Innovation
2024, Vol. 5, No. 4, pp. 109-114
https://doi.org/10.11648/j.innov.20240504.11

Research Article

The Analysis of the Characters of Wording for High School Classical Chinese

Sili Zhou*, Jie Li

Beijing Haidian District Teacher Training School Affiliated Experimental School, Beijing, China

Abstract

Classical Chinese carries the splendid culture of China. It is the important learning content for Chinese in high school. At the same time, the scoring rate is not good in the corresponding examination. Reading and comprehending classical Chinese in high school is the hardest part, especially for students who are not good at language learning. It is meaningful to analyze the characters of wording for classical Chinese. In this paper, NLP (Natural Language Processing) technology is applied to analyze the word usage for classical Chinese in high school. Firstly, ancient poems and literature in teaching materials in high school and classical Chinese exam questions of the past 15 years are collected. Secondly, raw data is preprocessed, cut words, and calculated word frequencies. Finally, top N words are generated and the experimental data is analyzed. According to the experimental results, the methods proposed in this paper can be used to analyze words for classical Chinese in high school. The Top N words are calculated and analyzed. It provides a good reference for high school students to learn classical Chinese.

Keywords

Classical Chinese in High School, Wording Analysis, Word Frequency Calculation

1. Introduction

Classical Chinese has a long history and contains the essence of traditional culture. It is the necessary learning content for high school students. At the same time, it is one of the test focuses of the college entrance examination. Reading and understanding classical Chinese are important parts of Chinese learning in high school. Classical Chinese and ancient poetry usually account for around 50 points in the college entrance examination, with a high proportion of 33.3%. Classical Chinese accounts for about 24-28 points, accounting for approximately 16.7%. Unfortunately, the study of classical Chinese is not easy. Several factors lead to the difficulty. Firstly, classical Chinese is very different from our everyday language. Secondly, the semantics of classical Chinese words are rich, and many words often have multiple meanings that are far from the words themselves. Therefore, high school students generally find it hard to learn classical Chinese.

In recent years, there has been some research on the frequency statistics and analysis of commonly used words in classical Chinese for middle school students. The paper [1] studied the teaching method of monosyllabic actual words in classical Chinese in the unified edition of high school textbooks. The paper [2] surveyed commonly used real words in classical Chinese literature for junior high school students. The paper [3] researched function word distribution and teaching methods. Literature [4-8] proposed the analysis methods of the frequency statistics of classical Chinese in

*Corresponding author: slzhoua@126.com (Sili Zhou)

Received: 26 August 2024; **Accepted:** 19 September 2024; **Published:** 10 October 2024

vocational colleges and junior high schools. In [9, 10], further research on actual words in classical Chinese was demonstrated. Literature [3, 11, 12] focused on function words in classical Chinese. Literature [13-15] studied classical Chinese exam questions for college entrance.

However, the analysis of the wording characters for the classical Chinese of high school is understudied. The topics in high school are wider than in middle school. At the same time, the content is harder. It is necessary to conduct targeted research to analyze the vocabulary of classical Chinese in high school, to help high school students improve their learning efficiency. In this paper, natural language processing technology is applied to analyze the content of high school classical Chinese. The vocabulary in classical Chinese is studied.

The structure of this article: Section 2 introduces the research method, Section 3 presents and analyzes the experimental results, and Section 4 provides the conclusion.

2. Materials and Methods

2.1. Overview Framework and Design

We collected ancient poetry and literature in textbooks and the recent 15-year classical Chinese exam questions for high school and analyzed the wording. The framework is in Figure 1.

