A Statistical Analysis of Chinese Writing Style in *Xin Qingnian* 新青年 (*New Youth Magazine*) in the 1910s and 1920s

Jack C. Yue[1] Curtis Lin[2]

Abstract

Big Data has been very popular in many fields since IBM first introduced the term in 2010. With the progressive increase of available information, the ability of efficiently processing it is a necessary skill. The information can appear in various forms, such as texts and pictures, which normally are not in the digital format. Texts, for example, have no specific format and require structurization before being analysed. The process of structurization depends on the study's objective and the data's attributes. In this study, we propose a procedure for analysing Chinese texts, based on the notion of Tukey's Explanatory Data Analysis and the concept of species diversity. To evaluate the proposed approach, we use the articles from *the New Youth Magazine* published from 1915 to 1926. We found that we can trace the change of Chinese writing style through the texts of *the New Youth Magazine* using the proposed method. It seems that more characters are common in classical (or literary) Chinese, but more two-character and multi-character words are common in modern Chinese.

Keywords: Text Mining, *Xin Qingnian (New Youth Magazine)*, Writing Style, Exploratory Data Analysis, Species Diversity

[1] Professor, Department of Statistics, National Chengchi University, Taipei, Taiwan, R.O.C. email: csyue@nccu.edu.tw

2 Graduate Student, Department of Statistics, National Chengchi University, Taipei, Taiwan, R.O.C.

1. Introduction

The improvement of technology, especially at computer science, can make our life more convenient. We are getting more information and the ability of efficiently processing such an increased amount of information is a necessary skill nowadays. In fact, it is believed that the information we come across in one day exceeds the information acquired in a lifetime by a British man in the 17th century[3]. We usually use the term Big Data to describe the era of this information explosion. Big Data was first introduced in 2010 by IBM (International Business Machines) and it has four important attributes, the so-called "4V's" (volume, velocity, variety, and veracity), which means that the data size is big, the speed of processing data is quick, the data comes from a great variety of sources, and the data quality needs to be considered carefully.

It is believed that all types of data, such as texts, sound, and pictures (which are referred to as unstructured data), can be digitalized and analysed. Unstructured data accounts for 90% of all data (Praveen 2017) but it requires extra effort to explore the information hidden under this kind of data. We use text mining (analysis of text data) as an example to demonstrate the analysis of text data. Most data analysis methods can only handle structured data or digital data (e.g., 0, 1, . . . , 9) and thus the first step of text mining is to transform the texts into digital data; this process is called structurization. Unfortunately, there are no standard methods for structurization, which can vary greatly according to the objectives of the research and the nature of texts. Still, some methods are frequently used in most studies (e.g., natural language process and latent semantic analysis). For analysing English texts, the first step of text mining would also include whitespace removal, stop words removal, and stemming procedure.

However, the process of analysing Chinese text is somewhat different. Chinese characters are made of square characters (i.e., each within the same unit) and many words are combinations of two or more different word stems. On the other hand, English words are made of 26 letters, can have different lengths, and are separated by whitespaces. In Chinese, as there are no spaces between words, stemming words is particularly difficult. This implies that the pre-processing work, such as removing whitespace and stemming, are not suitable. Because Chinese characters share the same unit, we can treat Chinese words as species. Thus, we will adapt the notion of species diversity/richness and/to design analysis tools for Chinese texts.

The proposed approach is based on the idea of Exploratory Data Analysis (EDA). A famous statistician, John W. Tukey, promoted EDA in 1970's and he thought that more emphasis should be placed on using data to construct research hypotheses and the

[3] Source: New York Times, March 23, 2011 "Too Much Information About 'Information'?" https://artsbeat.blogs.nytimes.com/2011/03/23/too-much-information-about-information/

direction of data analysis should not be solely decided by the researchers[4]. The idea behind EDA are foundationally "data driven" in that the errors due to choosing inappropriate statistical models can be greatly reduced if we can have a better understanding of the data. EDA is not a mere collection of techniques. EDA is a philosophy as to how we dissect a data set; what we look for; how we look; and how we interpret. In this study, we will introduce statistical tools for analysing Chinese texts and use them to uncover the/some important data attributes. We will use these tools to explore changes in the Chinese writing style in the period of 1915-1926 based on the articles from the *New Youth Magazine* (Xin Qingnian).

