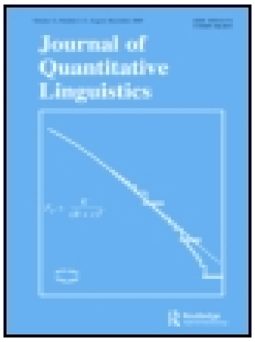

Journal of Quantitative Linguistics

ISSN: 0929-6174 (Print) 1744-5035 (Online) Journal homepage: https://www.tandfonline.com/loi/njql20

Prose, Verse and Authorship in *Dream of the Red Chamber*: A Stylometric Analysis

Haoran Zhu, Lei Lei & Hugh Craig

To cite this article: Haoran Zhu, Lei Lei & Hugh Craig (2020): Prose, Verse and Authorship in *Dream of the Red Chamber*: A Stylometric Analysis, Journal of Quantitative Linguistics, DOI: 10.1080/09296174.2020.1724677

To link to this article: https://doi.org/10.1080/09296174.2020.1724677

Published online: 09 Feb 2020.

JOURNAL OF QUANTITATIVE LINGUISTICS
https://doi.org/10.1080/09296174.2020.1724677

Prose, Verse and Authorship in *Dream of the Red Chamber*: A Stylometric Analysis

Haoran Zhu[a], Lei Lei[a] and Hugh Craig[b]

[a]School of Foreign Languages, Huazhong University of Science and Technology, Wuhan, People's Republic of China; [b]School of Humanities and Social Science, University of Newcastle, Newcastle, Australia

ABSTRACT
In this study, we provide a quantitative analysis of prose and verse in the classical Chinese novel, *Dream of the Red Chamber (DRC)*, and discuss the implications for the disputed authorship of the novel. Firstly, we examine the amount of verse in across the chapters of *DRC*, and compare the style of the verse and prose portions of *DRC*. Secondly, a Principal Component Analysis (PCA) of *DRC* is performed based on the prose portions of the novel. Lastly, we discuss the implications of our experimental results for authorship attribution as well as descriptive stylistic analysis of *DRC*. Our authorial analysis largely confirms the findings of some previous studies that the novel has two authors. Meanwhile, stylistic analyses of the prose portions of the novel yield new and interesting results, which demonstrates that stylometric tools can be used to facilitate descriptive studies of classical Chinese literature.

1. Introduction

Dream of the Red Chamber (henceforth referred to as *DRC)*, also known as *The Story of the Stone*, is revered as the most prominent of China's four great classic novels. It depicts the rise and fall of four aristocratic families in ancient China from the perspective of the major character Jia Bao-Yu, featuring the romantic tragedies of the younger generation in these noble families. *DRC* has long been regarded as an encyclopaedia of feudal China and probably the most influential of all Chinese novels (Mair, 2001). With the prominent position of *DRC* in Chinese literary scene, studies of the novel have developed into a particular academic field – Red-ology (Spence, 1990, pp. 106–110).

One unresolved question in Red-ology is the authorship of *DRC*. A traditionally accepted view according to Hu (1921) is that the first 80 chapters of the novel (henceforth referred to as Part A) were written by one author, Cao Xue-Qin, while the remaining 40 chapters (henceforth referred to

CONTACT Hugh Craig hugh.craig@newcastle.edu.au; craig1623@gmail.com

as Part B) were by a different author, Gao E. This conclusion was based mainly on qualitative analysis. In recent decades, quantitative means have been employed for authorship attribution of *DRC*. Chan (1981) may be the trail-blazer in this line of research. Chan (1981) divided the texts of *DRC* into three parts, conducted statistical correlation tests of nouns, verbs, and adjectives between these groups, and concluded that Part A and Part B were written by the same author. Another study by Cao (1985) also claimed that the novel has only one author based on an analysis of Chinese functional words.

However, more studies seemingly lend support to Hu (1921). For example, Chen (1987) adopted a frequency-based method for authorship attribution of *DRC*. Eighty-eight linguistic features were selected from the novel, including those at word level such as typical adverbs and interjections, as well as those at sentence level such as sentence length. His findings showed that the distribution of most linguistic features selected are different in Part A and Part B, contradicting the claim of Chan (1981). More recently, some studies have taken new approaches. For example, Tu and Hsiang (2013) adopted a text-mining approach, Du (2017) used a topic modelling approach and cluster analysis, and Fang (2017) applied a new linguistic index named 'motif'. All three studies suggested that the novel may have been composed by more than one author.

This authorial controversy has also manifested itself through the publishing history of *DRC*. Published editions have sometimes credited it to two authors, and sometimes to one. To be specific, there are at least three different ways that the authorship of *DRC* is assigned in published editions. The first is to identify Cao Xueqin as the only author (e.g. editions published by Cutural Books in 1970, China Times Publishing Company in 2016 and Chinese Overseas Publishing House in 2017). The second is to attribute the novel to two authors, Cao Xueqin and Gao E (e.g. editions published by Lejn Book in 1983, The Commercial Press in 2016, and People's Literature Publishing House in 2018). The third is to assign the authorship to Cao Xueqin, and acknowledge that compilation and revision were made by Gao E and Cheng Weiyuan (e.g. the Laureate Publishing House edition published in 1983).

