Research Article
A Quantitative Study on Dream of the Red Chamber: Word-Length Distribution and Authorship Attribution

Yue Yu,1 Wei Liu,2 and Ying Feng3

1School of Foreign Studies, University of Science and Technology Beijing, Beijing 100083, China
2School of Languages and Communication Studies, Beijing Jiaotong University, Beijing 100044, China
3Software Engineering Institute of Guangzhou, Guangzhou 510990, China

Correspondence should be addressed to Wei Liu; wliu@bjtu.edu.cn

Received 25 January 2022; Revised 12 February 2022; Accepted 18 February 2022; Published 10 March 2022

Academic Editor: Ning Cai

Copyright © 2022 Yue Yu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This paper investigates the distribution characteristics of word lengths in the Dream of the Red Chamber (DRC), measured in terms of the number of syllables or characters. The results show that the frequency distribution of words of different lengths in the DRC abides by the extended logarithmic distribution model. A comparison between the first forty, the middle forty, and the last forty chapters shows that the distribution of word lengths in these three parts does not differ significantly, which sheds light on the authorship of the novel.

1. Introduction

Considered to be the pinnacle of classical Chinese novels, Dream of the Red Chamber (DRC, also known as The Story of the Stone), has long been a favorite topic of discussion. The novel narrates the decline of a powerful Chinese family and vividly portrays the late imperial Chinese culture. With a perceptive and comprehensive observation of life and society in the 18th century, it is regarded as an encyclopedia of feudal China, drawing the attention of numerous researchers [1–9].

In recent decades, quantitative studies on the novel have attracted much attention. Many researchers use statistical methods to compare the first eighty chapters with the remaining forty, investigating whether it was the same author who composed them or if there were two different ones. This has long been a controversial issue. For example, Karlgren [10] compared the occurrence of thirty-two grammatical and lexical phenomena in the first eighty chapters with the remaining forty and concluded that they had one single author. Chan [11] calculated the word correlativeity between these two parts, including nouns, verbs, adjectives, adverbs, and function words, and reached a similar conclusion. Li and Li [12] conducted a statistical analysis of adverbs in DRC and also argued in support of single authorship.

However, many researchers claim that these two parts were composed by two or more different authors [13–18]. Li [13], for instance, calculated the frequencies of forty-seven function words in each chapter and performed a cluster analysis, which showed that the novel had even more than two authors. Wang [14] studied more than a hundred words and found clear diction differences between the two parts, claiming that more than one author had been involved. More recently, Zhu et al. [19] conducted a Principal Component Analysis on the prose portions of the novel, confirming the two-author claim.

Although many quantitative studies have been published on DRC, these studies have focused on the examination of specific words and their frequencies, such as function words and high-frequency words. Little is known about the overall word-length distribution in DRC. Some researchers believe that the frequency distribution of words of different lengths could shed light on the authorship problem. Mendenhall [20] compared the works of Shakespeare, Bacon, and Marlowe and found that the distribution pattern of word lengths in Shakespeare’s work was consistently different from Bacon’s. He claimed that it was implausible that Bacon
would have written works attributed to his more famous contemporary. However, Williams [21] argued that a difference in literary presentation, that is, a genre difference, could explain the differences in word-length distribution found by Mendenhall. Since then, word-length distribution has attracted the attention of more scholars.

Much research has been done on word-length distribution (WLD) in different languages [22–28]. The frequency distribution of words with different lengths would not be chaotic but follow specific rules. Moreover, two families of distributional models, Poisson and binomial, could fit most previously studied human languages [29]. Although much ink has been spilled on the topic of WLD, much of the work has been conducted on Indo-European languages. There has been less focus on Chinese, except for the recent studies by Wang [30], Chen [31], Chen and Liu [29], and so on.

Although it is questionable whether WLD can distinguish different authors, many researchers agree that it is mainly influenced by boundary conditions such as authorship, language, genre, size of texts, and time of emergence (see, e.g., [32]). Therefore, one can assume that if language, genre, size of texts, and time of creation are adequately controlled, authorship is most likely the factor responsible for the difference in WLD. In other words, for the first eighty chapters and the remaining forty chapters of DRC, where genre and creation time are consistent, if we choose texts of the same length, the difference in WLD could be attributable to a difference in authorship.

According to the above analysis, our research questions are as follows:

(i) Question 1: which model best fits the word-length distribution in the DRC?
(ii) Question 2: is there a significant difference in the word-length distribution between the first eighty chapters and the remaining forty chapters?

This paper addresses these questions based on a statistical analysis of the distribution frequencies of words of different lengths in each chapter and three groups of chapters (1 to 40, 41 to 80, and 81 to 120). The organization of this article is as follows: Section 2 describes the data and methods used in the study. Section 3 presents the results. Section 4 discusses the authorship attribution based on the findings. Section 5 concludes the paper.

