Hindawi
Complexity
Volume 2022, Article ID 9077360, 8 pages
https://doi.org/10.1155/2022/9077360

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 Feng**[3]

[1]*School of Foreign Studies, University of Science and Technology Beijing, Beijing 100083, China*
[2]*School of Languages and Communication Studies, Beijing Jiaotong University, Beijing 100044, China*
[3]*Software 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

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 correlativity 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).

Table 1: Fitting of extended logarithmic to word-length distribution in six sample texts.

$X[i]$	Text 20		Text 40		Text 60	
	$F[i]$	NP$[i]$	F[i]	NP$[i]$	$F[i]$	NP$[i]$
1	1104	1104.00	1084	1084.00	1121	1121.00
2	771	775.60	791	797.05	750	755.69
3	109	99.33	117	98.82	113	100.81
4	16	21.07	8	20.13	16	22.51
	$\theta = 0.2561$, $\alpha = 0.4480$, $X^2 = 2.1888$, $P(X^2) = 0.1390$ DF = 1, $C = 0.0011$, $R^2 = 0.9999$		$\theta = 0.2480$, $\alpha = 0.4580$, $X^2 = 10.7041$, $P(X^2) = 0.0011$ DF = 1, $C = 0.0054$, $R^2 = 0.9994$		$\theta = 0.2668$, $\alpha = 0.4395$, $X^2 = 3.3976$, $P(X^2) = 0.0653$ DF = 1, $C = 0.0017$, $R^2 = 0.9998$	
$X[i]$	Text 80		Text 100		Text 120	
	$F[i]$	NP$[i]$	$F[i]$	NP$[i]$	$F[i]$	NP$[i]$
1	1027	1027.00	1022	1022.00	962	962.00
2	863	855.51	874	869.27	913	912.72
3	87	98.97	85	92.92	105	105.54
4	22	15.27	19	15.81	20	19.74
5	0	2.65				
6	1	0.61				
	$\theta = 0.2314$, $\alpha = 0.4865$, $X^2 = 6.0484$, $P(X^2) = 0.0486$ DF = 2, $C = 0.0030$, $R^2 = 0.9998$		$\theta = 0.2138$, $\alpha = 0.4890$, $X^2 = 1.3447$, $P(X^2) = 0.2462$ DF = 1, $C = 0.0007$, $R^2 = 0.9999$		$\theta = 0.2313$, $\alpha = 0.5190$, $X^2 = 0.0062$, $P(X^2) = 0.9371$ DF = 1, $C = 0.0000$, $R^2 = 1.0000$	

Note. X[i] is the observed classes of word length; F[i] is the observed frequency; NP[i] is the calculated frequency of the extended logarithmic (θ, α) distribution model.

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 $P(x^2)$ or the discrepancy coefficient C ($C = (x^2/N)$). If $P \geq 0.05$, or in the case of long texts, $C \leq 0.02$, 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