A second aspect of Red-ology is the discussion of the role of verse in the novel. Many lyric poems are included in *DRC*, and offer a reflective, emotionally tinged perspective on the action recorded in the narrative. The characters engage in numerous poetry competitions and their creations are presented in the text. Each chapter begins with a couplet summarizing its content with epigrammatic brevity. These aspects of *DRC* have been discussed in a number of qualitative studies (Li, 2015; Ma, 2018; Soong, 1977; Wu, 2010).

It is noteworthy, however, that the comparison between verse and prose has been neglected in quantitative studies of *DRC*. Few have noted the possible stylistic difference between prose and verse, and none has conducted a stylometric analysis of the verse in *DRC*. Given that verse is a highly developed and specialized genre in classical Chinese literature, it is reasonable to assume that prose and verse carry distinct stylistic features. Work in other languages, such as English, has shown that verse and prose when used in the same literary work introduce important stylistic differences (Craig & Greatley-Hirsch, 2017). In this regard, further investigation is needed, in light of the possible prose-verse distinction, to verify the results from some of the aforementioned studies (Du, 2017; Fang, 2017; Tu & Hsiang, 2013).

In this paper we aim to contribute to Red-ology by providing summary statistics on the amount of verse in the main parts of the chapters of *DRC*; an estimate of the difference in style between the verse and prose portions of *DRC*; and an authorship study of *DRC* based on the prose portions of the novel.

2. Data

There have been dozens of versions of *DRC* texts in existence since its first printed edition in 1791. Following Du (2017) and Tu and Hsiang (2013), the present study employs the *DRC* texts provided by Yuanze University (downloaded from http://cls.hs.yzu.edu.tw/hlm/), which is believed to be the closest to the original version (Tu & Hsiang, 2013).

The downloaded data is composed of 120 files in plain text form, with each file corresponding to a chapter. The specific steps for the pre-processing of the data are as follows. Firstly, all punctuation marks are removed. Secondly, we manually marked out the verse portions in all 120 chapters, and save the prose portions and verse portions into two separate files, as the data for the comparison between prose and verse. Thirdly, all the prose portions from Part A are extracted and saved into a file, while those from Part B are saved into another file, thus providing the data for comparison between Part A and Part B.

Since *DRC* was written in traditional Chinese characters, we retain the characters in their original form without simplification. Traditional Chinese characters are currently not in use in Mainland China, but are still widely used in Hong Kong and Taiwan, and can be processed with UNICODE. Lemmatization, which is often a necessary step in pre-processing European languages, is not conducted in the present study because Chinese is an analytic language without inflections.

It should also be noted that, in many cases, tokenization needs to be conducted in studies on Chinese corpora, to convert a sequence of running characters into words. In the present study, however, we carry out experiments based on characters rather than words. That is, tokenization is not done in the present study, and we have two reasons for so doing. Firstly, tokenization tools are usually trained with modern Chinese texts (Chang, Galley, & Manning, 2008). Though these tools have achieved acceptable accuracy when working with modern Chinese, it should be noted that early modern Chinese is sharply different from modern Chinese in terms of lexis, syntactic structure, and semantics. Therefore, it would not be appropriate to apply modern Chinese tokenization tools to ancient Chinese (Han, Wang, Zhang, Fu, & Liu, 2018; Lee, 2012). Secondly, the character-based analysis will inherit quantitative features from the word-based analysis. For example, if an author has a strong preference for using a two-character word, 熙凤(Xi-Feng, the name of a female character in *DRC*), then when we conduct character-based analysis of texts written by this author without tokenization being done, the result would naturally show a frequent occurrence of both the characters 熙(Xi) and 凤(Feng). Therefore, for stylometric analysis of *DRC*, we believe using characters as the basic linguistic unit is an acceptable if not a perfect approach.

3. Methods

3.1. Principal Component Analysis

In this study, we mainly employ Principal Component Analysis (PCA) to analyse prose and verse in *DRC*. PCA is a widely used statistical tool for data mining and visualization (Abdi & Williams, 2010). The primary aim of the method is to reduce the dimensions of a data set that consists of a large number of interrelated variables, while preserving as much of the variation present in the data set as possible. That is, PCA transforms a large number of variables into a smaller set of new variables, which represent most of the variation in the original variables. These new variables, named principal components, are linear combinations of the original ones and are independent of each other. Data visualization can then be achieved by plotting the most important principal components in two- or three-dimensional charts. More technical details of PCA can be found in Hastie, Friedman, and Tibshirani (2001).

For a broader audience who need only a general understanding of this technique, Krzanowski (1988, pp. 53–56) provided a classical example for illustration of PCA. If there is a table involving the height and weight data of a group of people, we may generate a new variable *size* by simply summing the height and weight. The *size* may well retain much of the variation of the two original variables, since height and weight are often interrelated. As a general tendency, shorter people tend to be lighter, and taller people heavier. Therefore, by transforming the two variables into a new one, we could extract most of the information conveyed by the original data set, while using only one variable. This example shows the process of data reduction in the simplest form. Normally, there will be far more than two variables to be reduced to several principal components.

