International Journal of Knowledge and Language Processing
Volume 12, Number 1, **2021**

www.ijklp.org
KLP International ©2021 ISSN 2191-2734
pp.1--10

Digital Humanities Research of Jin Yong's Works Based on Quantitative Linguistics

Enshang Xia[1,2], Qinqing Tai[1], Qi Li[1], Jiangtao Li[1], Gaoqi Rao[1✉], Endong Xun[2✉1]

[1]Research Institute of International Chinese Language Education
Beijing Language and Culture University No.15,
Xueyuan Road, Haidian District, Beijing, China
raogaoqi@blcu.edu.cn

[2]Institute for Language Intelligence
Beijing Language and Culture University
No.15, Xueyuan Road, Haidian District, Beijing, China
xinbei122@sina.com

Selected paper from CLSW2019

ABSTRACT*. The authors of this research use various methods of quantitative linguistics to study Jin Yong's works. By taking Gu Long's works and the People's Daily as sources of comparison, the authors explore the linguistic features of Jin Yong's works through textual statistics, and use the statistical data to study the style as well as verify the present research results. In the analysis of lexical categories, Jin's works contain many nouns, verbs, adverbs and pronouns, demonstrating that his Kung Fu novels have abundant scenes and plots. In terms of syntax, Jin's novels incorporate more complicated and longer clauses than that of Gu's novels, which makes Jin's language style more colloquial. Moreover, there is little difference in sentence length dispersion in Jin's works, revealing a unified text rhythm. In terms of human geography, the authors delineate itineraries of the characters and find Jin's works to have the most broad literary geography distribution involving most of China's territory.*

Keywords: Jin Yong's works, quantitative analysis, digital humanities

1. **Introduction.** Academic research on literary works using quantitative methods in the era of big data has gained significant momentum and become more mature in the field of linguistics. These researches provide quantifiable characteristics to study the text style and authors' writing features. Li Xianping [1] once applied pattern recognition and exploratory

[1] ✉ Corresponding Author.

data analysis using function words as statistical features to study the relatedness of the different chapters of *The Story of the Stone*, and drew the conclusion that the work was not authored by Cao Xueqin alone. Liu Ying [2] further did quantitative work on *The Story of the Stone* through calculating and clustering, arriving at the same conclusion. Jin Di [3] compared Ge Fei's and Yu Hua's novels in word length, vocabulary density and sentence dispersion.

Jin Yong, who uses a modern and standard language style, also evokes distinct classical features in Chinese writing, earning a great reputation as one of the finest writers of Kung Fu novels. Xiao Tianjiu [4] analyzed part of Jin Yong's and Gu Long's novels based on text clustering and classification, concluding that there were differences in formality and richness of the language. Quantitative analysis applied to linguistic and humanities research is a profound and scientific supplement to the traditional stylistic research.

This paper offers an analysis of Jin Yong's works from the quantitative linguistic perspective and attempts to present his cultural connotations through the lens of the digital humanities.

2. **Corpora.** In this article, this study establishes a corpus covering 15 Jin Yong's novels (shorted as JY), 16 Gu Long's novels (GL) and the People's Daily (PD) from 1955-1970. GL is selected to show different writing styles. PD is chosen as the comparison to exhibit the characteristics of the Kung Fu novels. The authors process the data with BCC tools [5]: word segmentation, POS tagging, and proofread the data manually. Specific information of the corpora is as follows:

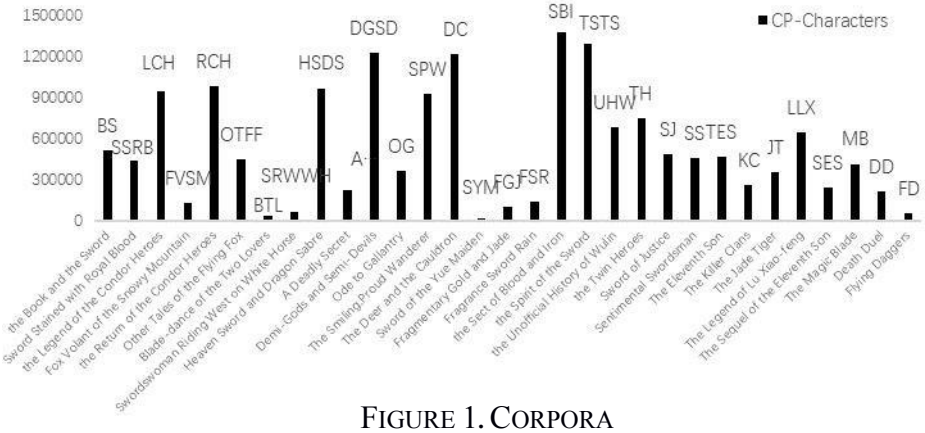

FIGURE 1. CORPORA

The term 'CP-characters' in Figure 1 includes Chinese characters and punctuation, and will be used for the same meaning in this paper. The name of each novel is denoted by their abbreviation in the following.