At the beginning of the 20th century, two Chinese writing styles coexisted: literary Chinese (*wenyanwen* 文言文), deriving from the classical language, and vernacular Chinese (*baihuawen* 白話文), from which modern Chinese would soon derive; literary Chinese remained the standard written language until the May Fourth Movement in 1919[5]. The May Fourth Movement can be treated as the Renaissance of modern China. It was directly related to 1911 Revolution (or Chinese Revolution) and it sped up the transformation of the Chinese writing style. Although modern Chinese has become the dominant writing style since 1919, it is believed that the change was gradual and quite a few well-known Chinese novels before the 20th century were written in vernacular Chinese, such as *The Scholars* (*Rulin waishi* 儒林外史) and *The Dream of Red Chamber* (*Honglou meng* 紅樓夢).

The goal of this study is to apply the EDA tools to distinguish the differences between literary and modern Chinese, based on species diversity measures, such as the richness and distribution of vocabulary (and multiple-character phrases), function words, and punctuation. We think that words in an article is like species in an ecosystem, and that is the explanation why ecological models can be applied in analysing texts (e.g., Zipf's law). In order to achieve ecological balance, every species/word has its role and there should be connections between species/words. These attributes can be used to differentiate and compare texts/populations. In fact, often we can uncover important information through careful and complete data exploration, and this step is crucial in developing research hypotheses.

Note that we will not include the meanings of characters or phrases in the data analysis in order to avoid disputes. Also, the study material chosen is *the New Youth Magazine*, published between 1915 and 1926, which is the most important publication

---

[4] Tukey (1962) defined data analysis as: "Procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data."

[5] In Taiwan, the term "literary Chinese" is used interchangeably with "classical Chinese" (Li and Dew, 2009). Likewise, we use "modern Chinese" and "vernacular Chinese" interchangeably in this study.

witnessing the change of Chinese writing style[6]. We will introduce the types of EDA tools in the next section; the empirical analysis of texts from *the New Youth Magazine* follows.

2. Methodology

Data analysis is a kind of mathematical induction. The goal is to uncover the pattern (or rule) underlying the data, if there are enough data. Basically, there are two types of data analysis: Exploratory Data Analysis (EDA) and Confirmatory Data Analysis (CDA).[7] As mentioned previously, the EDA is often referred to as a philosophy and its role is to figure out the characteristics and essence of data, to test underlying assumptions, and to develop parsimonious models. On the other hand, the role of CDA is to examine evidence, test hypotheses, and produce estimates. We will use an example to demonstrate how the EDA and CDA function in a data analysis.

Regression analysis is one of the most popular statistical models and it can be used to estimate or predict observations (or target variables). In other words, the goal is to construct a quantitative model, give or take some errors, i.e.,

$$y = f(x) + \varepsilon, \qquad \text{or} \qquad \text{Observation} = \text{Model} + \text{Error} \qquad (1)$$

where y is the target variable, x is the independent variable(s) which can be used to describe the target variable, $f(.)$ is the model, and ε is the error term. There are two keys in model construction: choosing the most relevant x's and building the most suitable $f(.)$. We intend to decide which x's are relevant to the target variable y at the EDA stage, and to develop the model $f(.)$, plugging all relevant independent variables x's, at the CDA stage.

However, not all studies go through the EDA stage and many would focus solely on the CDA stage. We think that the EDA is as important as the CDA and it should be operated with care. According to our experience, we found that the performance of quantitative models often depends on whether we can extract important variables at the EDA stage. Inappropriate choices of variables (including irrelevant variables and variables with poor data quality) cannot provide useful information, which is often referred as "garbage in, garbage out" in data analysis. In practice, the choices of these variables require domain knowledge related to the data, and experts' opinions

[6] New Youth was a Chinese magazine which played an important role in initiating the New Culture Movement. It was founded on September 15, 1915 in Shanghai and its headquarters moved to Beijing in January 1917 (Ash, 2009).