In the present study, we use the R function *prcomp* for PCA, using the correlation matrix of the variables, and the R package *ggfortify* for visualization. Our observations will be based on frequencies of characters. Thus, the process of dimensional reduction process in our specific case is illustrated in Figure 1. Suppose that we have an m-row n-column matrix with m observations of the frequency of n characters: $f_1, f_2, \ldots,$ f_n. After the mathematical process of dimension reduction, we will reduce the n variables to p ($p < n$) independent principal components which capture a large amount of the variance in the original table. The two most important principal components, PC1 and PC2 will be chosen to visualize data in a two-dimensional chart.

3.2. Sample Size and Number of Features

Figure 2 illustrates the specific steps of PCA using MFCs in *DRC*. Suppose that we have two corpora to be analysed, namely, $Corpus_1$ with a total of T_1 characters and $Corpus_2$ with T_2 characters. First, we divide $Corpus_1$ into M_1 non-overlapping segments of S running characters, and $Corpus_2$ into M_2 non-overlapping segments of S running characters, discarding in both cases any smaller segments that come at the end. Thus, we have:

$$M_1 = [T_1/S] \tag{1}$$

$$M_2 = [T_2/S] \tag{2}$$

where M_i is the largest whole number smaller than T_i/S.

Secondly, we generate a list of the N most frequent characters (MFCs) from the $M_1 + M_2$ segments. Thirdly, we calculate the frequencies of the N MFCs based on their occurrences in each of the $M_1 + M_2$ segments. Lastly, PCA is conducted on the $M_1 + M_2$ segments, with the frequencies of the N MFCs as features, and, in visualization, each segment is coloured according to whether they come from $Corpus_1$ or $Corpus_2$.

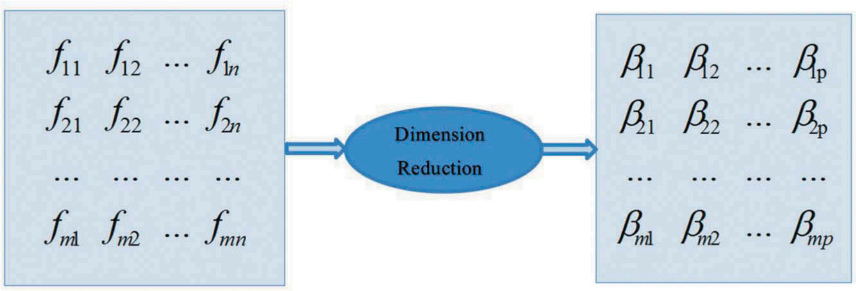

Figure 1. Dimensional reduction in Principal Component Analysis.

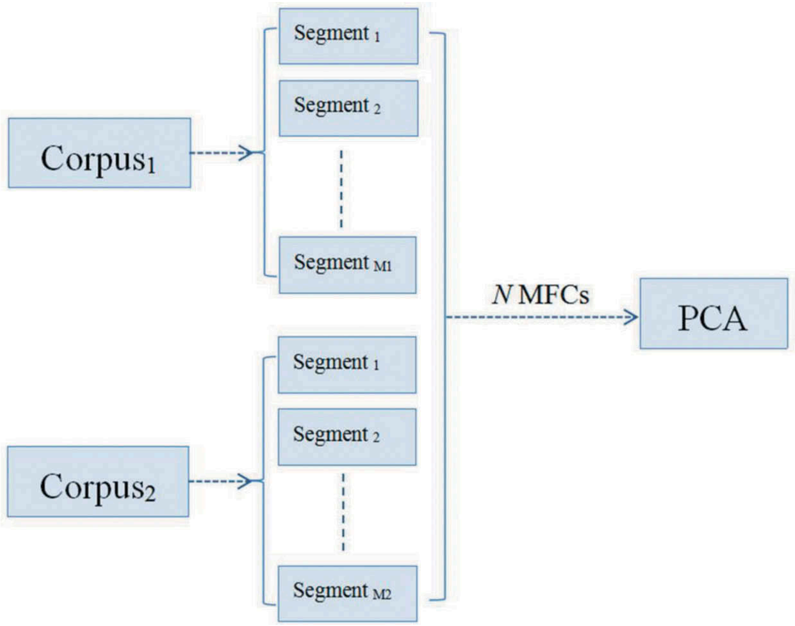

Figure 2. Using MFCs for Principal Component Analysis of *DRC*.

Two points are worth discussing. The first is about the sample size, i.e. the *S* in equations (1) and (2). According to previous studies, the minimum sample size for reliable stylometric analysis ranges from some 1000 words to 5000 (e.g. Burrows, 2002; Eder, 2015, 2017; Holmes, Gordon, & Wilson, 2001; Jockers, Witten, & Criddle, 2008; Luyckx & Daelemans, 2011), depending on language used and genre. Although Chinese is not included in these studies, the findings therein can serve as a reference for our study. Hence, we carry out PCA on different sample sizes, such as 1000, 2000, 3000, and 5000 characters, so as to improve the validity of our findings.

