities also gradually expands to the field of modern and contemporary literature. The application of digital humanities in the study of Chinese language and literature is shown in Figure 1, a framework that reflects the levels, techniques, processes and purposes of digital literary research.

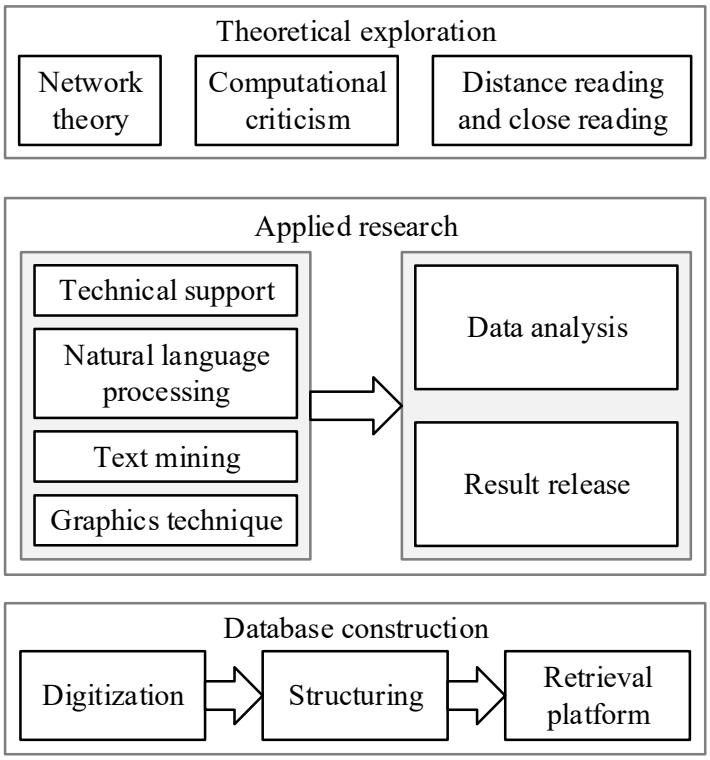

Fig. 1. The application of digital humanities in the study of Chinese language and literature

The construction of database is the basic work of digital humanities research, and the development and construction of the basic database of digital modern and contemporary literature has accumulated certain research experience, and its practice is quite large-scale. At the level of theoretical exploration, theories involving the ontology of digital humanities such as network

analysis theory, computational criticism, distant reading theory, and text analysis theory of multiple interpretations have emerged.

In the framework, the application layer, which is in the middle position, is the bridge between data development and theoretical conception. It takes the database content as the source of research, and makes the theory come to fruition through practice. Currently, the applications of digital technology in literary research include:

1) Measuring style research. Digital Humanities studies the writer's personality projected in the work by quantitatively analyzing the linguistic features. Examining the linguistic style of a text from a quantitative point of view, there are three issues that need to be included in the consideration: first, the choice of linguistic features, they must be quantifiable and appear stably. The second is the purpose of choosing linguistic features, i.e., it is clear why these features are chosen for the study and whether they can adequately reflect the writer's style. The third is the choice of research methodology, i.e., what kind of statistical methods and mathematical models to choose. The most basic and common style research is based on the statistics of language structure features.

On top of the basic feature statistics, more complex computational models and techniques can be used in order to carry out research such as work attribution judgment and sentiment analysis. For example, based on the improved TF-IDF algorithm to get word vectors, after extracting the text content with LSL technology, the cosine similarity is used to calculate the similarity between the texts, which is used to judge the problem of determining the authorship attribution of difficult texts.

2) Construct long-distance literary vision. Considering a large number of texts of long duration as a whole, the technical analysis presents a macroscopic judgment of literary history, and thus unearths the potential patterns that are difficult to be detected during the “close reading” of a single text.

3) Text/social network analysis. Graph technology is a technique for managing, analyzing and predicting data based on graph theory. In the field of humanities, graph technology is usually applied to network analysis, with the purpose of understanding the relationship between variables in a certain system, and then analyzing the narrative discourse and deeper connotations with the help of relational graphs.

In this paper, from the perspective of digital humanities, we adopt numerical methods to quantify the syntactic structure in Chinese language and literature.

4. Corpus construction and linguistic feature selection

Corpus-based linguistic characterization has now become an unavoidable research method in the study of syntactic structure of literary works, as well as a fundamental step in order to obtain the most authentic corpus, whether it is through the collection of corpus based on the automatic writing of web crawler scripts based on the Python language or the direct use of the existing more complete electronic version of the text, all of which provide a greater convenience for the later analysis of the data.

