

A Corpus-based Study on Language Style and Authorship Identification: Statistical Characteristics of Mo Yan's and Jia Pingwa's Works

Xiaoying Wang^a, Xiaonan Zhu^b

School of International Studies, Zhejiang University, Hangzhou 310058, China

^awhat_gonna_do@126com, ^binjoyce@163.com

Keywords: MO Yan, JIA Pingwa, structural characteristics, style.

Abstract. Since the 1970s, the corpus-based quantitative language research method has been introduced to Chinese stylistic studies. The paper proposes the method that applies statistical analysis of corpus data in language style comparison and authorship identification. The paper discovers 7 language structural characteristics which possess obvious distributional differences through the statistical analysis of 12 language structure characteristics in two sample corpora of 2 million words. This paper, employing quantitative and statistical approaches in authentic materials, brings greater objectivity in stylistic comparison and authorship identification.

1. Introduction

Buffon said, "Le style, c'est l'homme". This opinion, which is still popularly meaningful in theory, denotes that the writing style makes the man. When the theory relates specifically to literature works, the style can be interpreted as wording and phrasing, in other words, choice of language. It is the different frequency of basic language unit that determines various writing styles. Concrete analysis, consequently, can be made to these choices quantitatively through statistical methods. With real data, the precision of language characteristics research is guaranteed though style has long been deemed as abstract and subjective. This paper is a contrastive study of Mo Yan's and Jia Pingwa's works by employing quantitative and statistic approaches to two 2-million-word corpora of these two authors respectively. This research should give insights to stylistic studies and authorship identification.

2. Statistical methods and style studies

Adopting mathematical and statistical approaches in language research is one of the most important achievements of modern linguistics. Statistical probability, performed by computer, further expands the linguistic research scope and gives researchers a broader horizon and a multi-dimensional point of view. In-depth language and parole researches can therefore be conducted. The research in old-fashioned way which relies more on personal impression and judgment has been questioned for its ambiguity and subjectivity. The quantitative analysis as a remedy for this limitation, however, gives greater credibility. This method observes the quantitative relationship among language structures. Based on this principle, the rule and features of language stand a good chance to be discovered when substantial language samples are available. Corpus, in this regard, also occupies a decisive position. Humanities has evolved into a disciplinary era that quantitative methods of natural science are introduced [1].

In the area of Chinese language study, statistical methods are extensively used in metrical studies on Chinese words and phrases, and statistical style analysis. Qian Feng and Chen Guanglei, were the first two scholars in the field of rhetoric who promoted to establish Computational Stylistics. They put forward the proposition that methods of mathematics and computer techniques could be introduced to solve stylistic problems. An empirical research was made to compare Ba Jin's *We Met with Commander Peng Dehuai* and Ni Haishu's *Motor Tricycle* in the aspect of word, sentence, rhetoric, prosody and the art of composition [2]. Yan pointed out that language features, which in conventional sense could not be measured quantitatively, can be analyzed with computational and statistical

approaches[3]. Chen et al. and Huang & Liu, and other researches have conducted computer-aided statistical analysis to different texts[4-5]. The paper, inspired by these literatures, employs statistical and computational methods to Mo Yan's and Jia Pingwa's works with the expectation that inner relationship between style and language structure could be investigated further.

3. The present studies

Sample corpora of these two authors, Jia and Mo, are established and segmented. Altogether 12 specific language structures are extracted as stylistic features, and their ratio and percentage in the texts are calculated. Similarities and differences are explored based on the mean value.

3.1. Corpus

When collecting corpus, external factors should be considered. Similar language environment guarantees the comparability of the two authors, and Mo and Jia's works which share considerable commonalities, therefore, have a practical research value. Here are three main reasons for selecting Jia's and Mo's works as corpus. First, these two authors are age-matched. Mo was born in 1955 and Jia was born in 1952. Second, both of their works are with a strong spirit of clues on the local complex throughout the creation. From the 1980s, Mo's works, filled with nostalgic emotions and regarded as root-seeking literature, have been well received by Chinese readers. Jia's work is realistic. Jia met with huge favor by writing rural China's social confusion and conflicts with no rhetoric of words but great sentiments. Third, they are both famous contemporary writers with matching accomplishments. Mo and Jia won Mao Dun Prize in 2011 and 2008 respectively. Similar in age, writing environment, publication time of representative works and length of work, it is deemed that works of these two writers are comparable and worth studying.