The second is about the number of features, i.e. the *N* in Figure 2. Drawing upon the approach of previous studies (Craig & Greatley-Hirsch, 2017; Eder, 2017), which adopt round numbers of features for stylometric classification, we use 100 MFCs as features. Choosing a short list of MFCs in this way has at least two advantages. First, staying with the more frequent features offers abundant observations. Second, using a round number of features helps to avoid problems such as over-fitting and cherry-picking. Besides, we also try several different variable sets, such as 50, 75, or 150 MFCs, to check that there is no undue variation between these cases.

4. PCA of Prose versus Verse

In this section, we will compare the styles of prose and verse in *DRC*. What motivates us to do this is that previous stylometric studies on *DRC* did not distinguish between prose and verse. Mixing the two genres of prose and verse may lead to bias in the stylometric analysis of *DRC*, in that this approach neglected the possible influence from varying ratio of verse to prose in different chapters on stylistic analysis. In fact, when we calculate the proportion of verse, the results show that Part A contains more verse than Part B (See Table 1). Given such a difference in verse proportion, if the hypothesis stands that prose and verse have distinct styles; then, perhaps a better approach is to treat prose and verse separately in authorship identification.

Intuitively, we would expect that there should naturally exist stylistic differences between prose and verse. However, there have been previous studies showing that this assumption is not always valid. For example, Craig and Greatley-Hirsch (2017), already cited, compared the styles of prose and verse in early English drama, and found that 'comedies in prose are not immediately distinguishable stylistically from comedies in verse' while the prose and verse portions within comedies which mix prose and verse have some marked stylistic differences (Craig & Greatley-Hirsch, 2017, p. 69). Therefore, we conduct a comparative analysis of verse and prose in *DRC* to find out if they are distinct in styles.

To do this, we first extract all verse texts from *DRC*, which total 11,012 characters. We can divide the verse texts into eleven 1,000-character, five 2,000-character, three 3,000-character, or two 5,000-character segments, discarding any remaining smaller segments. Accordingly, we randomly extract from the prose texts of *DRC* eleven 1,000-character, five 2,000-character, three 3,000-character and two 5,000-character segments (all segments are non-overlapping running characters). These samples allow us to perform PCA four times with four different sampling sizes. For each PCA,

Table 1. Proportion of verse across *DRC*.

	Number of Characters in Prose	Number of Characters in Verse	Proportion of Verse
1–10	54,544	2563	4.70%
11–20	52,647	660	1.25%
21–30	58,706	1517	2.58%
31–40	56,735	1274	2.25%
41–50	58,777	910	1.55%
51–60	66,505	370	0.56%
61–70	67,482	886	1.31%
71–80	68,637	2160	3.15%
81–90	58,984	395	0.67%
91–100	56,916	153	0.27%
101–110	55,964	40	0.07%
111–120	61,026	84	0.14%

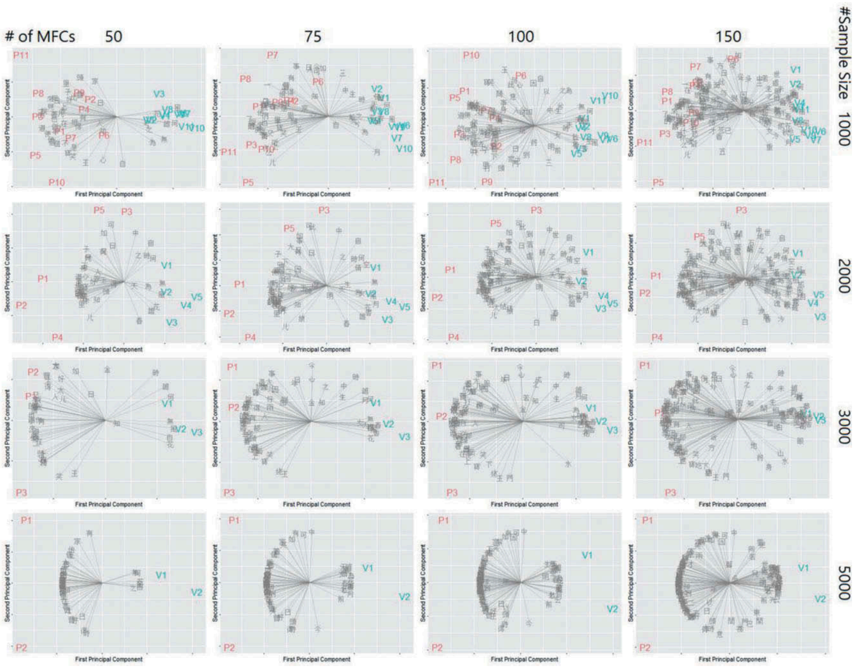

Figure 3. PCA bi-plots of prose and verse in *DRC* (Prose are coloured in red and verse in green).

we use 50, 75, 100, and 150 MFCs as features, respectively. Thus, PCA is conducted for a total of 16 times. Figure 3 shows the scatter-plots of the experimental results.