[7] Some may say that there are three types of data analysis (El Morr and Ali-Hassen 2019): descriptive analytics (DA), predictive analytics (PA), and prescriptive analytics. DA is to tell us what has happened, PA is to predict what would happen, and prescriptive analytics is to recommend the action following the data analysis.

frequently play an important role. This is especially the case for texts and unstructured data.

In this study, our goal is to investigate the change in Chinese writing style following the May Fourth Movement. We will distinguish the writing style via the EDA tools according to three directions: a) vocabulary diversity; b) function words, and c) punctuation.

First, concerning vocabulary diversity, it is believed that fewer characters are used in vernacular Chinese. One of the main characteristics of vernacular Chinese is to use existing words to effectively reduce the reading threshold of public (Ho et al. 2014). The analysis with respect to vocabulary diversity should provide evidence to evaluate whether modern Chinese uses fewer vocabulary and phrases. Thus, we propose some measurements to explore the underlying characteristics of literary and vernacular Chinese with respect to vocabulary/phrases. These measurements are related to species diversity and richness, such as numbers of vocabulary and phrases, distributions of vocabulary and phrases, and diversity indices of vocabulary and phrases.

We can directly count the numbers of vocabulary and two-character phrases, or use measures such as Type/Token Ratio (TTR, see Manschreck et al. 1981) as the diversity index. TTR is the number of different words (called Type) divided by the number of total words (or Token). TTR tends to decrease as the number of total words increases; therefore, in order to compare texts with the same number of words, we used the standardized TTR instead. The standardized TTR is to select 10,000 (50,000, or other numbers of) words randomly from the text and calculate the number of different words per 10,000 words. We repeat this process 1,000 times and compare the median (or average) TTR from 1,000 simulation runs, and the simulation results can be used to test if the vocabulary richness of two texts are the same.

The distributions of vocabulary and two-character phrases are another way to describe diversity and the proportions of most common characters and phrases are one way to check how dominant these characters or phrases are. Another possibility is to compute species diversity indices, such as Simpson's index and entropy (or Shannon's index), defined as

$$\theta_S = \sum_i p_i^2 \qquad \text{and} \qquad \theta_E = -\sum_i p_i \ln(p_i),$$

respectively, where p_i is the proportion of vocabulary (or phrase) i. Note that the larger the entropy is, the larger the species (or vocabulary) diversity. On the contrary, a smaller value of Simpson's index indicates larger diversity (Yue et al. 2001; Yue and Clayton 2005). Gini's index is also a good choice for measuring diversity or unevenness (Forcina and Giorgi 2005). However, it produces results similar to those of entropy and

Simpson's index and for that reason the results of Gini's index were omitted from this study.

Table 1. 10 Common Function Words in *the New Youth Magazine*

	Literary Chinese	Modern Chinese
Words	矣(yi3), 乎(hu), 焉(yan), 歟(yi2), 哉(zai), 耳(er3), 豈(qi3), 之(zhi), 乃(nai3), 無(wu2)	的(di), 是(shi), 們(men), 個(ge), 了(le), 和(han4), 麼(mo), 著(zhe), 嗎(ma), 吧(ba)

Coming to the function words, we adapted the idea of Ho et al. (2014) and Yue et al. (2016), and chose 10 common function words in classical Chinese and 10 in modern Chinese, as shown in Table 1. The proportion of classical/modern Chinese function words in each volume can be used to verify if their usage changes with time. Furthermore, we propose a measurement, borrowing the idea from CUSUM (Cumulative Sum)[8] to examine whether the usages of function words have a major change. Let z_i be the number of occurrences for function word z in the i^{th} article, for $1 \le i \le n$ and n is the number of articles. Define the statistics G* as

$$G^* = \frac{\sum_{i=1}^{t} z_i}{\sum_{i=1}^{n} z_i}$$

and the G* curve is the line connecting the points

$$\left(\frac{t}{n}, \frac{\sum_{i=1}^{t} z_i}{\sum_{i=1}^{n} z_i} \right) \text{ for } 1 \le t \le n .$$