4.1. Corpus building

The PDF electronic texts of 10 representative novels in Chinese language literature, notated as N1~N10, were selected in this study, and their sources were mainly libraries and related websites. The 10 novels of Chinese language literature were arranged according to the time of publication, and their syntactic structures were explored and quantitatively analyzed through vocabulary and sentences.

Since the text type initially downloaded was a PDF version, it needed to be converted to a different format to facilitate the corpus processing process afterwards. Using OCR text recognition software ABBYY, it was first successfully converted into a WORD document, and the first manual proofreading with the paper version of the book, the format conversion of garbled, misspelled words, spaces,

irregular punctuation and wrong paragraph ordering and other issues for correction, and finally exported with the paper book is basically the same as the TXT version of the book, and the corpus cleaning and denoising, and the use of Jieba Chinese The corpus is cleaned and denoised, and the Jieba Chinese thesaurus is used to label the words.

4.2. Language feature selection

Quantifiable statistical linguistic features can be categorized into five levels: character, vocabulary, sentence, paragraph, and part of speech, etc. In this paper, we intend to select the feature elements from the vocabulary and sentence levels as the indexes for the study of syntactic structure in Chinese language literature.

4.2.1. Lexical level. If the text is regarded as a collection of sentences consisting of a series of tokens, then lexical features are a kind of tokens of sentences, which include a series of indexes such as word length, word frequency, word class distribution, and lexical density.

1) Average word length. The average word length, i.e. the ratio between the total number of words and the total number of words, reflects the length of the language structure and lexical complexity of the article, and is used to measure the readability of the language and the ease of reading. The longer the average word length, the more complex the vocabulary.

2) Word Frequency. Frequency refers to the number of times a language unit appears in the text, word frequency refers to the number of times the vocabulary appears, and frequency refers to the probability of the appearance of a language unit in the text, reflecting its distribution in the text, and the formula is:

$$\text{Chance} = \frac{\text{Frequency}}{\text{Total}} \times 100\% \tag{1}$$

3) Distribution of word categories. Classification of words refers to the division of words into different types according to the principles of morphology, meaning and grammatical function, which is mainly based on grammatical function. Word distribution refers to the frequency and frequency of different words appearing in the text, which is an important feature and basis for reflecting the style of literary works.

4) Vocabulary Density. Vocabulary density refers to the ratio of real words appearing in the text to the total number of words, which is used to measure the amount of information in the text, is not affected by the length of the text, and has a certain degree of stability, which is directly proportional to the density of information and the difficulty of reading. The main bearer of text information is real words. The role of real words is to convey information, while the role of dummy words is to connect the real words and build the discourse. Therefore, through the proportion of real words in the text, we can know the information density of the work. The greater the vocabulary density, the greater the information content carried by the text, and the more difficult it is to read the text. The formula is as follows:

$$\text{Lexical density} = \frac{\text{Quantity of substantive word}}{\text{Total number of words}} \tag{2}$$

4.2.2. Sentence level. Sentence is a linguistic unit with a certain intonation, consisting of words, phrases and intonation, with the end of the sentence marked by a period, question mark, ellipsis, or exclamation mark, expressing the complete meaning. Its semantic structure can be divided into objective expression system (propositional structure, tense structure, sentence mode structure) and ideational system (modal structure, intonation structure). In this paper, we will analyze and discuss the average sentence length, sentence length dispersion and sentence class distribution.

1) Average Sentence Length. Sentence length, i.e. the length of a sentence. The larger the value, the

higher the text complexity and the lower the readability. On the contrary, the shorter the sentence length, the simpler and easier to understand the text. Sentence length is considered to be an important factor affecting the readability of the text, while sentence variation affects the rhythm of the text. Works with the right mix of long and short sentences are more attractive to readers and test the author's ability to manage sentence length and sentence style.

The average sentence length, i.e. the ratio of the sum of all sentence lengths (total number of characters in the text) to the number of sentences, reflects the complexity of sentence structure. They correlate with the readability of the text, i.e. the ease of reading. The average sentence length of the entire text is:

$$\text{Average sentence length} = \frac{\text{Total number of words in the text}}{\text{Sentence count}} \tag{3}$$

2) Sentence length dispersion. Sentence length dispersion indicates the ifference between each sentence of the text and the average sentence length, i.e., the degree of deviation or fluctuation of each sentence around the average sentence length, which can most intuitively reflect the degree of dispersion of the text. The greater the dispersion, the greater the difference between the lengths of the sentences, and the greater the change in sentence length, reflecting the strong narrative character of the text. The formula for calculating the discrete degree of sentence length is:

$$D_s = \sqrt{\frac{1}{N}\sum (L_i - \bar{L}_s)^2}\,(i = 1, 2, \cdots, N) \tag{4}$$

where D_s is the sentence length dispersion, L_i is the individual sentence length in the text, and L_s is the average sentence length.

