Study on the Differences in the Language Styles of Dream of Red Mansions Based on the Statistics of Lexical and Syntactic Features

Cuiting Hu and Yanqiu Shao

ABSTRACT

The author problem of Dream of Red Mansions has always been an important research topic of the academia. It has been basically a conclusion at present, that’s, the author of the first eighty of Dream of Red Mansions and the author of the latter forty are not the same person. In this paper, we want to use the computer science method and statistics to analyze the lexical features and syntactic structure of the first eighty and the latter forty of Dream of Red Mansions so as to explore the difference of language style, and the lexical features includes poems and idioms. From the analysis results, we find there are indeed differences in the lexical features and syntactic structures of the first eighty and the latter forty of Dream of Red Mansions, which not only corroborates the basic conclusion of the academia but also opens up a new perspective to research Dream of Red Mansions from the computer science.

INTRODUCTION

Measurement style is an important subject, and statistical analysis of literary works is an import application of statistical linguistics. Professor of Stanford University America called Efron and her students have made a study of Shakespeare’s works with statistical analysis method in 1976 and 1987[1][2]. At present, Generally speaking, previous researches on the language style of Dream of the Red Mansions mainly focus on lexical features to observe the characteristics of the first eighty chapters and the latter forty chapters of Dream of Red Mansions based on word frequency statistics and vector analysis. As early as 1952, the

1Cuiting Hu, Yanqiu Shao, Beijing Language and Culture University, Beijing, China.
Swedish scholar B. Karlgren[3] analyzed the using of 32 grammatical words and the using habit of oral vocabulary by statistical methods. In the 1980s, Chinese scholars Chen Bingzao[4] professor used statistical methods to conclude the first eighty and the latter forty were all written by Cao Xueqin, too. Li Xianping[5] combined the knowledge of computer and statistics, explored the appearance frequency of 47 function words and the writing styles. And Liu Ying and Xiao Tianjiu[6], Li Ruifang[7] all had done some studies, and found the writing habit difference between the first eighty and the latter forty in terms of the using of words and phrases. Zhang Weidong[8] and Chen Dakang[9] also got the same conclusion. Shi Jianjun[10] made use of the support vector machine technology in computer artificial intelligence to classify 120 chapters of Dream of Red Mansions with 44 classical Chinese characters as the eigenvector to prove the red scholar's inference over the years. While Wei Bocheng[11] applied the theory and method of "two overall equivalence tests" in statistics to prove the significant difference in some important scene descriptions of Dream of Red Mansions.

In addition, from the perspective of text clustering, text clustering is carried out on 120 chapters of Dream of Red Mansions. In this way, Xiao Tianjiu, Liu Ying[12] and Shi Jianjun[13] had done related work, and they argued that the first eighty and the latter forty of Dream of Red Mansions may not be made by one person.

However, most of these researches are done from the perspective of vocabulary, such as word length, unique words, high frequency words and so on, and few studies have been made from the viewpoint of syntactic structures. While the using of syntactic structures can reflect the writing habit of authors to a certain extent. So this paper will not only use the idioms and poems to study the differences of the first eighty chapters and the last forty chapters, but also use syntactic structures and use the relevant statistical methods and hypothesis testing to explore authorship problem of the first eighty and the latter forty of Dream of Red Mansions.

**Statistics and Analysis of Idioms in Dream of the Red Mansions**

Throughout the research on the style calculation of Dream of Red Mansions, there are many researches from the perspective of vocabulary, such as real words, function words, word categories, etc. However, many researches do not involve idioms. As a monumental work, "Dream of Red Mansions", the vocabulary, the stories and allusions contained in it are undoubtedly the most outstanding place in the whole work, while idioms are an indispensable part of this. Therefore, it is significant and valuable to conduct statistics and analysis of the idioms in Dream of Red Mansions for the study of its language styles.
**DATA STATISTICS OF IDIOMS**

We divide each ten chapters into a part as a research unit, and get 1217 idioms in Dream of the Red Mansions, 753 in the first eighty and 464 in the latter forty. As is shown in Figure 1.

![Figure 1. Idioms frequency distribution of dream of red mansions.](image)

In Figure 1, according to statistics, we can see the number of idioms from the first 40 chapters (from 1 to 40) of Dream of Red Mansions is 348, the number of idioms from the middle 40 chapters (from 41 to 80) in Dream of Red Mansions is 405, and the number of idioms from the last 40 chapters (from 81 to 120) is 464, which is the largest one. Just from the statistical results, we can’t simply draw conclusions, so we need the further testing.