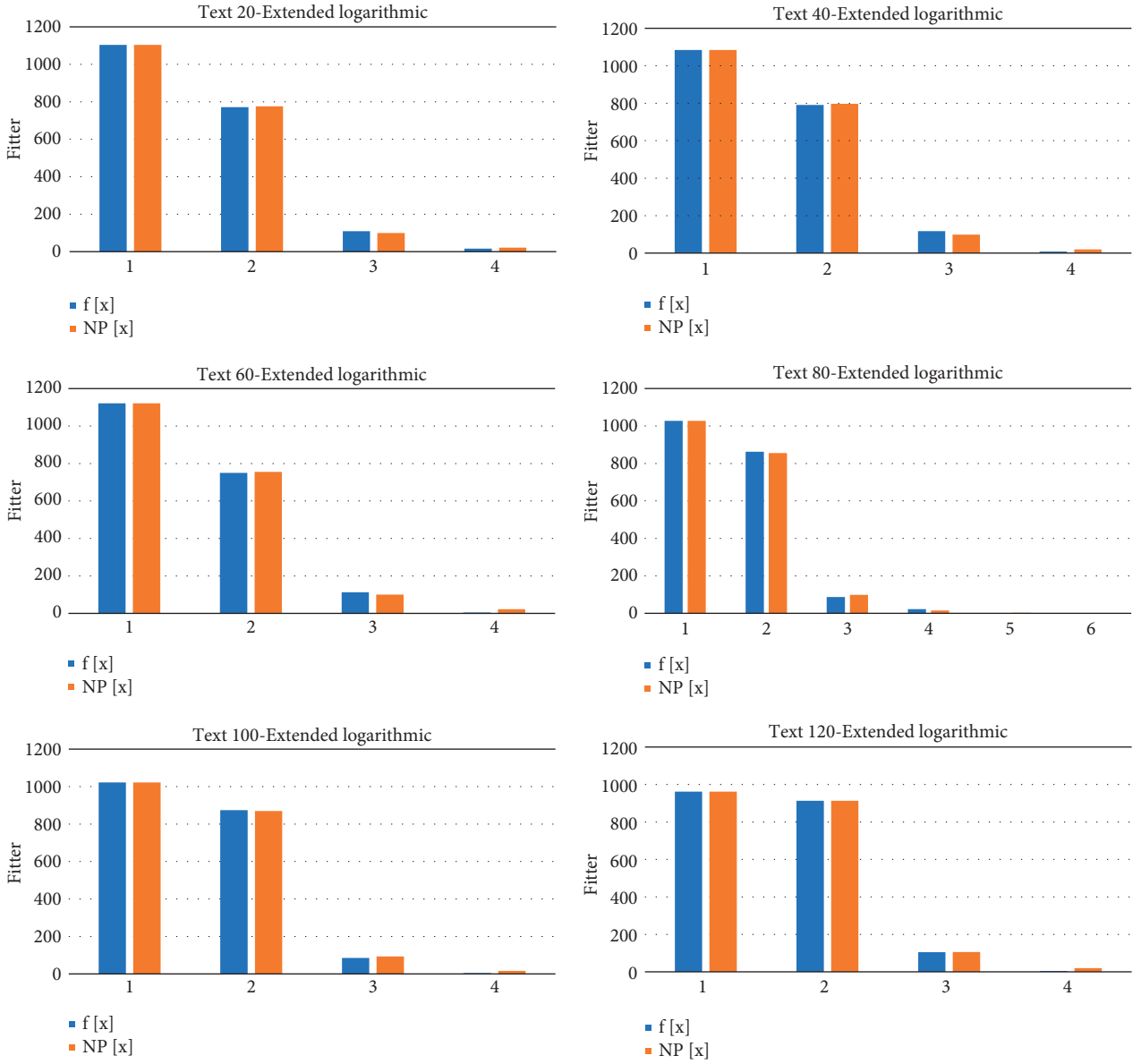

Figure 1: Fitting of extended logarithmic to word-length distribution in six sample texts.

$P(x^2)$, the best result was provided by the extended logarithmic (θ, α), 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:

$$\mathrm{DMWL} = \frac{\sum_{i=1}^{n} X_i F_i}{\sum_{i=1}^{n} F_i}. \quad (1)$$

Static mean word length:

Table 2: Dynamic and static mean word length of the three parts of DRC.

	Part I (texts 1–40)	Part II (texts 41–80)	Part III (texts 81–120)
DMWL	1.5805	1.5574	1.5828
SMWL	1.8584	1.8435	1.8649

Table 3: Using the Altmann-Fitter to fit the static word-length distribution data of the three parts of DRC.

DM	Part I				Part II				Part III			
	X^2	C	DF	R^2	X^2	C	DF	R^2	X^2	C	DF	R^2
A	283.62	0.0076	3	0.9983	226.77	0.0064	3	0.9986	273.85	0.0080	3	0.9983
B	128.70	0.0035	2	0.9997	78.32	0.0022	2	0.9998	136.41	0.0040	2	0.9997
C	119.37	0.0032	3	0.9999	70.65	0.0020	3	0.9999	128.02	0.0038	3	0.9999
D	132.13	0.0036	2	0.9996	82.18	0.0023	2	0.9997	139.57	0.0041	2	0.9996
E	278.09	0.0075	3	0.9984	216.25	0.0061	3	0.9987	271.35	0.0080	3	0.9984
F	153.90	0.0042	2	0.9999	122.12	0.0035	2	0.9998	120.13	0.0035	2	0.9999
G	164.54	0.0044	2	0.9991	95.62	0.0027	2	0.9995	162.67	0.0048	2	0.9992
H	130.28	0.0035	1	0.9998	78.72	0.0022	1	0.9999	138.47	0.0041	1	0.9998