3) Sentence class distribution. Sentence type, belonging to the concept of language, refers to the four types of sentences that can be categorized according to the type of tone function: declarative sentences, interrogative sentences, imperative sentences and exclamatory sentences. Different types of sentences express different tone and emotion, such as questioning sentences marked by the intonation word “?” and the question mark at the end of the sentence, can strongly express the reflection of doubt, affirmation or negation and other emotions. Exclamatory sentences, on the other hand, are marked by intonation words such as “ah” and “what” and exclamation marks, and usually express strong emotions such as excitement, surprise or anger.

4.3. Natural Language Processing

Natural Language Processing (NLP) is a field that studies how to enable computers to understand, process, and generate natural language, with the goal of enabling computers to understand and process natural language in the form of text, speech, and dialogues as humans do, involving a number of subfields such as lexical analysis, syntactic analysis, semantic understanding, machine translation, sentiment analysis, information retrieval, and dialog systems. The study mainly uses lexical analysis (i.e., jieba participle and TF-IDF value calculation) to characterize the syntactic structure of Chinese language and literature.

4.3.1. TF-IDF algorithm. TF-IDF function not only takes into account the situation that the feature words appear more often in a single text, but also takes into account the situation that the feature words appear less often in other texts, if these two situations are available at the same time, then the feature words selected in this way may carry more category information. Therefore, compared with Boolean function, square root function and logarithmic function, TF-IDF function is more comprehensive.TF-IDF function is one of the most used feature weighting methods.

The form of TF-IDF function goes through several stages, including: word frequency, inverse document frequency, product of word frequency and inverse document frequency, normalization of

the product of word frequency and inverse document frequency. The normalization of word frequency and inverse document frequency is the currently used form.

Word frequency is denoted by TF, which is the number of times a feature word appears in the text. A larger TF value for a feature word indicates a stronger representation of its category.

The inverse document frequency is denoted by IDF, which means that if the number of texts containing a feature word is smaller, the stronger the ability of the feature word to represent a certain category of documents. Correspondingly, its weight is also larger.The formula of IDF is:

$$IDF_j = \log(\frac{N}{n_j} + L) \tag{5}$$

where L is a constant, usually determined based on experiments. N is the total number of texts, and n_j is the number of texts in which feature word t_j occurs.

The TF-IDF formula is to multiply the word frequency with the inverse document frequency, and this approach balances the advantages of the two formulas. If a feature word has a high TF in a text and rarely occurs in other texts, then the feature word has good category differentiation ability.The TF-IDF formula is:

$$TF_i * IDF_j = TF_i * \log(\frac{N}{n_j} + L) \tag{6}$$

Considering from the length of the text, when the length of the text is long, even some feature words containing less category information may appear a lot in the text, and their TF values will be large. And when the length of the text is shorter, some words that have a lot of relative occurrences in the text will not have a large TF value. In summary, the original TF-IDF algorithm is standardized using the normalization function cosine.

The standardized TF-IDF function formula is:

$$TF_i * IDF_j = \frac{tf_i * \log(\frac{N}{n_j} + L)}{\sqrt{\sum_{t \in d_k} \left[tf_j * \log(\frac{N}{n_j} + L) \right]^2}} \tag{7}$$

4.3.2. Improved TF-IDF algorithm. The relative simplicity of the structure of the inverse document frequency IDF makes it insensitive in reflecting the degree of contribution of feature words to text categorization, while the TF-IDF algorithm does not take into account the distribution of feature words among categories, and does not take into account sufficiently the information of text categories contained in feature words. Therefore, the consideration of feature words to text category discrimination is added in the improvement process based on TF-IDF.

The TF-IDF weight calculation method and feature selection function are considered to be combined by using TF-IDF to express the frequency of feature items appearing in the text, and the feature selection function to express the relationship between feature items and text categories.

In the case of data imbalance, even if the number of documents containing a feature word in a "large category" is small, it may be larger than the number of documents containing the feature word in a "small category". Consider the introduction of a factor of λ, as shown below:

$$\lambda(t, c_i) = \frac{DF(t, c_i)}{D(c_i)} \tag{8}$$

Where, $DF(t, c_i)$ denotes the number of documents containing feature term t in category c_i, $D(c_i)$ denotes the total number of texts in category c_i, and λ is the ratio of the number of texts

containing feature term t in a category to the total number of documents in that category.