Figure 1 shows five possible patterns of G*-statistics. If the number of function words z is almost the same in all articles, then G* would look like a diagonal line (solid black line). If z_i is an increasing/decreasing function, then G* is a convex/concave curve (red lines). On the other hand, if z_i increases first and decreases later (denoted as in-decrease in Figure 1) then the G* curve looks like S-shape (black dotted line). The G* curve of the case of decreasing first and increasing later (denoted as de-increase in

[8] CUSUM involves the calculation of a cumulative sum. For example, suppose we have data $(x_1, x_2, x_3, x_4, \ldots)$ and the CUSUM is $(x_1, x_1 + x_2, x_1 + x_2 + x_3, x_1 + x_2 + x_3 + x_4, \ldots)$.

Figure 1) also looks like S-shape. We can use these five patterns to judge the usage of function words.

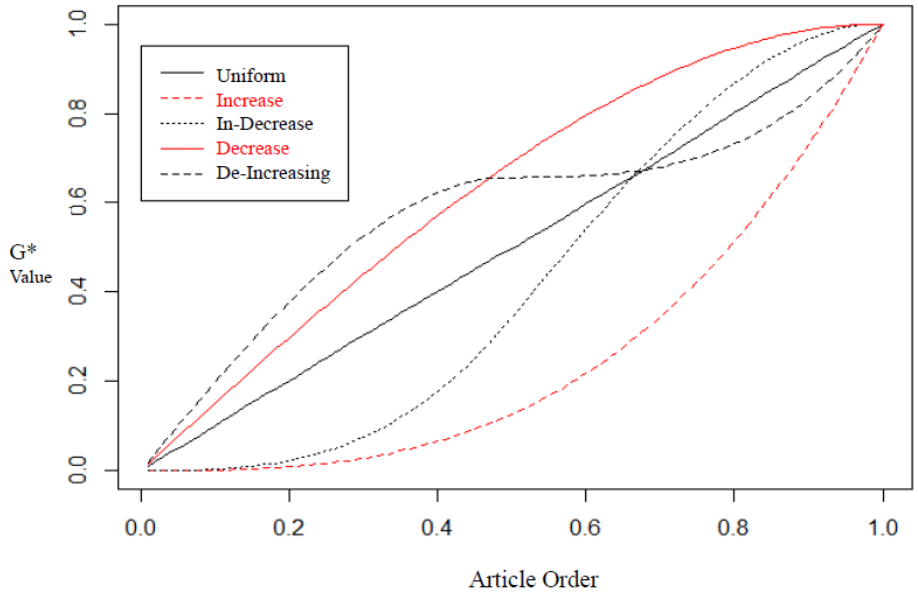

Figure 1. Five Different Patterns of G*-statistics

As for the third direction in EDA analysis, that is punctuation, although punctuation has been used in Chinese for more than 2,000 years, it was not unified until the Department of Education regulated its usage in 1920. In fact, the punctuation usually did not appear in classical Chinese writings and it was used by the readers to give their own interpretation while they read (i.e., personal comments). We first evaluated the trend of punctuation used in *the New Youth Magazine*, following by inspecting punctuation whose function is related to sentence breaking. Such punctuation includes periods, commas, semi-colons, question marks, and exclamation marks. In addition, we calculated the length of a sentence according to punctuation. It is believed that classical Chinese sentences are generally shorter.