Note. DM refers to the distribution model; *A* stands for positive Cohen–-Poisson (α, a); *B* stands for positive Cohen-negative binomial (k, p, a); *C* stands for extended logarithmic (θ, α); *D* stands for extended positive-negative binomial (k, p, a fixed); *E* stands for Shenton–Skees geometric (p, a); *F* stands for Dacey-negative binomial (k, p, a); *G* stands for Shenton–Skees logarithmic (a, b, θ); *H* stands for right truncated modified Zipf–Alekseev (a, b; $n = x$-max, α fixed).

$$\text{SMWL} = \frac{\sum_{i=1}^{n} X_i F_i'}{\sum_{i=1}^{n} F_i'}. \quad (2)$$

Here, i refers to the word length class, n refers to the number of word length classes; X_i is the length of class i, F_i is the number of tokens of class i; and F_i' is the number of types of class i.

Table 2 shows that both the dynamic and static mean word lengths of the last part are slightly longer than those of the other two parts. Moreover, the T-tests show that as far as the dynamic mean word length is concerned, there are significant differences between Parts I and II, and between Parts II and III ($P = 0.034 < 0.05$ and $P = 0.02 < 0.05$, respectively). There are no significant differences between Parts I and III. However, in terms of static mean word length, there are significant differences between Parts II and III ($P < 0.01$) and no significant differences were found between Parts I and II ($P = 0.119 > 0.05$) or Parts I and III ($P = 0.164 > 0.05$).

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.

Considering the values of C and R^2 and the number of parameters, extended logarithmic (θ, α) 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.

Table 4: Fitting extended logarithmic to static word-length distribution data of the three parts of DRC.

$X[i]$	Part I		Part II		Part III	
	$F[i]$	NP[i]	$F[i]$	NP[i]	$F[i]$	NP[i]
1	9966	9966.00	9810	9810.00	8930	8930.00
2	23236	23180.27	21968	21939.94	21614	21553.89
3	3096	3193.25	2865	2928.70	2868	2961.03
4	737	586.53	633	521.26	686	542.38
5	43	121.20	51	104.37	32	111.77
6	4	34.76	6	28.73	1	31.94

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 (θ,α) 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

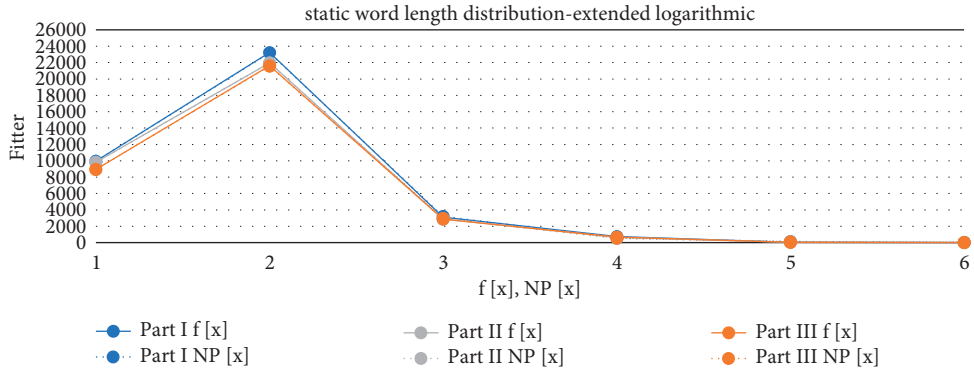

Figure 2: Fitting extended logarithmic to static word-length distribution data of the three parts of DRC.

Table 5: Using the Altmann-Fitter to fit the dynamic word-length distribution data of the three parts of DRC.