### 2. Data and Methods

Following many previous studies [19, 33, 34], the data for the current study were obtained from the DRC text provided by Yuanze University (https://cls.hs.yzu.edu.tw/hlm/), because this version is considered to be the closest to the original text. To obtain homogeneous text samples, we extracted a continuous body of text of 2000 words from each chapter. Following Wang [30] and Deng and Feng [35], we measured a word in terms of the number of characters, which in Chinese is basically equal to the number of syllables. We did the sample selection and the word length calculation using a Python script we programmed. In the end, 120 sample texts were retained, each consisting of 2000 words. In order to investigate whether there were significant differences between the first eighty and the remaining forty chapters, the 120 sample texts were divided into three even parts, namely, Part I (1–40 texts), Part II (41–80 texts), and Part III (81–120 texts).

The Altmann-Fitter 3.1 software was applied to fit the data obtained from these 120 sample texts to determine the best-fitting probability distribution model. Altmann-Fitter, widely used in quantitative linguistics [36], contains over 200 individual probability distributions and can automatically choose the best-fitting model. The goodness-of-fit was tested using the Chi-square test $\chi^2$ or the discrepancy coefficient $C = (\chi^2/N)$. If $P \geq 0.05$, or in the case of long texts, $C < 0.02$, the result is considered satisfactory [32, 37]. In addition, the determination coefficient $R^2$ was also used to analyze the fitting result. These parameter values can be easily obtained with the Altmann-Fitter. Moreover, SPSS was used to conduct the significance test of the difference.

### 3. Results

#### 3.1. Word-Length Distribution in DRC

To answer research question one, namely, which model best fits the word-length distribution in DRC, all probability distributions were fitted to 120 sample texts using the Altmann-Fitter. Based on the values of...
$P(x^2)$, the best result was provided by the extended logarithmic $(\theta, \alpha)$, where 99 out of 120 sample texts showed a satisfactory fitting result, among which 81 texts showed a very good fitting result ($P \geq 0.05$) and 18 texts presented an acceptable result ($0.01 \leq P \leq 0.05$). The fitting results of the model to the word-length distribution in six sample texts (randomly selected from our 120 sample texts) are shown in Table 1 and Figure 1.

For the six samples illustrated above, only text 40 has a negative fitting result, where $P(x^2) = 0.0011$, though the values of $C$ and the coefficient of determination $R^2$ are good. Further detailed exploration of Table 1 and Figure 1 shows that in DRC, word length measured by syllable or character numbers mainly ranges from 1 to 4 and most words consist of one or two characters.

### 3.2. Word-Length Distribution in Different Parts of DRC

In order to investigate whether there are significant differences between the first eighty and the remaining forty chapters in terms of word-length distribution (research question two) and to further investigate the problem of authorship, we divided the 120 sample texts into three equal parts. We compared them from two perspectives: mean word length and probability distribution model.

#### 3.2.1. Mean Word Length

Both dynamic and static mean word lengths were calculated, based on token and type, respectively, according to the following formulas.

**Dynamic mean word length:**

$$\text{DMWL} = \frac{\sum_{i=1}^{n} X_i F_i}{\sum_{i=1}^{n} F_i}$$  \hspace{1cm} (1)

**Static mean word length:**

...
3.2.2. Probability Distribution of Word Length in the Three Parts. We examined the static (based on type) and dynamic (based on token) word-length distribution of these three parts, and no significant differences were found. Using the Altmann-Fitter, we found that eight models produced very good fittings when the static word-length distribution was considered, as shown in Table 3.

Consider the values of $C$ and $R^2$ and the number of parameters, extended logarithmic ($\theta$, $\alpha$) is the most appropriate model to capture our data in Parts I, II, and III. As other researchers noted, according to Occam’s Razor, models with fewer parameters are preferable [31, 35]. This fitting result corresponds to those in Section 3.1, obtained on the basis of individual sample texts. Table 4 and Figure 2 show the fitting effects of the extended logarithmic model for the three parts of DRC.

Moreover, when dynamic word-length distributions are considered, twelve models show very good fittings, with 0.9992 as the lowest value of the determination coefficient $R^2$ and 1.0000 as the highest, as can be seen in Table 5.

Regarding the values of $R^2$ and $C$ as well as the number of parameters, as mentioned earlier, extended logarithmic ($\theta$, $\alpha$) presents the best-fitting result, which is consistent with the previous analysis, as illustrated in Table 6 and Figure 3.

An analysis of variance (ANOVA) was performed using SPSS to test the significance of the difference between each group of observed data in pairs, including the static and dynamic distribution data, as already shown in Tables 4 and 6, respectively. The significance degree is usually judged using the $P$ value. Typically, $P < 0.05$ is regarded as a statistically significant result, and $P < 0.01$ is regarded as a statistically highly significant result. In our tests, the $P$ values ($P > 0.05$ for all cases) further show no significant differences between the three parts in terms of word-length distribution data.