1) $\mathrm{TF}-\mathrm{IDF}*\lambda\mathrm{IG}$. The basic idea of information gain is to consider that although some words do not appear in the document, the meaning represented by the word matches the meaning in the text and the word is very helpful for the category representation of that text. Therefore, this time the word can be extracted and saved as a feature word of that text.

First, the feature words that occur more frequently in a single text but less frequently in other texts of the document set are selected using TF-IDF. Then, the words that do not appear in the sample but can express the meaning of the text and contribute greatly to discerning the category of the text are identified by the information gain method. Finally, λ factors are introduced to combine them, so that the feature words that can represent the sample can be more comprehensively identified, which in turn reduces the time of classification and improves the efficiency of the classifier.

The improved formula is:

$$w(t_i,d_j)=\frac{tf_{ij}*\log(\frac{N}{n_i})*\lambda IG}{\sqrt{\sum_{i\in d_j}\left[f_{ij}*\log(\frac{N}{n_i})*\lambda IG\right]^2}} \tag{9}$$

2) $\mathrm{TF}-\mathrm{IDF}*\lambda\mathrm{CHI}$. The main idea of the CHI algorithm is to find the words that are more relevant to the category and select these words as feature vectors for classification. The larger the value of the feature word chi-square statistic, the more relevant the feature word is to the category.

TF shows the frequency of occurrence of the feature word, and CHI shows the relationship between the feature word and the category by introducing the λ factor. Combining them, the improved algorithm favors feature words that occur more frequently and can contain a large amount of category information.

The improved formula is:

$$w(t_i,d_j)=\frac{tf_{ij}*\log(\frac{N}{n_i}+L)*\lambda CHI}{\sqrt{\sum_{i\in d_j}\left[tf_{ij}*\log(\frac{N}{n_i}+L)*\lambda CHI\right]^2}} \tag{10}$$

5. Quantitative analysis of syntactic structure in Chinese language literature

By quantifying the linguistic features of lexical categories in the corpus, both the lexical usage preferences in Chinese language literature and the distribution of sentence categories in the text can be quickly accessed, and the complexity of the sentence vocabulary can be accurately analyzed. Therefore, this chapter will examine the syntactic structure of the selected novels in Chinese language literature by combining the linguistic feature calculation formula and natural language processing techniques in the above section.

5.1. Lexical category analysis

5.1.1. Word length statistics. Word length is strictly defined in the field of linguistics, which refers to the number of syllables included in a word. In Chinese, a Chinese character is a syllable, so the word length in Chinese refers to the total number of Chinese characters in a word.

1) Average word length. After counting the linguistic feature of average word length in a corpus of 10 Chinese language literary novels, the statistics of average word length are shown in Figure 2. The

average word length of the selected Chinese language literary novel corpus ranges from 1.369 to 1.522, and N3 has the smallest average word length of 1.369, which indicates that the novel uses the most short words. N7 has the largest mean word length of 1.522, indicating that the novel uses the most long words. The difference in average word length between the two is 0.153.