**DIFFERENCE TEST OF IDIOMS**

Put 1-80 chapters as A, 81-120 as B, then A and B are two independent overall. We want to explore whether A and B have significant differences in the number of idioms. First of all, using the Shapiro-Wilk normal test method, we conducted a normality test on the two sets of data to get $p_A = 0.5342 > 0.05$ and $p_B = 0.02386 < 0.05$. Therefore, A is generally normal distribution, but B is not. Based on the above, we use rank sum test of non-parametric test to verify the previous guess. This method is mainly used to compare the differences between two independent samples.

First the null hypothesis $H_0$: there is no difference in the use of idioms between the first 80 chapters and the latter 40 chapters, that is, $\theta_A = \theta_B$. Hypothesis $H_1$: the first 80 chapters and the last 40 chapters are different in the use of idioms, that is, $\theta_A \neq \theta_B$.

In order to complete the test, we need to sort these 12 groups of data from small to large, as is shown in the following table I: (For rank equal, average rank is used.)
**Statistics and Analysis of Poems in Dream of the Red Mansions**

Poetry culture is one of the most brilliant parts of Dream of the Red Mansions. The rich poetry culture in Dream of Red Mansions can demonstrate the author's talent and creativity. Therefore, the statistics and analysis of these poems can not only observe the change of the using number of poems, but also can see the author's change of language styles.

**DATA STATISTICS OF POEMS**

We count the poems of Dream of Red Mansions, the following figure below is our data:

In Figure 2, by observing the graph above, we can intuitively see that the using number of poems in the last forty chapters is far less than the first eighty chapters.

---

**TABLE 1. THE FREQUENCY SEQUENCE OF EVERY TEN CHAPTERS.**

<table>
<thead>
<tr>
<th>Data</th>
<th>Group mark</th>
<th>Rank</th>
<th>Data</th>
<th>Group mark</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>74</td>
<td>A</td>
<td>1</td>
<td>105</td>
<td>B</td>
<td>7</td>
</tr>
<tr>
<td>80</td>
<td>A</td>
<td>2</td>
<td>107</td>
<td>B</td>
<td>8</td>
</tr>
<tr>
<td>85</td>
<td>A</td>
<td>3</td>
<td>114</td>
<td>A</td>
<td>9.5</td>
</tr>
<tr>
<td>90</td>
<td>A</td>
<td>4</td>
<td>114</td>
<td>A</td>
<td>9.5</td>
</tr>
<tr>
<td>92</td>
<td>A</td>
<td>5</td>
<td>115</td>
<td>B</td>
<td>11</td>
</tr>
<tr>
<td>104</td>
<td>A</td>
<td>6</td>
<td>137</td>
<td>B</td>
<td>12</td>
</tr>
</tbody>
</table>

Test statistics: \( W_A = 1+2+3+4+5+6+9.5+9.5 = 40 \); \( W_B = 7+8+11+12 = 38 \); take significant level \( \alpha = 0.05 \), rejection region \( W = \{ W \leq W_{0.025}^A(m, n) \text{ or } W \geq W_{1-\alpha}^A(m, n) \} \), look up the table to get \( W_{0.025}^A(8, 4) = 14 \), \( W_{1-\alpha}^A(8, 4) = W_{0.075}(8, 4) = 38 \), that is, rejection region \( W = \{ W \leq 14 \text{ or } W \geq 38 \} \).

According to the principle of rank sum test, because the number of cases of the two sets of data is different, we should choose the smaller number of cases as the test statistic. Due to \( W_A = 38 \geq 38 \), so we should reject the null hypothesis, that means, there are some differences in the use of idioms in the first 80 chapters and the latter 40 chapters of Dream of Red Mansions. Therefore, we can infer that the author of the first eighty chapters and the author of latter forty chapters of Dream of Red Mansions are not the same person and it is consistent with the expected conclusion.

Statistics and Analysis of Poems in Dream of the Red Mansions

Poetry culture is one of the most brilliant parts of Dream of the Red Mansions. The rich poetry culture in Dream of Red Mansions can demonstrate the author's talent and creativity. Therefore, the statistics and analysis of these poems can not only observe the change of the using number of poems, but also can see the author's change of language styles.
Based on the above statistics, we can initially infer that there are some significant differences in the use of poems between the first eighty and the latter forty. Again, we do the relevant test to prove this inference.