4. Discussion

The above analysis allows us to obtain an overview of the word-length distribution in DRC. Based on the fitting results of both individual sample texts (Section 3.1) and groups of sample texts (Section 3.2), the best-fitting model for the
word-length distribution in DRC, measured in terms of syllable or character, would be extended logarithmic \((\theta, \alpha)\). This finding is inconsistent with some previous studies [31, 38], which have shown that mixed Poisson provides the best-fitting results for written Chinese. The inconsistency may be due to the difference in units of measurement used, as in their study, the researchers took components as the unit of measurement, while in the present study, we take syllable/character as the unit of measurement. Chen and Liu [29] point out that the most appropriate unit of measurement for written Chinese are components. “The component is the constructing units of characters which have more than one stroke” [29:10]. In their study, measuring written Chinese based on characters did not yield satisfactory results. The current study suggests that defining Chinese words in terms of characters could also be appropriate. Besides, this study indicates that there may not be a uniform best-fitting model for word-length distribution in written Chinese, as they supposed. In other words, works composed by different authors may conform to different models.

As shown in Section 3.2, the differences between the three parts of DRC yielded interesting results. We have noticed some differences in the mean word length between these parts, while no significant differences were found in word-length distribution.

As mentioned earlier, opinions are divided on whether word length properties can identify different authors. Mathematician and logician de Morgan believes so (see, e.g., [36]). It is believed that if two texts are written by different authors, even on similar topics, the difference between average word lengths would be more significant than for two texts written by a single author, even if the topics are different (see, e.g., [36]). As mentioned earlier, Mendenhall [20] found empirical evidence of the above claim based on an analysis of word-length distribution patterns in works by Shakespeare, Bacon, and Marlowe. Other researchers have
questioned this claim [21, 39, 40]. The authors of [41]: 12 argue that "word length need not, or not only, or perhaps not even primarily, be characteristic of an individual author's style, rather word length, and word length frequencies maybe dependent on a number of other factors, genre being one of them" (see also [39, 42]).

In this study, the most frequently discussed boundary conditions such as language, genre, time of the composition of the texts, and size of the text samples were controlled. Therefore, we can assume that significant differences in word length were largely due to a different authorship. In Section 3.2, we examined both the mean word length and the frequency distribution of words with different lengths in three parts of the DRC. In Section 3.2.1, we showed significant differences between Parts II and III in terms of static mean word length (based on the frequency of types), but no significant differences were found between Parts I and II, and Parts I and III. If the above assumption is on the right track, the statistics leads to one conclusion, namely, that Parts I and II, as well as Parts I and III, were written by a single author, while Parts II and III were written by different authors, which is paradoxical. This, in turn, shows that static mean word length is highly not characteristic of an individual author's style, which is consistent with the study by Wei et al. [41]. As for the dynamic mean word length (based on the frequency of tokens), we found significant differences between Parts I and II, and Parts II and III, but there are no significant differences between Parts I and III. This leads to the unlikely conclusion that Parts I and III were written by the same author, and Part II by a different author, since it is generally accepted that Parts I and II were written by Cao Xueqin. Based on the results of our calculations, we can conclude that the average word length is not an indication of authorship.

However, regarding word-length distribution patterns, the three parts present the same regularities. The best-fitting model to describe them is extended logarithmic, and no statistically significant differences have been found. If word-length distribution is an indication of authorship, as we assumed, we can conclude that DRC was written by one and the same author.

In addition, to strengthen the above claim, we did an additional experiment to examine the word-length distributions of works that were chronically similar and shared the same style but composed by different authors. We randomly selected five sample texts from LCMC, a modern written Chinese corpus. The genre of the sample texts is novel, and the number of each text is roughly 2000. Besides, they were composed contemporaneously by different authors. We calculated the word lengths of each text and examined the distribution model using the Altmann-Fitter. It was found that four out of five texts had appropriate distribution models and there were discrepancies between them, as shown in Table 7.

This experiment further displays that when the emergence time and the style of texts are adequately controlled, the difference of word-length distribution is very likely an indication of a difference of authorship, which supports our hypothesis that the same word-length distribution may be attributable to identical authorship.

5. Conclusion

The statistical characterization of the word-length distribution in DRC was mainly carried out from two perspectives: an exploration of fitting models based on 120 individual sample texts of 2000 words from 120 chapters and a comparison of both the mean word length and the word-length distribution patterns in three groups of
sample texts, namely, Part I (1–40), Part II (41–80), and Part III (81–120).

Extended logarithmic (θ, α) was found to be the most adequate theoretical distribution model fitting word-length distribution in DRC, both in individual sample texts and in groups of texts. The results show that a syllable or character can be accepted as the unit of measurement for written Chinese.

Moreover, significant differences were found between different parts of DRC in terms of mean word length, but paradoxical or implausible conclusions could arise if the authorship attribution is judged based on mean word length differences. This, in turn, proves that mean word length is not an indication of authorship.

In addition, no significant differences in word-length distribution were found between the different parts of DRC, according to both model fitting results and the variance analysis tests. It suggests that DRC might be written by one single author.

**Data Availability**

The data used to support the findings of this study are available from the corresponding author upon reasonable request.

**Conflicts of Interest**

The authors declare that they have no conflicts of interest.

**Acknowledgments**

The research was funded by the 2021 BJTU Fund for Teaching Reform and the Fundamental Research Funds for the Central Universities (Grant no. FRF-BR-20-06B).

**References**