3. **Parts of speech (POS).** Parts of speech (POS) are the grammatical classifications of words based on their functions. The distribution of POS indicates different styles and authors' writing features. In this section, the authors quantify the word tokens and word types to analyze the characteristics of Jin's novels.

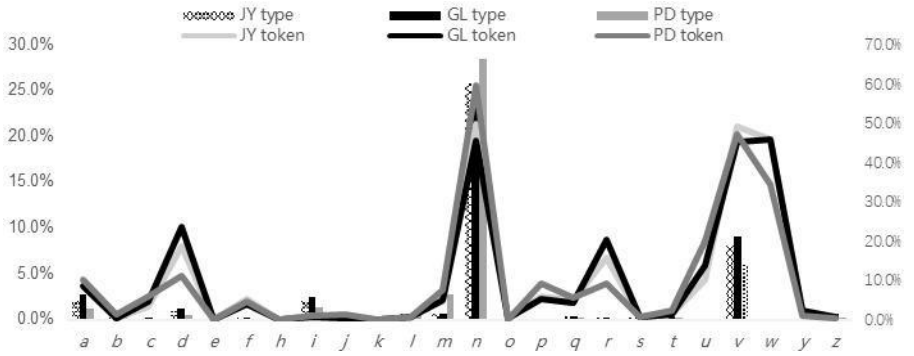

FIGURE 2. JIN YONG, GU LONG, PEOPLE DAILY'S TYPES AND TOKENS

The horizontal axis represents the 24 different POS in the texts (BCC tag), while the vertical axes on both sides stand for the proportion of the POS.

Word tokens, which represent the total number of words of a certain POS, are shown in the polygonal chart (values on the left). The top 6 POS in JY are noun (*n*), verb (*v*), adverb (*d*), pronoun (*r*), auxiliary word (*u*) and adjective (*a*). Tag *v* and *n* hold the first two places of all works. Novels with high proportion of verbs often describe more physical and psychological activities. Generally, adverbs are used to modify the verbs, thus the number of adverbs is relatively consistent with the number of verbs. Adjectives are mainly used to describe characters and surroundings. The large number of adjectives indicates that Jin pays more attention to the details in description.

Compared with JY, numerals (*m*) are frequently used in PD. Words such as 'The 30th conference' and '2980 senators' often appear in newspapers, so that the numerals account for a large part of the text. Moreover, the number of pronouns in PD is significantly less than that in the novels. The narration of newspapers emphasizes clarifying the times, places and characters; therefore, demanding definite reference. However, the novels focus on the development of plots and tend to simplify the complexity when referring to things, which also reflects the principle of least effort in language. As personal pronouns in oral language appear more frequently than in written language, the abundant usage of personal pronouns indicates that the Kung Fu novels, though in written form, embody a colloquial style to a certain extent.

Word types are represented by the bar chart (values on the right). The most commonly used POS in JY are noun (*n*), verb (*v*), idiom (*i*), adjective (*a*), adverb (*d*) and temporary words (*l*). The tags *n* and *v* are plentiful both in tokens and types. Nouns express the names of characters, things, times and places, covering a wide range of areas and suggesting the delicate and extensive portrait. In addition, the infrequent usage of verb types in PD indicates that although the number of verbs is large, the types are concentrated.

The following conclusions can be drawn from the above analysis: Firstly, compared with People's Daily, Jin's novels show the characteristics of Kung Fu novels, such as the large amount of *d* and *r*, richer types in *v*, *a* and *i*, manifest the Kung Fu novels' delicate description and vivid plots. Secondly, there are distinctions between Jin's and Gu's novels, such as Gu's novels various usage of *d* and *a*, while Jin Yong conceives exquisite places

and characters.

4. **Sentence Analysis.** A sentence is the linguistic unit which carries sentence tone, expressing relatively complete meaning. For the linguistic features of sentence, the analysis focuses on the sentence length distribution and the dispersion of sentence length.