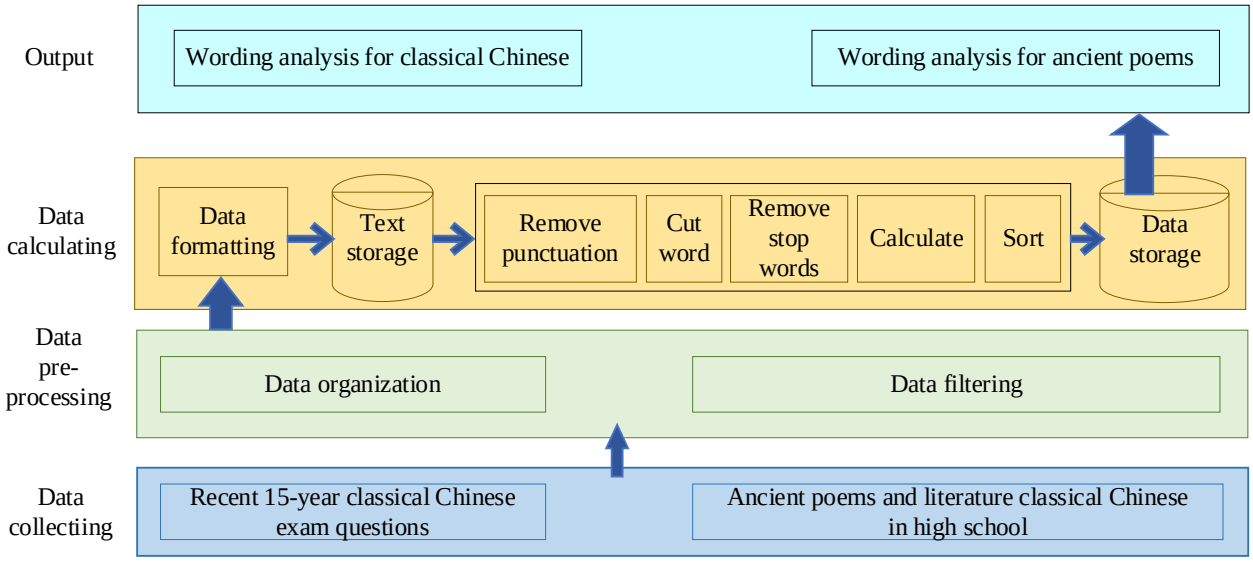

Figure 1. *Overview Framework and Design.*

2.2. Data Collection and Preprocessing

(1) Data collection

The data sources include two parts (as shown in Table 1):

Table 1. *Data sources list.*

Category	Content	Data size (word number)
Classical Chinese exam content	2008-2023 Classical Chinese content in the college entrance examination	12257
Ancient poetry and literature	72 Essential ancient poems and literature in high school textbooks	22342

(2) Data preprocessing

The data collected has different formats containing different elements. Such as images and advertisements (content unrelated to classical Chinese and poetry). It needs to be preprocessed. The preprocess includes two steps: remove invalid content and annotate raw data.

Table 2. *Annotation fields for Classical Chinese content.*

Annotation fields	Description
Year	The year of the college entrance examination questions
Region	The region of the exam paper, such as Beijing
Type	Ancient poems or classical Chinese
Dynasty	The dynasty of the text
Author	The author of the poem or article
Title	The title of the poem or article
Content	The content of the poem or article
Exam question stem	Exam questions and options
Answer	The answers to the exam questions

The annotated data is stored in CSV format. An example is Figure 2.

年份	地区	类别	题干				题目	答案
			朝代	作者	题目	内容		
2008	北京	文言文			汉书·韩延寿	延寿字长公，燕人也。霍光擢延寿为谏大夫，徙颍川。颍川多豪强，难治。先是，赵广汉为太守，患其俗多朋党，故构会吏民，令相告讦，颍川由是以为俗，民多怨仇。延寿欲更改之，教之礼让。恐百姓不从，乃历召〇1郡中长老为乡里所信向者数十人，设酒具食，亲与相对，接以礼意，问以谣俗、民所疾苦，为陈和睦亲爱销除怨咎之路。长老皆以为便，可施行，因与议定嫁娶丧祭仪品，略依古礼，不得过法，百姓遵用其教。数年，徙为东郡太守，黄霸代延寿居颍川，霸因其迹而大治。 延寿为吏，上礼义，好古教化，所至必聘其贤士，以礼待用，广谋议，纳谏争；修治学官〇2，春秋乡射，陈钟鼓管弦，盛升降揖让，及都试讲武，设斧钺旌旗，习射御之事。治城郭，收赋租，先明布告其日，以期会为大事，吏民敬畏趋向之。又置正、五长，相率以孝弟，不得舍奸人。闾里仟佰有非常，吏辄闻知，奸人莫敢入界。其始若烦，后皆便安之。接待下吏，恩施甚厚而约誓明。或欺负之者，延寿痛自刻责：“吾岂其负之，何以至此？”吏闻者自伤悔，门下掾自刭，人救不殊〇3，因瘖不能言。延寿闻之，对掾史〇4涕泣，遣吏医治视，厚复其家。 延寿尝出，临上车，骑吏一人后至，敕功曹议罚白。还至府门，门卒当车，愿有所言。延寿止车问之，卒问：“今旦明府早驾，久驻未出，骑吏父来至府门，不敢入。骑吏闻之，趋走出谒，适会明府登车。以敬父而见罚，得无亏大化乎？”延寿举手舆中曰：“微子，太守不自知过。归舍，召见门卒。卒本诸生，闻延寿贤，无因自达，故代卒，延寿遂待用之。在东郡三岁，令行禁止，断狱大减，为天下最。	下列语句中加点的词语的解释，不正确的一项是 A．或欺负之者，延寿痛自刻责 欺负：压迫、侮辱 B．略依古礼，不得过法 过法：逾越法规 C．霍光擢延寿为谏大夫 擢：提拔 D．门卒当车，愿有所言 愿：希望	A