3.2. Method

Two sample corpora are built for comparison study. They are of similar scale of around 200 million words: Mo's corpus includes 17 works (1,976,260 words) and Jia's corpus includes 19 works (1,925,943 words). After text collection, ICTCLAS 2008 and AntConc are used for segmentation, tagging and performing language characteristics analysis statistically.

3.3 Data analysis

AntConc is applied to extract and perform statistical analysis in the corpus. Data distribution of word length, sentence length, type token ratio, adverb ratio, noun ratio, pronoun ratio, auxiliary word ratio, punctuation ratio, declarative sentence ratio, interrogative sentence ratio, exclamatory sentence ratio and hapax ratio is observed (See Table 1).

Word length = number of Chinese characters (punctuations excluded)/number of words

Sentence length = number of Chinese characters (punctuations excluded)/number of sentence

Type token ratio = number of types/number of tokens

Adverb ratio = number of adverbs/number of words

Noun ratio = number of nouns/number of words

Pronoun ratio = number of pronouns/number of words

Auxiliary word ratio = number of auxiliary words/number of words

Punctuation ratio = number of punctuations/number of Chinese characters

Declarative sentence ratio = number of declarative sentences/number of all sentences

Interrogative sentence ratio = number of interrogative sentences/number of all sentences

Exclamatory sentence ratio = number of exclamatory sentences/number of all sentences

Hapax = words appearing once only

Table 1 Language Structure Distribution of Mo & Jia Corpora

Language structure	Mo's works	Jia's works
Word length	1.5645	1.5972
Sentence length	28.9899	30.8631
Type token ratio	0.0281	0.0249
Adverb ratio	0.0807	0.1020
Noun ratio	0.2300	0.2334
Pronoun ratio	0.0890	0.0927
Auxiliary word ratio	0.1026	0.1020
Punctuation ratio	0.1090	0.1214
Declarative sentence ratio	0.7675	0.5928
Interrogative sentence ratio	0.1245	0.1988
Exclamatory sentence ratio	0.1081	0.2084
Hapax ratio	0.0083	0.0075

Word length and sentence length. The word length of Jia's works is 2.09% longer than that of Mo. At sentence level, on average, Jia's sentence is 1.8732 longer than Mo's. There is only a slight difference of 6.46%.

Type token ratio. Type token ratio (TTR) is an important measurement indicator of lexical richness, and it is an index widely borrowed in judging writing proficiency. As can be observed from the table, Mo's TTR is 0.0281 higher than Jia's. A gap of 11.39% implies that Mo's works are more rich in vocabulary.

Word. In terms of adverb, a frequency of 0.1020 is observed in Jia's works, 26.39% higher than that of Mo. Jia is more inclined to use adverbs. The frequency of noun of Jia is 0.0034, namely 1.48% higher than Mo. In the matter of pronoun, the frequency is 0.0037 higher than Mo's. In the matter of auxiliary word, a frequency gap of 0.0006 is discovered that Mo uses 0.59% more auxiliary words than Jia. In general, difference in vocabulary use between two authors is not significant.

Punctuation. The distinction in punctuation is more salient. A frequency gap of 0.0124 is found that Jia uses 11.38% more punctuation than Mo. This finding is in line with what is discovered with sentence length.

Hapax. In Mo's works, more hapax can be seen. The hapax frequency of Mo's work is 0.0008, 10.67% more than Jia's. It complies what is found in TTR. Hapax, like TTR, plays a major role in determining vocabulary richness. More hapax indicate a higher level of lexical richness. With more hapax and TTR explored in Mo's works, it can be inferred that Mo's words are greatly more diverse than Jia's.