4.1. **Sentence Length Distribution.** In written language, punctuation marks such as the full stop (.), the exclamation mark (!), the question mark (?) and the ellipsis (…) are generally used to indicate the end of a sentence. Therefore, these punctuation marks are used as node marks when calculating sentence length. This section analyzes the distribution of sentence length of Jin's 15 novels according to the number of CP-characters and clauses.

4.1.1. **Sentence Length Distribution based on characters.** Figure 3 depicts the number of CP-characters in sentences and their frequencies. The horizontal axis represents the numbers of CP-characters in sentences, while the vertical axis shows the frequency. The following figure takes RCH as an example.

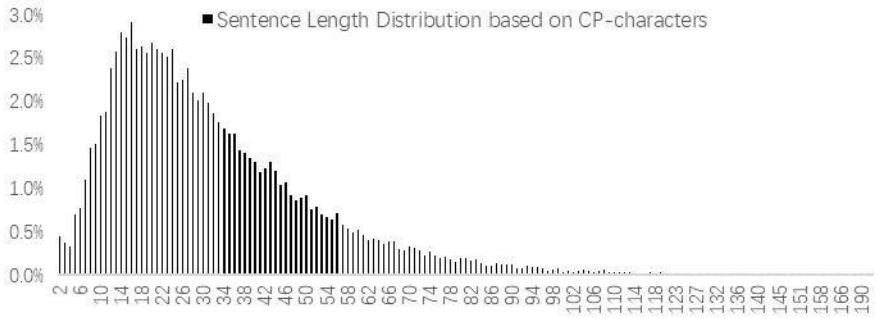

FIGURE 3. SENTENCE LENGTH DISTRIBUTION BASED ON CP-CHARACTERS OF RCH

As is shown in Figure 3, the sentence length of RCH ranges from 2 to 219 CP-characters. While the overall distribution shows a rapid rise and slow decline, there are also some peaks and troughs in this trend. A valley also appears between CP-characters 2 and 4. Statistical results reveal that all the novels except BTL have a local peak when the sentence has 2 characters. Most sentences are in the pattern of "personal pronoun + ellipsis" or "mood word or interjection + exclamation mark", such as "你..." "爹..." "哎!". This pattern reveals the discontinuity of the language in the dialogue, which makes the dialogue scene more real and vivid, and reflects the colloquial style of the novels. Corresponding to the local peaks, the troughs often occur when there are 4 CP-characters, and most are "three-character words/phrases + punctuation", such as "怎么办?" "这一次..." "田伯光!". Disyllabic words account for the majority in modern Chinese. Therefore, the trisyllable words have a lower probability of collocation, which is likely an explanation for the small trough when the sentence length is 4.

The section also considers a holistic comparison between the novels and the People's Daily.

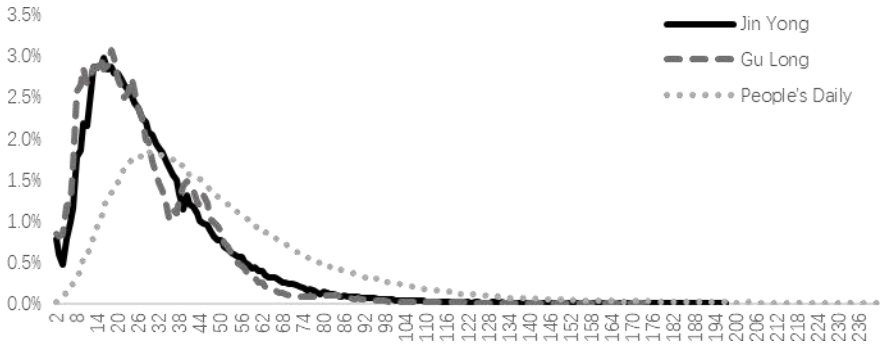

FIGURE 4. THE FREQUENCY OF SENTENCE LENGTH BASED ON CP-CHARACTERS

Figure 4 exhibits that the data fluctuation of the PD is much slower than that of the novels. When the length is longer than 31, the proportion of these sentence in PD is always higher than that of JY. It can be inferred that the PD has more longer sentences and fewer short sentences. Consequently, the language rhythm appears more gentle and steady, relative to the writing style of news. On the contrary, the language of JY is changeable and rhythmic.