DM	Part I				Part II				Part III			
	X^2	C	DF	R^2	X^2	C	DF	R^2	X^2	C	DF	R^2
A	159.13	0.0020	2	0.9996	155.35	0.0019	2	0.9996	132.27	0.0017	2	0.9997
B	389.53	0.0049	3	0.9992	361.69	0.0045	3	0.9993	341.08	0.0043	3	0.9994
C	274.19	0.0034	3	0.9997	231.26	0.0029	3	0.9998	225.45	0.0028	3	0.9998
D	68.92	0.0009	2	1.0000	35.47	0.0004	2	1.0000	67.05	0.0008	2	1.0000
E	96.35	0.0012	3	1.0000	57.92	0.0007	3	1.0000	100.00	0.0012	3	1.0000
F	508.79	0.0064	1	0.9990	493.78	0.0062	1	0.9991	430.86	0.0054	1	0.9992
G	66.87	0.0008	2	1.0000	33.03	0.0004	2	1.0000	63.73	0.0008	2	1.0000
H	415.33	0.0052	3	0.9992	376.08	0.0047	2	0.9994	361.54	0.0045	2	0.9994
I	148.31	0.0019	3	0.9998	114.11	0.0014	3	0.9998	132.62	0.0017	3	0.9998
J	384.14	0.0048	2	0.9992	351.96	0.0044	2	0.9993	336.50	0.0042	2	0.9994
K	70.14	0.0009	2	1.0000	35.51	0.0004	2	1.0000	70.38	0.0009	2	1.0000
L	81.60	0.0010	1	1.0000	44.19	0.0006	1	1.0000	83.76	0.0010	1	1.0000

Note. A: hyper-Pascal (k, m, q); B: hyper-Poisson (a, b); C: positive Cohen–Poisson (a, α); D: positive Cohen-negative binomial (k, p, α); E: extended logarithmic (θ, α); F: extended positive binomial (n, p; α fixed); G: extended positive-negative binomial (k, p; α fixed); H: Dacey–Poisson (a, α); I: Shenton–Skees geometric (p, a); J: Dacey-negative binomial (k, p, α); K: Shenton–Skees logarithmic (a, b, θ); L: right truncated modified Zipf–Alekseev (a, b; $n = x$-max,α fixed).

Table 6: Fitting extended logarithmic to dynamic word-length distribution data of the three parts of DRC.

$X[i]$	Part I		Part II		Part III	
	$F[i]$	NP[i]	$F[i]$	NP[i]	$F[i]$	NP[i]
1	39753	39753.00	41118	41118.00	39330	39330.00
2	34933	35004.78	33969	34046.36	35508	35583.91
3	4485	4253.30	4182	4050.98	4408	4256.32
4	781	721.52	674	642.67	720	678.82
5	44	134.57	51	114.70	33	121.79
6	4	33.83	6	27.28	1	29.16

word-length distribution in DRC, measured in terms of syllable or character, would be extended logarithmic (θ, α). 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

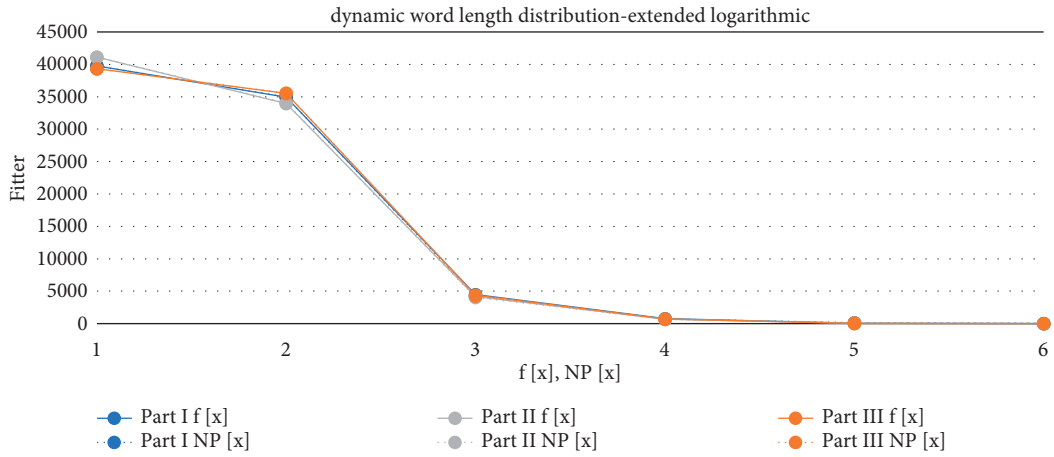

Figure 3: Fitting extended logarithmic to dynamic word-length distribution data of the three parts of DRC.

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.

Table 7: Word-length distribution of the sample texts from LCMC.