Obviously, the results show a clear-cut distinction between prose and verse in *DRC*, with a concentration of prose segments on the left, and verse on the right. The prose segments are close to characters only used for narrative purposes such as 了(modal particle signifying the change of situation), 的(ablative cause suffix), 這(this or these in oral Chinese), and 那(that or those in oral Chinese), while the verse segments are close to common images such as 花(flower), 風(wind), 月(the moon) as well as 兮(interjection mainly used in ancient poems). Thus, it can be concluded that prose and verse in *DRC* do demonstrate distinct styles. Given the uneven distribution of verse part across the novel as shown in Table 1, it is necessary to treat prose and verse separately in authorship attribution.

5. PCA of Prose in Part A and Part B

In this experiment, prose parts of DRC are extracted, totalling 716,923 characters (excluding punctuation marks). 484,033 characters are from

Part A, and 232,890 from Part B. With incomplete tail segments cut off, we have 232 non-overlapping 1,000-character segments from Part B. These can also be divided into 116 2,000-character, 77 3,000-character or 46 5,000-character segments. Similar to the sampling approach in Section 4, we randomly select 232 1,000-character, 116 2,000-character, 77 3,000-character and 46 5,000-character segments from prose in Part A, to make comparable samples to those from Part B. Then, PCA is performed 16 times, with 50, 75, 100, and 150 MFCs as features, respectively. The results are summarized in the matrix in Figure 4.

Two observations can be made from Figure 4. Firstly, segments from Part A and Part B are generally plotted apart. While there is considerable overlap between Part A and B in the first row, the two Parts are plotted further apart as the sample size increases to 3,000 or 5,000 characters. This shows that prose from Part A and B do have distinct styles. Secondly, while larger sample size sets the segments from the two Parts further apart, the choice of feature numbers seemingly makes little influence. This shows that a larger sample size (>3000 characters) may better reveal the stylistic differences between these segments. Meanwhile, the increase in a number of features from 50 to 150 MFCs adds limited if any stylometric information.

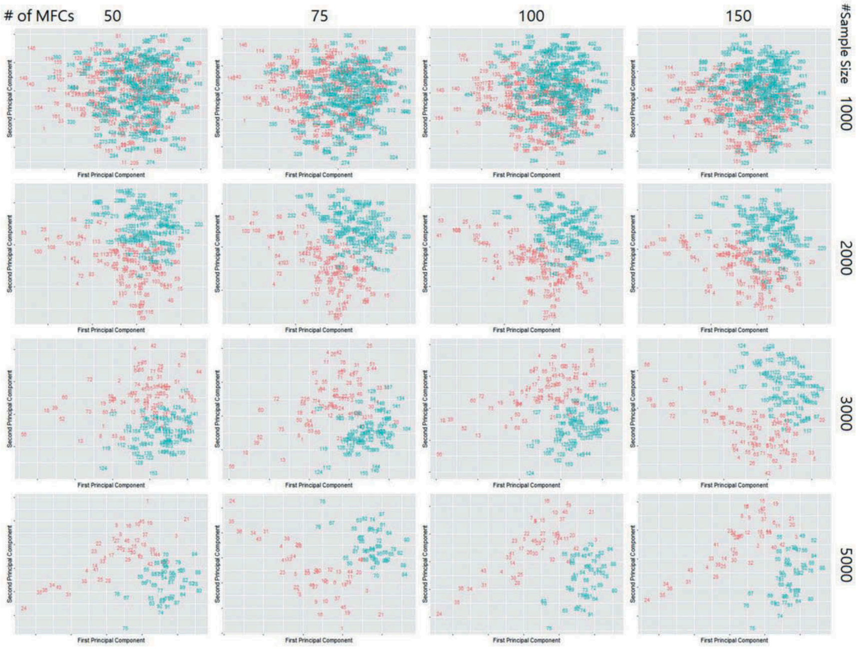

Figure 4. PCA bi-plots of prose from Part A and Part B in *DRC* (Prose from Part A are coloured in red and those from Part B in green).

6. Statistical Tests for Stylistic Analysis of Prose in Part A and Part B

It can be seen from Section 5 that PCA helps uncover and visualize the stylistic distinctions between Part A and B. To provide more statistical details, and to analyse the linguistic preference of Part A and B (in terms of usage of characters), we apply *t*-tests to compare the frequencies of MFCs in the two parts.

To this end, we employ as data the 46 5,000-character segments from Part A and the 46 5,000-character segments from Part B, which were already used in Section 5 for PCA. A list of 100 MFCs is generated from the 92 segments in total. For each of the 100 MFCs, *t*-tests are applied to compare their frequencies in the 46 segments from Part A and the 46 segments from Part B. As a result, 53 out of the 100 MFCs show a significant difference between Part A and Part B (See Figure 5 for the 100 MFCs and their P-values).

To make a comparison, we shuffle the 92 segments to sequence them in random orders, and apply t-tests the same way as mentioned above to examine the differences in frequencies of the 100 MFCs in the first 46 segments and the last 46 segments in the new sequence. We repeat this process 50 times. See Tables 2 and 3 for the number of MFCs showing significant differences for each time and their descriptive statistics.