4.1.2. **Sentence Length Distribution based on clauses.** Judging from the number of clauses separated by a semicolon (;) in Figure 5a, we observe that the trend of the novels and newspaper is relatively consistent. The horizontal axis represents the numbers of clauses in a sentence, while the vertical axis shows the percentages of the clauses. From the internal comparison of Kung Fu novels, the graph reveals that JY have more sentences containing 3-5 clauses, which suggests that the sentence structure of JY is comparatively complicated. Kung Fu novels have more sentences incorporating 2 or 3 clauses, while the proportion of the sentence containing 3 or more clauses is higher in PD, which indicates that the People's Daily tends to use complex clauses and focuses on the division of paragraphs. It reflects the diversity between different styles.

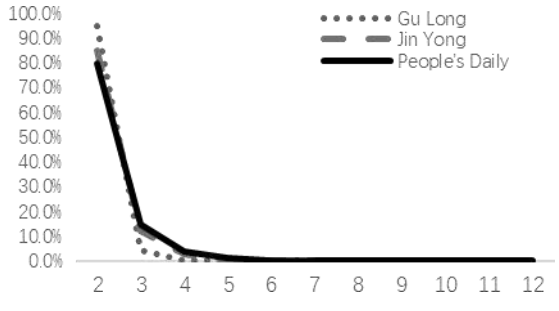

FIGURE 5A. THE CLAUSES SEPARATED BY SEMICOLON

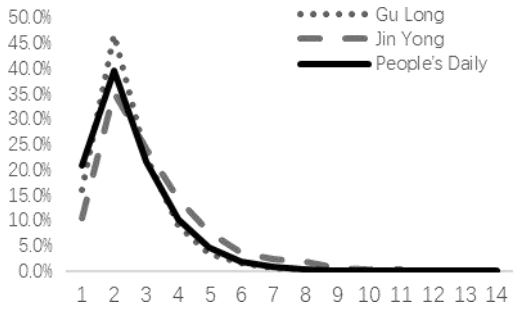

FIGURE 5B. THE CLAUSES SEPARATED BY COMMAS

In addition, judging from the number of clauses separated by a comma (,) in Figure 5b, sentences containing 3 or more clauses demonstrates that the proportion in JY is greater than that in GL. This shows that Jin's novels contain more clauses than Gu's novels.

Combined with sentence length distribution, the trend of Kung Fu novels is relatively consistent, which suggests that Jin's sentences are more fragmented than Gu's, indicating Jin's language style is more colloquial.

4.2. **The Dispersion of Sentence Length.** The dispersion of sentence length (D) refers to the degree to which the sentence length deviates from the average sentence length in the text, reflecting the rhythm of the language. If the dispersion is low, the text is regular, repetitive and rhythmic. In contrast, provided that the dispersion of sentence length is high, the rhythm is variable. The dispersion is calculated by the average sentence length, where the average sentence length (L_0) is figured by dividing the sum of the lengths of all sentences in the text by the total number of sentences (N). The specific formula is as follows:

$$D = \sqrt{\frac{1}{N}\sum(L_i - L_0)^2} \tag{1}$$

The dispersion of sentence length of the novels is computed according to the above formula and shown in Figure 6.

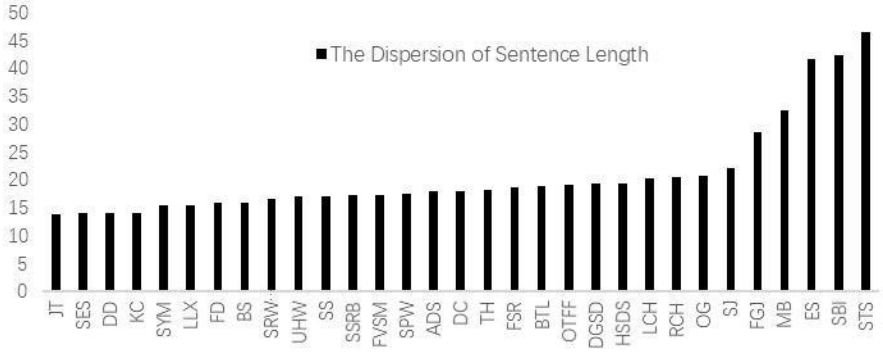

FIGURE 6. THE DISPERSION OF SENTENCE LENGTH

In Figure 6, the horizontal axis represents the titles of all 31 texts, while the vertical axis represents the corresponding values. There is little difference in the dispersion ranging from 16 to 20 among JY, while the dispersion of sentence length in GL is quite different, falling between 13 and 46. Therefore, compared with Gu's novels, Jin's novels are more consistent in text rhythm, maintaining strong and stable rhythm.