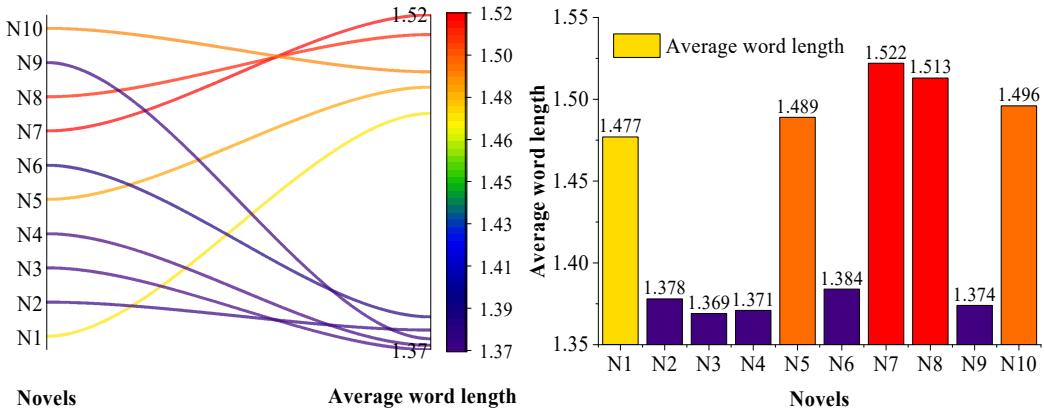

Fig. 2. Statistics on the average word length

2) Word length distribution. On this basis the word length analysis is further refined to count their word length distribution. In modern Chinese, a word is usually a syllable, and X words are X-syllable words, and the average word length can be shown by syllable words. After counting the corpus of Chinese language literature novels, the distribution of word lengths is shown in Figure 3. Monosyllabic words, disyllabic words, trisyllabic words and four-syllabic words occupy a high proportion in the selected Chinese language literary novels, especially the cumulative frequency of monosyllabic words and disyllabic words occupies an absolute proportion, and the cumulative proportion can reach more than 90% in the Chinese language literary novels corpus. Trisyllabic words and four-syllabic words are used in fewer numbers and types, with a cumulative frequency of less than 2.8%, indicating that they are not flexibly distributed in the text. It can be seen that in the syntactic structure of the selected Chinese language literary novels, the use of monosyllabic and disyllabic words is more concentrated, and there are more short words, so the texts are concise and fluent, and easy to read.

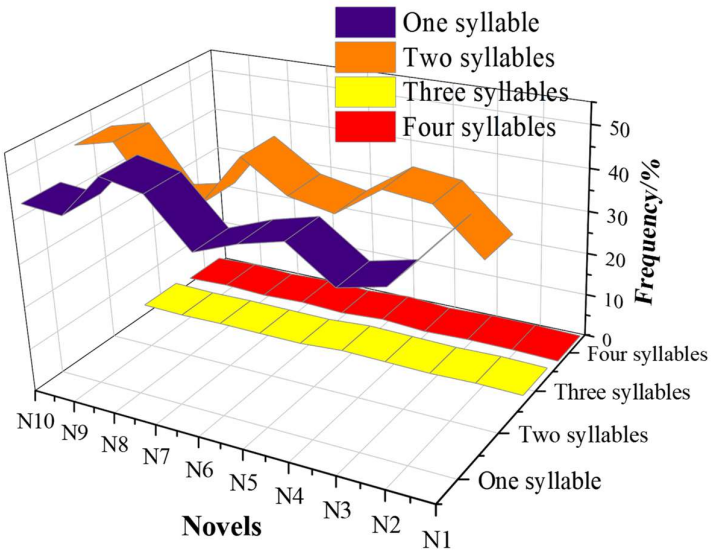

Fig. 3. Distribution of word length

5.1.2. Word frequency analysis. Based on the keyword extraction method of improved TF-IDF, this time, in terms of keyword statistics of the selected Chinese language literature novels, word frequency, weight and lexicality are taken as the foothold, and the deactivated words are removed to obtain the top 20 keywords in the novels. Table 1 shows the statistics of the top 20 keywords, which are mainly nouns, verbs, pronouns, adverbs and a small number of conjunctions and number words, and the proportion of nouns and pronouns in the 20 keywords is the largest, which is 30% and 25%.

Table 1. Statistics of the top 20 keywords

Number	Keywords	Word frequency	Weight	Word class
1	Shunko	1432	1.000	Nouns
2	Self	1358	0.848	Pronoun
3	One	1160	0.831	Numerals
4	Or	1139	0.829	Conjunction
5	That	1127	0.802	Pronoun
6	This	1055	0.683	Pronoun
7	They	1049	0.51	Pronoun
8	Family	1036	0.475	Nouns
9	Some	909	0.444	Pronoun
10	Feel	838	0.431	Verbs
11	Know	808	0.423	Verbs
12	Sammai	786	0.385	Nouns
13	Even	781	0.364	Adverb
14	Stage	717	0.298	Nouns
15	Home	651	0.297	Nouns
16	Not	563	0.255	Conjunction
17	Come back	385	0.251	Verbs
18	Same thing	378	0.203	Pronoun
19	Suddenly	360	0.189	Accessory word
20	And then	306	0.145	Conjunction

5.1.3. Distribution of word classes. The frequency of different word classes in the text effectively reflects the linguistic features of the text, and also proves the word preference of the text authors in their word formation. The structure of high-frequency word class usage in the selected novels in Chinese language literature is shown in Figure 4, where the use of word classes is verbs>nouns>adverbs>pronouns>numerals>prepositions>adjectives> others, which account for 31.56%, 22.80%, 12.79%, 11.97%, 6.29%, 5.54%, 4.74% and 4.29%, respectively. Among them, verbs, nouns, pronouns, number words and adjectives are real words, and adverbs and prepositions are imaginary words. Chinese language literature novels express character behavior and portray character through a large number of real words, and a small number of imaginary words can connect words and sentences to make the overall structure tighter.