It can be seen that in the shuffled blocks, a significantly smaller number of the MFCs yield P-values that are lower than 0.05: the mean is 5.28 with a standard deviation of 3.01, and none of the 50 randomization experiments reveals more than 15 MFCs showing significant differences between the first 46 segments and the last 46. In comparison with 53 MFCs showing significant differences in the original unshuffled 92 blocks, this confirms that the first 80 and the last 40 chapters do have distinct stylistic features.

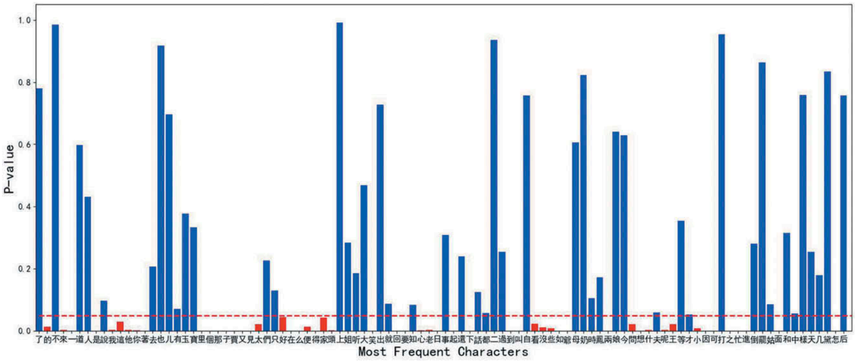

Figure 5. The 100 MFCs and corresponding P-values when applying t-tests to character frequencies in Part A and B (Blue bars signify P-values above 0.05 and red bars signify P-values below 0.05; the red dash line indicates the statistical significance level of 0.05).

Table 2. Numbers of MFCs showing significant differences for the 50 shuffled sequences.

13	4	7	2	4	6	4	5	4	7
4	3	6	2	2	1	3	7	5	7
5	0	3	3	8	6	5	3	2	8
14	10	7	3	3	2	9	4	6	5
1	5	4	6	11	6	5	11	5	8

Table 3. Descriptive statistics of the number of MFCs showing significant differences.

N	Minimum	Maximum	Mean	Standard Deviation
50	0	14	5.28	3.01

For a more detailed stylistic description of the novel, we pick out the 53 characters with P-values smaller than 0.05, and investigate their frequencies in the 46 segments belonging to Part A and in those from Part B. See Appendix for the frequencies of the 53 characters in Part A and Part B of the novel, and the ratio of these frequencies in the two parts. Several points are noteworthy in the findings.

The first is that conjunctions are used more frequently in Part A, for example, 因(because), 又(and, in addition) 正(a polysemy which, in most cases, means while/when) and 如(a polysemy which, in most cases, means if). This suggests that in terms of writing style, the author of Part A paid more attention to logical links and cohesion, while Part B is featured by a relatively loose writing style with less use of conjunctions.

More noticeable, however, is that more pronouns are used in Part A than in Part B, such as 你(second-person pronoun), 我(first-person pronoun), and 他(third-person pronoun), perhaps showing a tendency to emphasize individuality. In contrast, characters such as 賈(Jia, the family name), and 家(family) appear more frequently in Part B, which indicates that when writing Part B, the author may have placed more value on the honour and dignity of the aristocratic family rather than the individuals. In addition, Part B also contains more honorifics such as 爺(Lord) and 太(Duchess). These together present a contrast between the pursuit of individuality and the hierarchy of the family. This contrast is also consistent with the different ways that Jia Bao-Yu, the hero, is portrayed in the earlier and later parts of the novel. Such a comparison between individuality and hierarchy might be another form of what previous studies called the ‘dual worlds’ of Jia Bao-Yu, ‘worlds of innocence and experience, youth and non-youth, utopia and reality, idealism and realism’ (Edwards, 1990, p. 69). These contrasts correspond to the differences between the characteristic preferences and emphases of one author and another which have been detected elsewhere on the basis of contrasting profiles of function word use (e.g. Craig, 1999).

Our study is a quantitative one and we rely on the numerical evidence presented above for our contention that Part A and Part B have different authors. However, there are some aspects which are not susceptible to stylometry, which focuses more on tendencies maintained over large extents, but are evident to the unaided reader of the novel paying attention to the more local level. These are worth adducing in a more comprehensive view of the authorship problem. For instance, the changes that occur on Jia Bao-yu are often too abrupt to be part of a natural progression of the narrative by a single author. In the 80th Chapter, Jia Bao-Yu is portrayed as a fun-loving dandy who is eager for freedom from confinement, and awkward when allowed to play outside. In the following Chapter 81, however, he declares that 'I shall follow in the footsteps of Old Sire Jiang'(Cao, 1973), a metaphor indicating that he is willing to become a politician and help to govern the country.[1] Such a contradiction between two adjacent chapters is hard to explain, if the statement stands that they belong to the same author.

In the area of attitudes and behaviours, again more evident to the reader than to the stylometrician, there are a number of logical inconsistencies when one puts together Part A and B. For instance, before the 80th chapter, Jia Bao-Yu is described as a rebel against the feudal family, a playboy who is not interested in Confucian classics, detests arranged marriage, and would rather spend time in the company of the young females. Into the last 40 chapters, however, Jia Bao-yu devotes himself to learning and pursuit of a mundane life; he starts to respect Confucianism and filial piety; he seeks to become a government official; he even accepts an arranged marriage to Xue Baochai, all of which are logically impossible for his previous self.