5. **Characters' Routes and Geography Distribution.** The time and spatial distributions of literary works are of great significance to the study of the authors' style and background. Jin's novels are divided into periods from the Spring and Autumn period to the Qing Dynasty. At the same time, his characters' activities are set across wide and diverse places and he also portrays many natural scenery and places of historic interest. This section uses text computing and mapping to visualize the routes of the heroes and present Jin Yong's Jianghu. The term "Jianghu" refers to a society in which heroes travel, teach skills and uphold a strict code of honor.

5.1. **Routes.** This section calculates the occurrence that the character name appears with the location name in each paragraph and sorts the data according to frequency. Such co-occurrence of names and places is regarded as the itineraries of characters in the texts.

TABLE 1. HIGHEST CO-OCCURRENCE

Character	Place-frequency	Kind
Wei Xiaobao	Raksha(Russia):432 Beijing:302 Yunnan:209 Yangzhou:197 Taiwan:191 Mt. Wutai:97 Moscow:86 Is. Shenlong:83	254
Ling Huchong	Mt. Heng:244 Mt. Hua:236 Shaolin Temple:140 Mt.Song:85 Mt. Hen:57 Mt. Wudang:49 Meizhuang:45	136
Zhang Wuji	Persia:155 Mt. Wudang:106 Shaolin Temple:89 Guangming Ding:89 Menggu:79 Mt. Wudang:57 Is. Binghuo:45	115

Among the 15 novels, Wei Xiaobao in DC travels to the most extensive areas, with a total of 254 different places. Wei Xiaobao is an intricate character in Jin's novels. Not only does he work in brothels and a palace, but also he participates in a secretive organization and becomes the front-line commander to subside war. Therefore, it is reasonable that he travels to so many places. Linghu Chong in SPW, ranks the second, and visits 136 places such as Mt. Heng, Mt. Hua, Shaolin Temple and so on. Although Ling yearns for freedom, but as the leader of the Hengshan group, he has close contact with different Wulin Schools.

The authors also mark the data of co-occurrence on Baidu maps. Due to the diverse historical background in each novel and the indistinct boundary of ancient cities, we eventually choose to use modern maps and make adjustments when encountering ambiguous membership of the territory. For example, "Menggu" in JY should be Inner plus Outer Mongolia now, but uniformly marked as the present Outer Mongolia instead. In addition, most descriptions about geography are Jin's artistic imagination, such as "Mt. Tie Zhang" and "Is. Shen Long" are places in the literary geography, which do not exist in reality. Even "Mt. Hua" which appears in several novels is the "selective description" of the real Mt. Hua, the authors mark it down on the map to present the protagonist's traveling routes. Taking LCH as an example, the places where Guo Jing travels (dark) and Guo Jing & Huang Rong couple visit together (light) are marked. The farthest place they travel to is Huarazimo (an old city destroyed by Genghis Khan in the novel) in Central Asia where Guo follows Genghis Khan to attack Samarkand, while Huang gives advice to him. Their activities are mainly concentrated in East and Central China, and they go out in pairs and travel to many places along the way.

FIGURE 7A. LCH MAP

FIGURE 7B. HSDS MAP

In HSDS, Zhang Wuji (dark) and Zhao Min (light) are marked. Compared with other novels, it is apparent that the scope of characters' activities is more scattered, even to the edge of the Chinese territory.

5.2. **Geography Distribution.** This paper extracts the high-frequency places from 15 novels, which are considered to be depicted by Jin Yong and are closely related to the story. As can be seen in Table 2, places involved in his works are almost all over China and also include foreign countries such as Russia, Uzbekistan.

TABLE 2 GEOGRAPHY DISTRIBUTION

District	Province
East China	Shandong, Jiangsu, Anhui, Zhejiang, Fujian
South China	Guangxi, Guangdong, Taiwan
Central China	Hubei, Hunan, Henan, Jiangxi
North China	Hebei, Shanxi
Northwest	Ningxia, Xinjiang, Qinghai, Gansu, Shanxi
Southwest	Yunnan, Sichuan, Guizhou, Chongqing, Xizang
Northeast	Heilongjiang, Jilin, Liaoning
Other Countries	Russia, Uzbekistan

With the help of mapping tools, a three-dimensional map of DC is generated to observe the places distribution, in which the more the place appears, the darker the region is. This novel is about fighting for the throne, telling much about wars in different places, so the regional distribution is scattered. Among them, Raksha (now named Russia) appears most—Wei Xiaobao helps the Russian princess seize power. After that, he is sent to Raksha again for helping with a foreign invasion. Therefore, Raksha accounts for a large part in the book. In addition, Beijing and Taiwan are all closely related to the storyline.