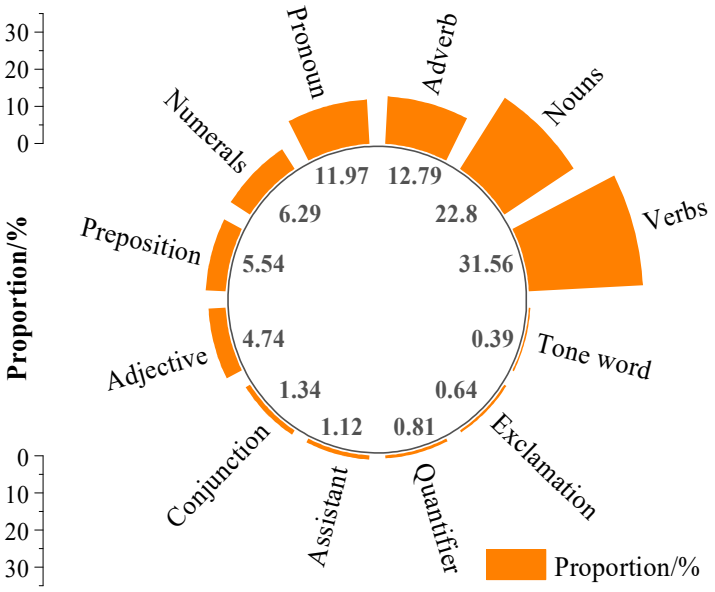

Fig. 4. High frequency part of speech usage in novels

5.1.4. Vocabulary density. Vocabulary density refers to the proportion of real words in the text occupied by all words. In this paper, nouns, verbs, adjectives, number words, quantifiers, pronouns are categorized as real words, and adverbs, prepositions, conjunctions, auxiliaries, exclamations, and intonations belong to the imaginary words category. The lexical density statistics of the selected novels in Chinese language literature by categorizing the results of word division separately are shown in Figure 5.The lexical densities of the 10 novels range from 55.54% to 72.74%, with a larger proportion of real words, higher lexical densities, and a larger amount of information carried, and a high degree of lexical richness of the texts. Among them, the vocabulary density of N7 and N4 reaches more than 70%, indicating that the vocabularies of these two novels are richer and more diverse.

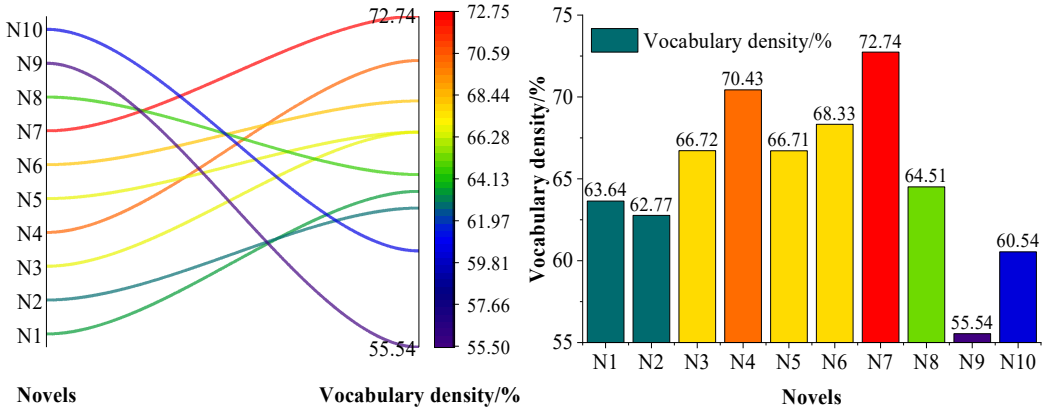

Fig. 5. Vocabulary density statistics for each novel

5.2. Sentence category analysis

5.2.1. Average sentence length. The average sentence lengths of the 10 novels in the corpus were counted, and the results of the average sentence lengths of the novels are shown in Fig. 6.The average sentence lengths of the 10 novels range from 9.65 to 12.78, and the longest average sentence lengths are N5 and N3, with values of 12.78 and 12.67, and the shortest is N4. This reflects the fact that the average sentence lengths of the selected novels are relatively similar, with more short sentences in

the N4 text and a more readable text, whereas the syntactic components and sentence structures of N5 and N3 would be relatively more complex, more difficult to read and less readable.