Figure 2. *An example of annotated data.*

2.3. Wording Analysis

After pre-processing, the contents of classical Chinese and ancient poems can be extracted. But some additional characters are included. It needs further processing. The steps are as follows:

(1) Remove punctuation and special characters. The regular expression is used and generated as below:

punc = '[，。？《》：（）,.?():""“”·；\n‘ 【 】、！！‘’]'

The method of “sub” is called from the “re” module of Python. It can be used to remove punctuation and special characters. Corresponding codes are as follows:

new_txt = re.sub(punc, "", txt)

(2) The method of “cut” from the “jieba” module is used to cut words.

init_words = jieba.lcut(txt, cut_all=False)

(3) Remove stop words. According to the stop words list, the words are filtered to remove stop words. Examples of stop words are “之、【1】、①、1…”.

(4) Calculate the total number of words and each word frequency. Key points are as below:

```
for word in init_words:
    if word not in stopwords:
```

```
            if word != '\t':
                #print(word)
                words.append(word)
                # outstr += " "
    number = len(words)
```

3. Experimental Results

We designed a configurable parameter that can control the "N" number of Top N words. When applying N=30, the Top 30 frequency words are shown in Figure 3.

NO.	Top 30 of Classical Chinese content in college entrance examination questions			Top 30 of ancient poems and literature in Chinese language textbooks		
	Word	Word Frequency	Percentage(%)	Word	Word Frequency	Percentage(%)
1	也	189	3.24	也	230	2.26
2	曰	132	2.26	而	182	1.79
3	而	131	2.24	曰	82	0.81
4	为	71	1.22	不	71	0.7
5	不	48	0.82	为	69	0.68
6	与	48	0.82	与	64	0.63
7	者	48	0.82	有	63	0.62
8	矣	45	0.77	则	55	0.54
9	其	43	0.74	以	53	0.52
10	有	39	0.67	矣	51	0.5
11	天下	34	0.58	其	49	0.48
12	以	34	0.58	者	49	0.48
13	则	34	0.58	于	42	0.41
14	是	31	0.53	是	41	0.4
15	衡山	28	0.48	天下	40	0.39
16	皆	27	0.46	吾	37	0.36
17	吾	27	0.46	在	34	0.33
18	人	26	0.45	无	34	0.33
19	使	25	0.43	皆	32	0.31
20	械器	24	0.41	人	31	0.3
21	于	23	0.39	我	27	0.27
22	不能	20	0.34	欲	26	0.26
23	得	18	0.31	亦	25	0.25
24	所	18	0.31	兮	24	0.24
25	管子	18	0.31	非	23	0.23
26	礼义	16	0.27	焉	22	0.22
27	对	16	0.27	故	22	0.22
28	无	16	0.27	又	22	0.22
29	或	15	0.26	以为	21	0.21
30	诸侯	14	0.24	所	21	0.21

Figure 3. *Top 30 Words, word frequency, and percentage.*

We compared the top 10 words. It is found that 80% of them in the college entrance examination questions overlap with the top 10 words in high school textbooks (marked in green in Figure 3). It shows that the top 10 high-frequency words in the 72 ancient texts that must be memorized in high school Chinese textbooks highly overlap with the top 10 high-frequency words in college entrance examination questions. This result shows that high school students need to importance ancient poetry and literature in their textbooks. In addition, it is found that high-frequency words, due to their frequent usage, often contain multiple semantics. It needs more time to understand and master. A recommended method is to require in-depth understanding in conjunction with example sentences.

At the same time, it is observed that there are some limitations in the "jieba" module for word segmentation in classical Chinese. For example, the correctness of the word cutting for the term '械器' remains to be debated. We realize that the accuracy of word cutting significantly impacts the final frequency statistics of words. Therefore, in the next work, further research will be conducted to improve the accuracy of word cutting in classical Chinese. Another content is to generate example sentences for Top N words. This can help high school students better learn and master classical Chinese.

4. Conclusions

Classical Chinese and ancient poetry carry the long history and splendid culture of our country. They contain brilliant wisdom and humanistic ideas that have shone through the ages. As high school students, we should master classical Chinese semantics and expression norms. This can help us master classical Chinese and absorb the excellent ideas of the ancients.