Sentence. The works of Jia and Mo show more differences in terms of sentence. Mo uses 29.47% more declarative sentences than Jia, with a frequency gap of 0.1747; Jia tends to use 59.68% more interrogative sentences than Mo, with a larger frequency gap of 0.0734; when it comes to exclamatory sentences, a more significant frequency difference of 0.1003 is observed that there is 92.

The above statistical data demonstrate that among all the 12 language structure characteristics, salient difference is observed in terms of TTR, adverb, punctuation, hapax, declarative sentence, interrogative sentence and exclamatory sentence. In contrast, Mo's works enjoy a higher lexical richness and shorter sentences. Mo is inclined to use more declarative sentences. Jia's works are featured with more adverbs, interrogative sentences and exclamatory sentences.

In order to guarantee the reliability of the experiment, two test corpora of smaller scale are established. Other works of the two authors are collected as text source. TTR, adverb ratio, declarative sentence ratio, interrogative sentence ratio, exclamatory sentence ratio and hapax ratio are examined for a second time (See Table 2).

Table 2 A Statistical Comparison of Small-Scale Test Texts and Corpus

Language structure	Mo's small-scale test texts	Jia's small-scale test texts	Mo's copora	Jia's copora
Type token ratio	15.7982	21.3772	35.5830	40.1634
Adverb ratio	0.0561	0.0925	0.0807	0.1020
Punctuation ratio	0.1114	0.1112	0.1090	0.1214
Declarative sentence ratio	0.8436	0.5962	0.7675	0.5928
Interrogative sentence ratio	0.0750	0.2098	0.1245	0.1988
Exclamatory sentence ratio	0.0815	0.1940	0.1081	0.2084
Hapax ratio	0.0232	0.0173	0.0083	0.0075

The data obtained from the test corpus further verify what have been discovered in the sample corpora. Considering all the findings, it can be confirmed that 7 language structure characteristics, i.e. TTR, adverb, punctuation, hapax, declarative sentence, interrogative sentence and exclamatory sentence are distinguishing features of Mo's and Jia's works. The representativeness of Mo's works lies higher lexical diversity, brief sentence and highly frequent declarative sentences. Jia, however, prefers to use longer sentences. More interrogative and exclamatory sentences are utilized in psychological description and emotional expressing.

4. Summary

Altogether 12 language structure features of Mo Yan's and Jia Pingwa's works are reached after statistical researches are performed in the two 200 million words sample corpora. The contrastive analysis demonstrates 7 saliently different features. The smaller scale test corpus further backs up the findings of sample corpus. It thus can be seen that TTR, adverb, and the ratio of adverb, punctuation, declarative sentence, interrogative sentence, and exclamatory sentence are the distinctive features of Mo's and Jia's writing style. The traditional way of judging writing style is denounced for its lack of objectivity. Corpus and statistical method can be applied as remedy for this deficiency, and it is one crucial approach in the modern Chinese stylistic studies[5]. Quantitative research method based on corpus, one innovative research strategy, will provide higher accuracy and reliability.

References

- [1]. Cao, C. On language stylistic statistics. *Journal of Tianjin University*. Vol. 4(1988), p. 70-75.
- [2]. Qian, F. & Chen, G. Suggestion on the development of Chinese computational stylistics studies. *Rhetoric studies and Chinese rhetoric studies*. Shanghai: Fudan University Press.1983.
- [3]. Yan, S. On quantitative research methods in linguistics and applied linguistics. *Journal of PLA University of Foreign Languages*. Vol. 24 (2001) No. 5, p. 4-5.
- [4]. Chen, X., Li, W., & Wang, Y. Application of quantitative characteristics in comparison of language style and author judgment—Triple Gates of Han Han and Never Flowers in Never Dreams of Guo Jingming as examples. *Computer Engineering and Applications*, Vol 48(2012) No. 3, p. 137-139.
- [5]. Huang, W., & Liu, H. Application of quantitative characteristics of Chinese genres in text clustering. *Computer Engineering and Applications*, Vol. 45(2009) No, 29, p. 25-27.