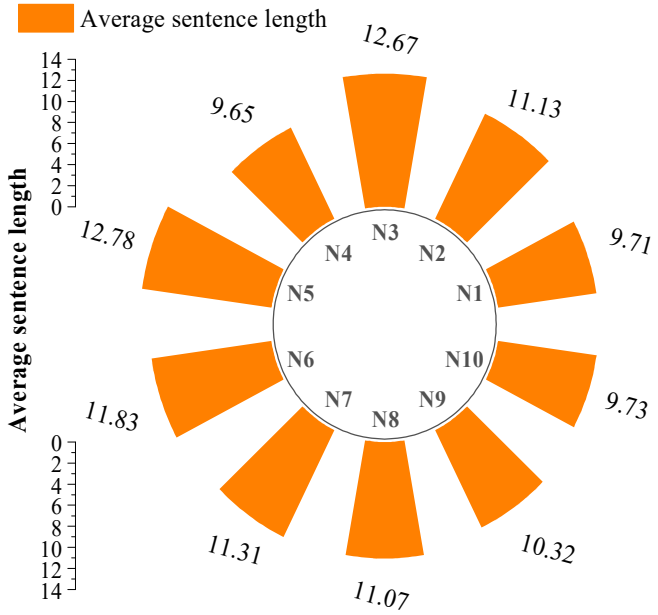

Fig. 6. The average sentence results of novels

5.2.2. Sentence length dispersion. The sentence dispersion in the selected novels of Chinese language literature is calculated, and the results of sentence dispersion are shown in Figure 7. It can be observed through the overall values in the graph that the difference between each value in the graph is not large, and the difference between the maximum value and the minimum value is only 0.46, which may be related to the fact that the genres of the selected works of Chinese language literature are novels.The sentence discrete degree of N9 has the maximum of 1.48, which indicates that it has a large change in sentence length and more free-structured prose sentences, and it reflects that this text has a stronger narrative character.The sentence discrete degree of N6 has the N6 has the smallest sentence discrete degree of 1.02, indicating that its sentence length varies little, the more neatly structured whole sentences, the more syllables are well-proportioned, the text expresses less rhythmic ups and downs, and the narration is smoother and softer.

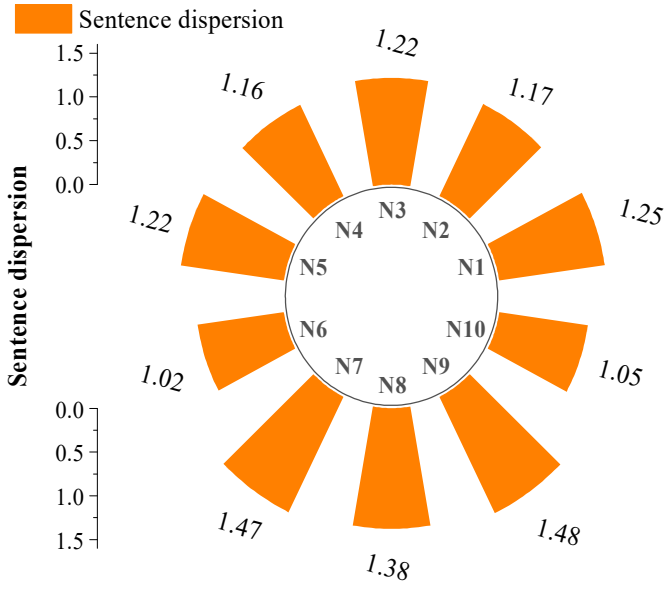

Fig. 7. The result of the discrete degree of the sentence

5.2.3. Distribution of sentence types. Different types of sentences have different tone and intonation, this paper counts the proportion of declarative, interrogative, supplicative and exclamatory sentences in the selected Chinese language literature in the text, and their p-values are 0.024, 0.017, 0.085, and 0.023, respectively, of which the p-value of supplicative sentences is greater than 0.05, which is not significant, so only declarative, interrogative, and exclamatory sentences were Comparison. Attempting to analyze the differences between the tones implied by sentences in Chinese language literature, the proportions of declarative, exclamatory, and interrogative sentences in Chinese language literature in the text are plotted.

A comparison of the distribution of sentence categories in the novels is shown in Figure 8. In the selected Chinese language novels, the most widely used sentences are declarative sentences, which account for 15.21% to 29.98% of the works respectively, the highest percentage, and exclamatory and interrogative sentences both account for less than 1% of them, and the percentage of declarative sentences in the works of novels exceeds the sum of the percentages of exclamatory and interrogative sentences.