This article focuses on the content of classical Chinese and poems in high school. Firstly, it collects ancient poetry and classical Chinese content from high school textbooks and the past 15 years of college entrance examination questions; Then cuts words and calculates word frequency statistics; Finally, the statistical results are analyzed and high-frequency words are provided. The experimental results show that this method can analyze words in high school classical Chinese automatically, providing a reference for high school students to improve their learning efficiency in classical Chinese.

Abbreviations

NLP Natural Language Processing

Author Contributions

Sili Zhou: Conceptualization, Resources, Methodology, Software, Validation, Writing – original draft, Writing – review & editing

Jie Li: Data curation, Investigation, Project administration, Writing – review & editing

Funding

The work is supported by the “Beijing Youth Top-notch Talent Training Program”.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Shuyi Lu. Research on the Teaching of Monosyllabic Real Words in Classical Chinese in the Unified High School Compulsory Textbook [D], Shanghai Normal University. 2023. https://doi.org/10.27312/d.cnki.gshsu.2023.002143

[2] Qian Wu. Research on Teaching of Commonly Used Real Words in Classical Chinese for Junior High School Students in the Department of Compilation and Translation [D]. Shaanxi University of Technology, 2022. https://doi.org/10.27733/d.cnki.gsxlg.2022.000239

[3] Chen Qi. Research on the Distribution and Teaching of Function Words in Classical Chinese in the Unified High School Compulsory Textbook [D]. Shanghai Normal University. 2022. https://doi.org/10.27312/d.cnki.gshsu.2022.000631

[4] Xueyang Liu. Research on Classical Chinese Vocabulary Teaching in Five Year Vocational Chinese Language Based on Word Frequency Statistics: A Case Study of the 2011 Su Education Press Chinese Language Textbook [J]. Education Science Forum. 2020, (36): 53-57.

[5] Xiaoyu Ge. Research on Teaching Common Words in Middle School Classical Chinese Based on Frequency Statistics of Commonly Used Words [D]. Liaoning Normal University. 2018.

[6] Hong Zhang. Frequency Statistics and Common Word Analysis of Classical Chinese in Middle School [D]. Huazhong Normal University, 2018.

[7] Ying Zhang. Comparative Study on Selection and Annotation of Senior Classical Chinese Textbooks in Mainland and Taiwan —Taking Mainland Edition and Taiwan Longteng Edition as examples [D]. Fujian Normal University, 2023. https://doi.org/10.27019/d.cnki.gfjsu.2023.002182

[8] Yixuan Zhang. Research on Word List Compilation and Word Teaching of Classical Chinese in Senior High School [D]. Soochow University, 2022. https://doi.org/10.27351/d.cnki.gszhu.2022.004386

[9] Tao Xue. Research on General Notional Word Teaching of High School Classical Chinese Based on Learning Task Group [D]. Sichuan University, 2022. https://doi.org/10.27703/d.cnki.gsclg.2022.000137

[10] Liang Zhong. Suggestions for Teaching High School Classical Chinese Real Words Based on Word Frequency Statistics [J]. Journal of Ningbo Institute of Education, 2013, 15(06): 130-133.

[11] Lan Chen. Study on Classical Chinese Adverb Teaching of Compulsory Textbooks in Senior High Schools [D]. Guizhou Normal University, 2024. https://doi.org/10.27048/d.cnki.ggzsu.2024.001204

[12] Songmiao Zhou. Research on the Teaching of Classical Chinese Function Words in the People's Education Press High School Chinese Textbook [D]. Liaoning Normal University, 2017.

[13] Xunyang Ren. Research on Evaluation Strategy of Classical Chinese Teaching in Senior High School Under the Background of New College Entrance Examination [D]. Shaanxi University of Technology, 2024. https://doi.org/10.27733/d.cnki.gsxlg.2024.000282

[14] Jie Wen. Focusing on Ability and Literacy to Understand Ancient Poetry and Prose - Analysis of Ancient Poetry and Prose Reading Test Questions and Preparation Suggestions for the 2024 Nine Province Joint Entrance Examination [J]. Guangxi Education, 2024, (08): 35-38.

[15] Xinrong Li. Focusing on the connection between teaching and examination, consolidating core competencies - the characteristics of classical Chinese essay topics in the 2023 National College Entrance Examination [J]. Yu Wen Tian Di, 2024, 31(02): 13-16.

Research Field

Sili Zhou: Natural language processing, Artificial intelligence, Software engineering
Jie Li: Information Technology, Education