7. Conclusion

To conclude, this study employed PCA to conduct stylometric analyses of *DRC*. In contrast to previous research, the present study has highlighted the division of the novel into prose and verse components and analysed them separately. Several findings of interest were revealed. Firstly, we examined the proportion of prose and verse in *DRC*, and found that Part A contains proportionally more verse than Part B. A PCA experiment on prose and verse in *DRC* showed distinct styles in the two genres, suggesting that it is necessary to distinguish the two genres in authorship attribution. Secondly, our analysis of the prose texts of *DRC* showed that Part A and Part B are different in writing styles, providing evidence to the claim that the novel has more than one author. Thirdly, we also performed a descriptive stylistic analysis and found that sentences in Part A are more logically coherent than those in Part B, and a contrast between individuality and hierarchy was found.

This study also provides implications for the choice of sample size in authorship attribution of literary works. As mentioned earlier, plenty of research on this topic has been conducted on other languages except for Chinese. Although it is not the theme of the present study to detect the optimum sample size for authorship attribution, our results do show that variation in sample size has an impact on the stylometric analysis of Chinese texts. Specifically, setting the sample size at 3000 characters per segment or above seems to produce better results in a PCA of the prose of *DRC* (See Figure 4). However, classifying prose and verse seemingly requires fewer characters per segment. With 1000 characters per sample, prose and verse show a clear-cut division on the PCA scatter-plot (See Figure 3). Clearly, more experiments on other texts apart from *DRC* are still needed in future research, for a definitive answer to the question of how to determine the minimum sample size for stylometric analysis of Chinese texts.

Overall, this study demonstrates that PCA can be used as an effective tool for the distant reading of the Chinese novel. We also prove that, as previous studies (e.g. Craig & Greatley-Hirsch, 2017) suggested, computational stylistic methods represented by PCA can be applied to the stylistic description as well as to authorship attribution, as shown in Sections 5 and 6. While plenty of studies have adopted computational means for the descriptive genre and style analysis of literary works in languages such as English (Craig & Greatley-Hirsch, 2017; Craig & Kinney, 2009), there is still a research gap in this regard when it comes to Chinese literary works. Therefore, it is our hope that the present study can serve as a reference and provide implications for future descriptive analysis of Chinese literary works using quantitative means.

Note

[1] Old Sire Jiang, or Jiang Tai Gong, was a famous politician in ancient China. Originally a fisherman, Jiang Tai Gong later became the Prime Minister of the Zhou Dynasty and helped it to take control of the entire country.

Acknowledgments

This work was supported by the National Social Science Fund of China (Grant No. 15BYY179) and the Australian Research Council (Grant No. DP160101527). The authors appreciate the reviewers' insightful comments and suggestions.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the National Social Science Fund of China [15BYY179] and the Australian Research Council [DP160101527].

ORCID

Haoran Zhu http://orcid.org/0000-0001-8219-6147
Lei Lei http://orcid.org/0000-0002-3366-1855
Hugh Craig http://orcid.org/0000-0002-9336-1678

References

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *WIREs Computational Statistics*, *2*(4), 433–459.

Burrows, J. (2002). 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, *17*(3), 267–287.

Cao, Q. (1985). The last 40 chapters of A Dream of Red Mansions is not written by Cao Xueqin. *Studies on A Dream of Red Mansions*, *7*(1), 281–312. (in Chinese).

Cao, X. (1973). *The story of the stone*. (D. Hawks, & J. Minford Trans.). Harmondsworth: Penguin.

Chan, B. (1981, June). *The authorship of the Dream of the Red Chamber based on a computerized statistical study of its vocabulary*. Paper presented at 1st International Forum on Dream of the Red Chamber, Madison, WI.

Chang, P. C., Galley, M., & Manning, C. D. (2008). Optimizing Chinese word segmentation for machine translation performance. *Proceedings of Workshop on Statistical Machine Translation*, *3*, 224–232.

Chen, D. (1987). The authorship of the last forty chapters: A mathematical linguistic perspective. *Studies on A Dream of Red Mansions*, *9*(1), 293–318. (in Chinese).

Craig, H. (1999). Authorial attribution and computational stylistics: If you can tell authors apart, have you learned anything about them? *Literary and Linguistic Computing*, *14*(1), 103–113.

Craig, H., & Greatley-Hirsch, B. (2017). *Style, computers, and early modern drama: Beyond authorship*. Cambridge: Cambridge University Press.

Craig, H., & Kinney, A. F. (2009). *Shakespeare, computers, and the mystery of authorship*. Cambridge: Cambridge University Press.

Du, K. (2017, August). *Authorship of Dream of the Red Chamber: A topic modeling approach*. Paper presented at 2017 Digital Humanities Conference, Montreal, Canada.

Eder, M. (2015). Does size matter? Authorship attribution, small samples, big problem. *Digital Scholarship in the Humanities*, *30*(2), 167–182.