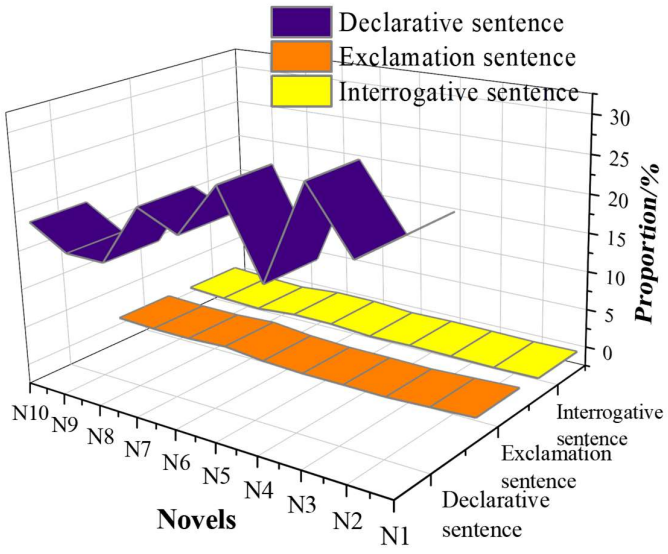

Fig. 8. Contrast of sentence class in novels

6. Conclusion

The study analyzes the application of digital humanities research in Chinese language literature, collects a number of Chinese language novels to establish a corpus, and selects appropriate linguistic features, and explores the syntactic structure of the Chinese language literature corpus by combining the calculation method of each linguistic feature and the improved TF-IDF algorithm. Statistical analysis shows that the average word length of the selected Chinese language novel materials is 1.369~1.522, and their monosyllabic and polysyllabic words are more widely used, and the cumulative proportion of both of them in the selected novels is greater than 90%, which indicates that there are more short words in the utterances of the sample literary works, and the readability is good. Meanwhile verbs and nouns have the highest frequency of occurrence in them, with a percentage of 31.56% and 22.80% respectively. In the category of sentence analysis, the average sentence length and sentence dispersion of the selected works of fiction in Chinese language do not differ much, with the average sentence length ranging from 9.65 to 12.78, and the difference of sentence dispersion is less than 0.46, and declarative sentences have the highest distribution among them, accounting for 15.21% to 29.98%. This paper adopts numerical measurement analysis to realize the quantitative analysis of the syntactic structure of Chinese language literature, mining the lexical linguistic features and sentence linguistic features of the Chinese language literature corpus, and demonstrating the applicability of numerical methods in the study of syntactic structure of Chinese language literature.

References

[1] Wang, D. (2018). Multilingualism and Translanguaging in Chinese Language Classrooms. Springer.

[2] Owen, S. (2020). Readings in Chinese literary thought (Vol. 30). BRILL.

[3] Wu, Y. (2017). The Interfaces of Chinese Syntax with Semantics and Pragmatics. Taylor & Francis.

[4] Pas, V. J., & Paul, W. (2018). The Syntax Of Complex Sentences In Mandarin Chinese: A Comprehensive Overview With Analyses. Linguistic Analysis, 42(1), 63-161.

[5] Niu, R., & Osborne, T. (2019). Chunks are components: A dependency grammar approach to the syntactic structure of Mandarin. Lingua, (224), 60-83.

[6] Cortina, J. M. (2020). On the Whys and Hows of Quantitative Research. Journal of Business Ethics, 167(1).

[7] Mohajan, H. K. (2020). Quantitative Research: A Successful Investigation in Natural and Social Sciences. Journal of Economic Development, Environment and People, 9(4), 50.

[8] Jamieson, M. K., Govaart, G. H., & Pownall, M. (2023). Reflexivity in quantitative research: A rationale and beginner's guide. Social and Personality Psychology Compass, 17(4).

[9] SÜRÜCÜ, L., & MASLAKÇI, A. (2020). VALIDITY AND RELIABILITY IN QUANTITATIVE RESEARCH. Business & Management Studies: An International Journal, 8(3), 2694.

[10] Chen, H., & Wang, Y. (2023). How does language evolve as a multi-level system? A quantitative exploration of written Chinese. Language Sciences, 98, 101554.

[11] Song, J., Yang, C., Sun, Y., Qu, Y., Ma, K., & Cai, H. (2024). Quantitative Research on Chinese Sentences Structure Based on Pattern Grammar. SAGE Open, 14(3), 21582440241268641.

[12] Yu, Q., & Wen, R. (2022). A corpus-based quantitative study on the interpersonal functions of hedges in Chinese and German academic discourse. Heliyon, 8(9).

[13] Chen, C. J. (2023). Verb Polysemy and Compounding Productivity in Chinese: A Quantitative Study. In Chinese Language Resources: Data Collection, Linguistic Analysis, Annotation and Language Processing (pp. 383-395). Cham: Springer International Publishing.

[14] Guus, S., & Fred, V. (2023). Numerical methods in scientific computing. TU Delft Open.