![Poems frequency distribution](image)

**Figure 2. Poems frequency distribution of Dream of Red Mansions.**

**DIFFERENCE TEST OF POEMS**

Simultaneously, due to the smaller sample size, we use the T-test in parametric tests to prove the conclusion. The null hypothesis $H_0$: there is no difference in the use of poems between the first 80 chapters and the latter 40 chapters, that means, $u_A = u_B$. Test statistics:

$$t = \frac{\bar{X} - \bar{Y}}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

After calculation, we get $t=3.689$, when $\alpha = 0.05$, two-tail test, look up the table to find $t_{1-\alpha/2}(n_1 + n_2 - 2) = 2.2281$. Because $t=3.689 > 2.2281$, please reject the null hypothesis while support hypothesis, that means, there are some differences in the use of poems in the first 80 chapters and the latter 40 chapters of Dream of Red Mansions.

By comparing the variance, it can be seen that there are obvious differences in the two parts because the number of poems every 10 chapters fluctuates a lot. The language styles in the first 80 chapters is not as uniform as the last 40 chapters, and the using number is much higher than the last 40 chapters. However, the using number of poems by the author in the last 40 chapters is close, we can see that the author has a more unified style in the writing process and the author in the latter is
Statistics and Analysis of Syntactic Structures in Dream of the Red Mansions

Syntactic structures are not only the core component of the Chinese grammar system, but also the characteristics of Chinese grammar. Therefore, we can also explore whether there are differences in the syntactic structures between the first eighty chapters and the latter forty chapters of Dream of Red Mansions.

We use language technology platform (LTP) natural language processing system developed by Harbin Institute of Technology Social Computing and Information Retrieval Research Center to analyze and label sentence, word segmentation, part of speech tagging, named entity recognition and syntactic structures of all the texts of Dream of Red Mansions, and we reprocess and analyze the results obtained.

DATA STATISTICS OF SYNTACTIC STRUCTURES

We extract six common syntactic structures, namely ADV (adverbial-verb structure), ATT (modifier-head construction), VOB (verb-object structure), COO (coordinate structure), ABV (subject-predicate structure), CMP (verb-complement structure), and we count the using frequency of the six kinds of syntactic structures in all 120 chapters. In order to be able to get a clearer observation, we get the frequency diagram of the number of the six syntactic structures contained in each ten chapters, as is shown in Figure 3:

Observing the above data, especially the frequency diagram of the six kinds of syntactic structures in Figure 4, we can see that although the frequencies of the first eighty chapters fluctuate but relatively flat, relatively speaking, the frequencies of the latter forty chapters fluctuate greatly, and what the most important is that there is a significant change from the eighth group to the ninth group from the line chart, so
we can speculate that there is a significant difference in the use of six syntactic structures.

**DIFFERENCE TEST OF SYNTACTIC STRUCTURES**

In order to make this conclusion more scientific, we use the selected six kinds of syntactic structures to carry out one by one verification.

First take ADV (adverbial-verb structure) as an example: Since the sample size is small, we use the T test. As above, we get $t = 2.539$, take significant level $\alpha = 0.05$, and we get critical value $t_{1-\alpha/2}(n_1 + n_2 - 2) = 2.2281$, because $t = 2.539 > 2.2281$, so we should reject the null hypothesis $H_0$ while support hypothesis $H_1$, that is, there are some differences in the use of ADV (adverbial-verb structure) in the first 80 chapters and the latter 40 chapters of Dream of Red Mansions.

Similarly, we can get the same conclusion in ATT (modifier-head construction), VOB (verb-object structure), COO (coordinate structure), ABV (subject-predicate structure), CMP (verb-complement structure) of the first 80 chapters and the latter 40 chapters.

**CONCLUSIONS**

This paper explores the language styles of the first eighty and the latter forty chapters of Dream of Red Mansions from the aspect of the lexical features and syntactic structures, aiming at using the method of statistical linguistics to provide data support for the author problem of Dream of Red Mansions.

In terms of vocabulary, this paper chooses the idioms, poems and syntactic structures in Dream of the Red Mansions, respectively counting the using number and combining with the difference test to prove the inference, and we can get the conclusion that the first eighty chapters and the latter forty chapters of Dream of Red Mansions have a significant difference in the two levels, and the paper provides the evidence of lexical features and syntactic structures for the conclusion that the author of the first eighty chapters and the latter forty chapters is not the same person.

**REFERENCES**


677
9. Dakang Chen. From Mathematical Linguistics to See the Author of the Latter Forty Chapters - to Discuss with Mr. Chen Bingzao [J]. Dream of Red Mansions Tribune, 1987 (1).