3. EDA of *the New Youth Magazine*

In this section, we will proceed with the data analysis on *the New Youth Magazine* which has 11 volumes, published between 1915 and 1926. Many think that the writing style of Volume 1 is classical Chinese and that Volume 7 (and after) is written in modern Chinese[9].

a) vocabulary diversity:

Table 2 shows some selected vocabulary diversity, such as the numbers of words and different words, the proportion of words in the top 500 most common words, and the proportion of two-word phrases in the top 500 most common phrases. It is obvious that there are more total words but fewer different words in the later volumes, which indicates a decreasing vocabulary richness. Similarly, both the proportion of top 500 words and top 500 phrases increase with time as well, and it is especially obvious for the two-word phrases. This matches our expectation since most people think that the use of two-word and multiple-word phrases are one of the main features of modern Chinese.

Table 2. Vocabulary Diversity of the *New Youth Magazine*

Volume	# of Words	# of Vocabulary	% of Top 500 Words	% of Top 500 Phrases
1	248,833	4,379	76.60%	39.46%
2	291,848	4,344	76.66%	37.82%
3	290,038	4,227	78.92%	41.79%
4	305,020	4,298	79.59%	43.99%
5	343,519	4,125	81.06%	48.53%
6	389,407	3,848	83.08%	52.30%
7	586,942	3,850	83.34%	53.09%
8	461,731	3,753	83.75%	54.70%
9	437,748	3,745	84.06%	54.85%
10	342,778	2,980	87.50%	62.04%
11	489,223	3,093	86.67%	62.29%

[9] According to Ash (2019), the *New Youth* was the first publication to use all vernacular Chinese beginning with the May 1918 issue, Volume 4, Number 5. However, based on our analysis, there were still a lot of articles using literary Chinese in Volumes 4, 5, and 6.