Eder, M. (2017, August). *Short samples in authorship attribution: A new approach*. Paper presented at 2017 Digital Humanities Conference, Montreal, Canada.

Edwards, L. (1990). Gender imperatives in Honglou meng: Baoyu's bisexuality. *Chinese Literature: Essays, Articles, and Reviews*, *12*, 57–69.

Fang, Y. (2017). L-motif TTR for authorship identification in Hongloumeng and its translation. In H. Liu & J. Liang (Eds.), *Motifs in language and text* (pp. 87–108). Monton: De Gruyter.

Han, X., Wang, H., Zhang, S., Fu, Q., & Liu, J. S. (2018). Sentence segmentation for classical chinese based on LSTM with radical embedding. *arXiv Preprint arXiv, 1810.03479*, 1–8.

Hastie, T., Friedman, J., & Tibshirani, R. (2001). *The elements of statistical learning; Data mining, inference and prediction*. New York: Springer Verlag.

Holmes, D. I., Gordon, L. J., & Wilson, C. (2001). A widow and her soldier: Stylometry and the American civil war. *Literary and Linguistic Computing*, *16* (4), 403–420.

Hu, S. (1921). *Textual research on the Dream of the Red Chamber*. Beijing: Beijing Publishing Group. (in Chinese).

Jockers, M. L., Witten, D. M., & Criddle, C. S. (2008). Reassessing authorship of the Book of Mormon using delta and nearest shrunken centroid classification. *Literary and Linguistic Computing*, *23*(4), 465–491.

Krzanowski, W. J. (1988). *Principles of multivariate analysis: A user's perspective*. Oxford: Clarendon Press.

Lee, J. (2012, April). A classical Chinese corpus with nested part-of-speech tags. In *Proceedings of Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities 2012* (pp. 75–84). Red Hook, NY: Curran Associates, Inc.

Li, Y. (2015, November). *An investigation into the hybridity phenomenon of poetry translation in Hong Lou Meng*. Paper presented at International Conference on Education, Language, Art and Intercultural Communication, Kaifeng, China.

Luyckx, K., & Daelemans, W. (2011). The effect of author set size and data size in authorship attribution. *Literary and Linguistic Computing*, *26*(1), 35–55.

Ma, S. (2018, March). *A case study of chrysanthemum inscriptions and poems on the tragic deepening of the the story of the stone with the talented beauty novel in the early Qing dynasty*. Paper presented at International Conference on Arts, Linguistics, Literature and Humanities, Shenzhen, China.

Mair, V. H. (2001). *The Columbia history of Chinese literature*. New York: Columbia University Press.

Soong, S. C. (1977). Types of misinterpretation: Some poems from Red Chamber Dream. *Renditions: A Gateway to Chinese Literature and Culture*, *7*, 73–92.

Spence, J. (1990). *The search for modern China*. New York: W. W. Norton & Company.

Tu, H. C., & Hsiang, J. (2013, July). *A text-mining approach to the authorship attribution problem of Dream of the Red Chamber*. Paper presented at 2013 Digital Humanities Conference, Lincoln, NE.

Wu, S. (2010). *A verbal paradise visualized: An ekphrasistic study of the Daguanyuan in Cao Xueqin's Hongloumeng* (Doctoral dissertation). Retrieved from https://docs.lib.purdue.edu/dissertations/AAI3444792/

Appendix. Frequencies of the 53 MFCs in Part A and Part B

Character	Frequency in the sample of Part A (F_a)	Frequency in the sample of Part B (F_b)	F_a/F_b
笑	1553	571	2.720
之	786	293	2.683
忙	694	343	2.023
因	711	406	1.751
兩	794	504	1.575
面	583	389	1.499
可	661	447	1.479
小	688	475	1.448
如	845	587	1.440
日	1039	734	1.416
子	1893	1424	1.329
又	1835	1411	1.300
你	2586	1998	1.294
下	967	748	1.293
個	1933	1521	1.271
他	2710	2221	1.220
我	3262	2709	1.204
一	4020	3371	1.193
這	2617	2354	1.112
來	3523	3905	0.902
的	4921	5457	0.902
好	1256	1449	0.867
沒	704	844	0.834
是	3050	3683	0.828
看	705	861	0.819
便	1189	1454	0.818
問	577	709	0.814
頭	1051	1299	0.809
家	1090	1358	0.803
些	646	818	0.790
在	1173	1486	0.789
要	875	1137	0.770
起	754	981	0.769
見	1384	1805	0.767
得	1088	1425	0.764
什	542	713	0.760
太	1347	1830	0.736
心	785	1068	0.735
呢	523	719	0.727
進	432	595	0.726
著	1721	2371	0.726
那	1396	1934	0.722
回	858	1217	0.705
老	750	1078	0.696

(*Continued*)

(Continued).

Character	Frequency in the sample of Part A (F_a)	Frequency in the sample of Part B (F_b)	F_a/F_b
想	514	777	0.662
里	1384	2181	0.635
怎	334	527	0.634
王	463	738	0.627
么	976	1658	0.589
叫	586	1043	0.562
賈	1177	2111	0.558
到	530	1110	0.477
爺	426	994	0.429