Sample text	Distribution model	$P(X^2)$	R^2
LCMC-K02	Dacey-negative binomial	0.0284	0.9999
LCMC-K04	Extended positive Poisson	0.1942	1.0000
LCMC-K06	Poisson	0.0785	0.9990
	Hyper-Poisson	0.0854	0.9999
	Jackson–Nickols	0.6306	1.0000
	Extended logarithmic	0.7486	1.0000
	Dacey–Poisson	0.0731	0.9998
	Shenton–Skees geometric	0.2919	1.0000
LCMC-K10	Extended positive Poisson	0.4460	1.0000

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

[1] S. Hu, *Textual Research on the Dream of the Red Chamber*, Beijing Publishing Group, Beijing, China, 1921, in Chinese.

[2] C. R. Zhou, *New Textual Criticism of Dream of the Red Chamber*, Tang Di Publishers, Shanghai, China, 1953, in Chinese.

[3] D. J. Liu, *The Ideology and Characters of Dream of the Red Chamber*, Classical Literature Publishers, Shanghai, China, 1956, in Chinese.

[4] S. C. Wu, *Exploration of the Origins of Dream of the Red Chamber*, People's Publishers, Beijing, China, 1961, in Chinese.

[5] X. H. He, *An Appreciation Dictionary of Poems in Dream of the Red Chamber*, The Forbidden City Publishing House, Beijing, China, 1990, in Chinese.

[6] K. L. Wang, *A Discussion of the Characters of Dream of the Red Chamber*, Beijing Publishing Group, Beijing, China, 2004, in Chinese.

[7] G. F. Lin, *Hongloumeng Zongheng Tan*, Culture and Art Publishing House, Beijing, China, 2005, in Chinese.

[8] A. C. Yu, *Rereading the Stone: Desire and the Marking of Fiction in Dream of the Red Chamber*, Princeton University Press, Princeton, 2018.

[9] X. X. Shen, *Metaphor: A New Horizon on the Language of Dream of the Red Chamber*, Fudan University Press, Shanghai, China, 2019, in Chinese.

[10] B. Karlgren, "New excursions in Chinese grammar," *Bulletin of the Museum of Far Eastern Antiquities*, vol. 24, pp. 51–80, 1952.

[11] B. C. Chan, "The authorship of the Dream of the Red Chamber based on a computerized statistical study of its vocabulary," in *Proceedings of the Paper presented at 1st International Forum on Dream of the Red Chamber*, Madison, WI, January 1981.

[12] G. Q. Li and R. F. Li, "Study based on statistics of word frequency-research on only author of the Dream of the Red Chamber," *Journal of Shenyang Institute of Chemical Technology*, vol. 20, no. 4, pp. 305–307, 2006.

[13] X. P. Li, "A new study on Hongloumeng," *Fudan Journal (Social Sciences Edition)*, vol. 5, pp. 3–16, 1987, in Chinese.

[14] W. H. Wang, "Lexical differences between the first 80 chapters and the later 40 chapters in A Dream of Red Mansions," *Research in Ancient Chinese Language*, vol. 33, no. 3, pp. 35–40, 2010, in Chinese.

[15] J. J. Shi, "Research on the author of A Dream of Red Mansions based on support vector machine," *Studies on A Dream of Red Mansions*, vol. 33, no. 5, pp. 35–52, 2011, in Chinese.

[16] T. J. Xiao and Y. Liu, "Analysis of words and N-gram models in A Dream of Red Mansions," *Modern Library and Information Technology*, vol. 31, no. 4, pp. 50–57, 2015, in Chinese.

[17] L. Ye, "Analysis of authors of A Dream of Red Mansions based on quantitative stylistic features clustering," *Studies on A Dream of Red Mansions*, vol. 38, no. 5, pp. 312–324, 2016, in Chinese.

[18] G. Q. Qin and C. G. Gu, "Author identification of A Dream of Red Mansions based on sentence classification model," *Software Guide*, vol. 20, no. 4, pp. 26–31, 2021, in Chinese.

[19] H. Zhu, L. Lei, and H. Craig, "Prose, verse and authorship in Dream of the Red Chamber: A stylometric analysis," *Journal of Quantitative Linguistics*, vol. 28, no. 4, pp. 289–305, 2021.

[20] T. C. Mendenhall, "A mechanical solution to a literary problem," *Popular Science Monthly*, vol. 60, pp. 97–105, 1901.

[21] C. B. Williams, "Mendenhall's studies of word-length distribution in the works of Shakespeare and Bacon," *Biometrika*, vol. 62, no. 1, pp. 207–212, 1975.