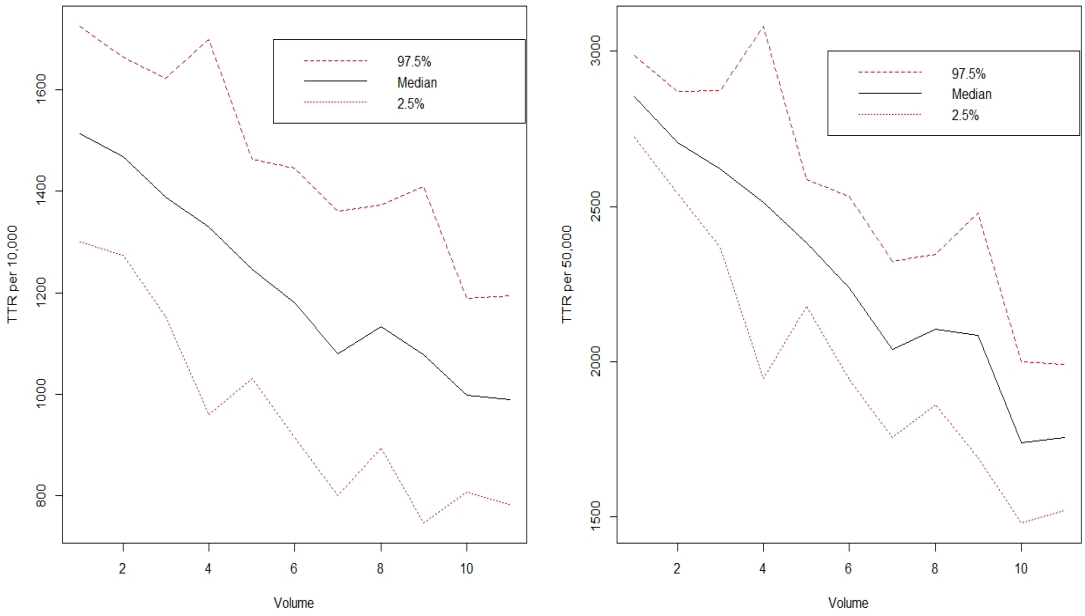

Figure 2. Standardized TTR per 10,000 and 50,000 Words

Generally speaking, the writing style of articles in Volume 1 is classical Chinese and it gradually evolves to modern Chinese by Volume 7. We can also compute the standardized TTR to check the slowly reducing trend of vocabulary diversity. As mentioned previously, 1,000 computer simulation runs are used to estimate the numbers of different words per 10,000 and 50,000 words (Figure 2). Usually we would choose the median of 1,000 runs as the estimate of TTR and treat 2.5% and 97.5% as the lower and upper bounds of 95% confidence interval. The computer simulation can also be used to obtain the (Monte Carlo) distribution of standardized TTR values for each text, for verifying whether two texts have the same vocabulary richness. It seems that the number of different words in 10,000 or 50,000 decreases steadily from Volume 1 to Volume 11, similar to those in Table 2. Entropy and Simpson's index of single words and two-word phrases also indicate a declining trend in diversity. Since the results of Simpson's index are similar, we will only provide the results of entropy. Figure 3 shows the entropy values of single words and two-word phrases and both are decreasing, which indicate a shrinking diversity in the variety of words and two-word phrases.