FIGURE 8A. DC MAP

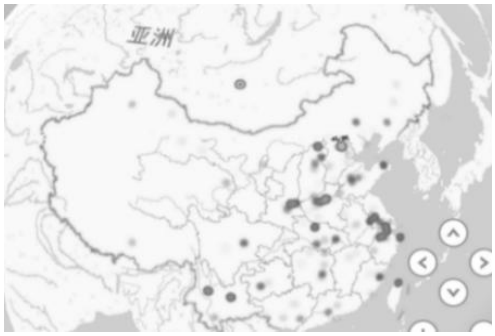

FIGURE 8B. JY GEOGRAPHY DISTRIBUTION

The places in the 15 novels are collected on one heat map. The darker the color, the more frequent the places appear in JY.

We can see that Mongolian grassland is the darkest part, since Han-Mongolian ethnic relations are frequently mentioned in JY. In the first part of LCH, the Mongolians extend their warm, kind and honest nature. While later, for the sake of ethnic interests, the Mongolians show their merciless and aggressive side. Genghis Khan even orders the destruction of the cities. The multiplicity of Mongolian images is presented correspondingly with different historical periods and different relationships among people. It is in this meticulous and holistic depiction of Mongolian soldiers and heroes in different novels that we can see Jin Yong's profound and progressive national consciousness as a writer and historian. In addition, Beijing, Taiwan, Mt. Hua and other places also have high density, in line with our impression of JY. His broad geographical and diachronic vision are well shown through the large-scale space-time migration in his novels. Particularly, the wide range of geographical locations implies that Jin Yong's Jianghu has a strong inclusion of the civilization of East Asia and the surrounding regions.

6. **Conclusions.** Jin Yong is a writer who has a profound influence in Chinese life and social culture, his works can be analyzed from different perspectives. With the help of linguistic features in the texts, style can be measured by a more objective way. The authors can not only use the results to verify the literary analysis of Jin's works, but also discover some new characteristics. According to the analysis of POS, Jin's novels show the common characteristics of Kung Fu novels with copious descriptions and vivid plots. From the internal comparison of novels, the various use of nouns indicates Jin's unique conception of places and characters, while Gu Long focuses more on the description of actions and scenes. According to the analysis of sentences, on the one hand, there are more longer sentences in Jin's novels than Gu's novels, which implies that Jin's sentence structure is complicated and includes more insertions in sentences, thus the language of his works is more colloquial. On the other hand, the sentence length dispersion of Jin's works is lower than Gu's, indicating that Jin's text rhythm is steady and regular. Finally, according to the analysis of the geography distribution, we visualize literary geography span from domestic to foreign countries, presenting the historical and humanistic connotations of the works.

Acknowledgments. This paper is supported by the Research Funds of National Language Committee (YBI135-90), MOE Key Research Center Project (16JJD740004) and the Funds of Beijing Advanced Innovation Center for Language Resource (TYR17001).

REFERENCES

[1] X. P. Li, New Theory of Dream of Red Mansions. *Journal of Fudan University*, vol.29, no.5, pp.3-16, 1987.

[2] Y. Liu and T. J. Xiao, A Stylistic Analysis of A Dream of Red Mansions, *A Dream of Red Mansions Journal*, vol.34, no. 4, pp.260-281, 2014.

[3] D. Jin, A Corpus-based Quantitative Study of Ge Fei's and Yu Hua's Novels, Master thesis, Nanjing Normal University, 2018.

[4] T. J. Xiao and Y. Liu, A Stylistic Analysis of Jin Yong's and Gu Long's Fictions Based on Text Clustering and Classification, *Journal of Chinese Information Processing*, vol.29, no.5, pp.167-177, 2015.

[5] E. D. Xun, G. Q. Rao, X. Y. Xiao and J. J. Zang, The construction of the BCC Corpus in the age of Big Data, *Corpus Linguistics*, vol.3, no.1, pp.93-109, 2016.