[22] G. Wimmer, R. Köhler, R. Grotjahn, and G. Altmann, "Towards a theory of word length distribution," *Journal of Quantitative Linguistics*, vol. 1, no. 1, pp. 98–106, 1994.

[23] P. Meyer, "Relating word length to morphemic structure: A morphologically motivated class of discrete probability distributions," *Journal of Quantitative Linguistics*, vol. 6, no. 1, pp. 66–69, 1999.

[24] S. Barbaro, "Word length distribution in Italian letters by pier paolo pasolini," *Journal of Quantitative Linguistics*, vol. 7, no. 2, pp. 115–120, 2000.

[25] H. Pande and H. S. Dhami, "Model generation for word length frequencies in texts with the application of Zipf's order approach," *Journal of Quantitative Linguistics*, vol. 19, no. 4, pp. 249–261, 2012.

[26] G. Altmann, "Aspects of word length," in *Issues In Quantitative Linguistics 3*, R. Köhler and G. Altmann, Eds., pp. 23–38, RAM-Verlag, Lüdenscheid, Germany, 2013.

[27] I.-I. Popescu, S. Naumann, E. Kelih et al., "Word length: aspects and languages," in *Issues In Quantitative Linguistics 3*, R. Köhler and G. Altmann, Eds., pp. 224–281, RAM-Verlag, Lüdenscheid, Germany, 2013.

[28] H. Chen, J. Y. Liang, and H. T. Liu, "How does word length evolve in written Chinese?" *PLoS ONE*, vol. 10, no. 9, p. e0138567, 2015.
[29] H. Chen and H. Liu, "How to measure word length in spoken and written Chinese," *Journal of Quantitative Linguistics*, vol. 23, no. 1, pp. 5–29, 2016.
[30] L. Wang, "Word length in Chinese," in *Issues In Quantitative Linguistics 3*, R. Köhler and G. Altmann, Eds., pp. 39–53, RAM-Verlag, Lüdenscheid, Germany, 2013.
[31] H. Chen, *Quantitative Studies of Chinese Word Length*, Ph.D. Dissertation, Zhejiang University, Hangzhou, China, 2016.
[32] P. Meyer, "Word–Length distribution in Inuktitut narratives: Empirical and theoretical findings," *Journal of Quantitative Linguistics*, vol. 4, no. 1–3, pp. 143–155, 1997.
[33] H. C. Tu and J. Hsiang, "A text-mining approach to the authorship attribution problem of Dream of the Red Chamber," in *Proceedings of the Paper presented at 2013 Digital Humanities Conference*, Lincoln, NE, June 2013.
[34] K. Du, "Authorship of dream of the Red Chamber: A topic modeling approach," in *Proceedings of the Paper presented at 2017 Digital Humanities Conference*, Montreal, Canada, August 2017.
[35] Y. C. Deng and Z. W. Feng, "A quantitative linguistic study on the relationship between word length and word frequency," *Journal of Foreign Languages*, vol. 36, no. 3, pp. 29–39, 2013, in Chinese.
[36] H. T. Liu, *An Introduction to Quantitative Linguistics*, The Commercial Press, Beijing, China, 2017, in Chinese.
[37] W. Röttger, "Distribution of word length in Ciceronian letters," *Journal of Quantitative Linguistics*, vol. 3, no. 1, pp. 68–72, 1996.
[38] H. T. Liu, *Advances in Quantitative Linguistics*, Zhejiang University Press, Hangzhou, China, 2018, in Chinese.
[39] P. Grzybek, E. Stadlober, E. Kelih, and G. Antić, "Quantitative text typology: The impact of word length," in *Classification-The Ubiquitous Challenge*, C. Weihs, Ed., pp. 53–64, Springer, Heidelberg, Berlin, 2005.
[40] F. Lian and Y. Li, "Word length distribution in German Texts during the 17th-19th century," *Journal of Quantitative Linguistics*, vol. 28, no. 2, pp. 117–137, 2021.
[41] A. Y. Wei, Q. Lu, and H. T. Liu, "Word length distribution in Zhuang Language," *Journal of Quantitative Linguistics*, vol. 28, no. 3, pp. 195–222, 2021.
[42] E. Kelih, G. Antić, P. Grzybek, and E. Stadlober, "Classification of author and/or genre? The impact of word length," in *Classification -The Ubiquitous Challenge*, C. Weihs and W. Gaul, Eds., pp. 498–505, Springer, Heidelberg, Berlin, 2006.