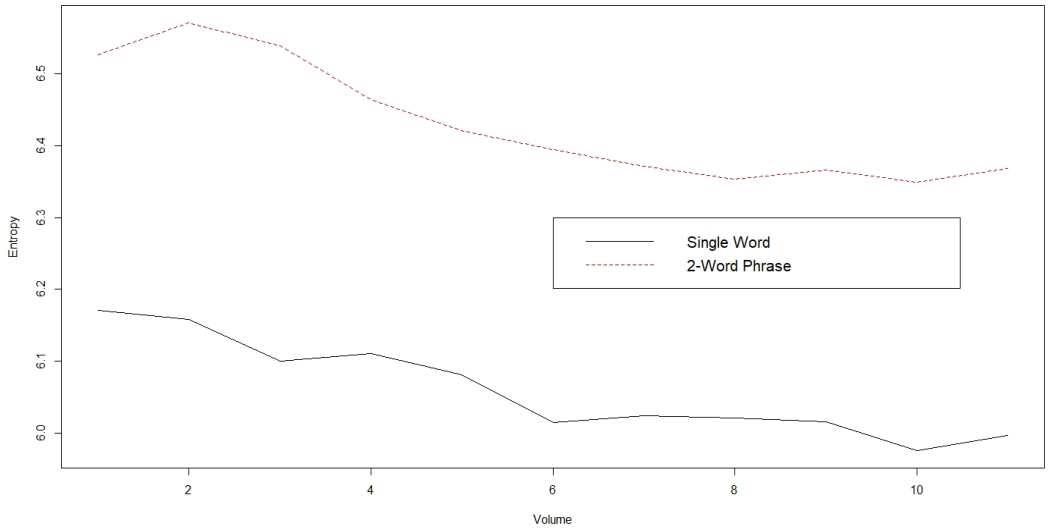

Figure 3. Entropy of Single Words and 2-word Phrases

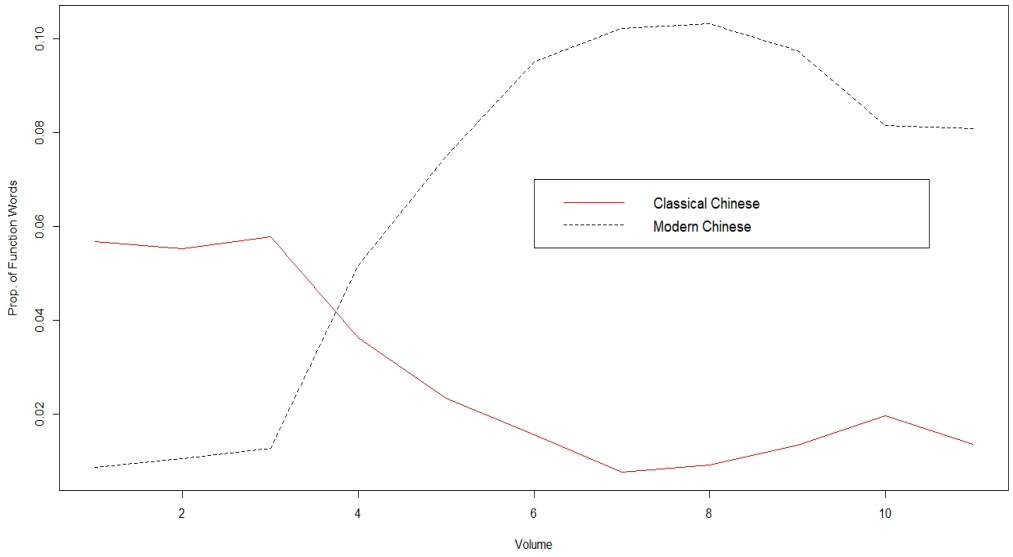

Figure 4. Proportion of Classical and Modern Chinese Function Words

b) function words

Next, we calculated the usage of classical and modern Chinese function words. Figure 4 shows the proportion of these two types of function words. It is obvious that there are more classical Chinese function words in the first three volumes and modern Chinese function words dominate since Volume 4. In fact, the use of 10 modern Chinese function words is about three times as many as that of 10 classical Chinese

function words. We can also describe the time trend of these function words via G* statistics. We choose the top two classical and modern Chinese function words as a demonstration (Figure 5). The trends of top two classical Chinese function words, "zhi 之" and "wu 無", are decreasing and uniform, respectively, compared to the increasing trend of top two modern Chinese function words "de 的" and "shi 是". The usage of "無" is especially interesting, and the uniform pattern indicates that its usage does not change with time which is quite unique among all function words.

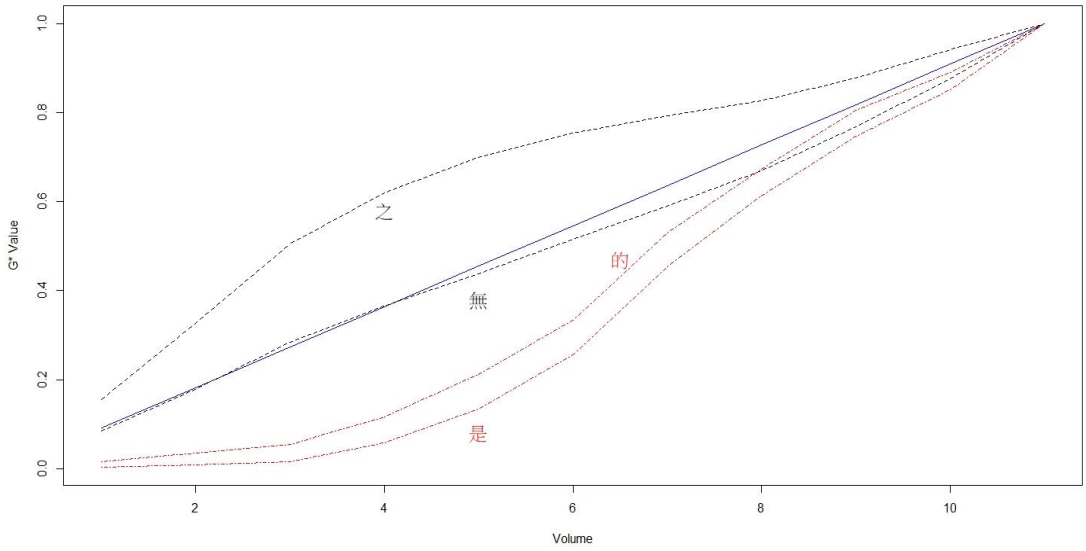

Figure 5. G* Values of Two Classical and Modern Chinese Function Words

c) punctuation:

The punctuation can also provide valuable information regarding Chinese writing style. Similarly to English, periods and commas are used most often in Chinese. However, commas are seldom used in the first three volumes of *Xin qingnian*, while they become the dominant punctuation (about two times the rate of periods) in Volumes 7-11 (Figure 6). The usage of "、" (or "*ton*"), which is a slight-pause mark, is quite surprising. The proportion of *ton* among all punctuations is about 35% (or 1/3) in Volume 3. This is an astonishing result, and deserves more exploration. One possible explanation is that maybe people were not accustomed to the punctuations and some might confuse *ton* with commas.

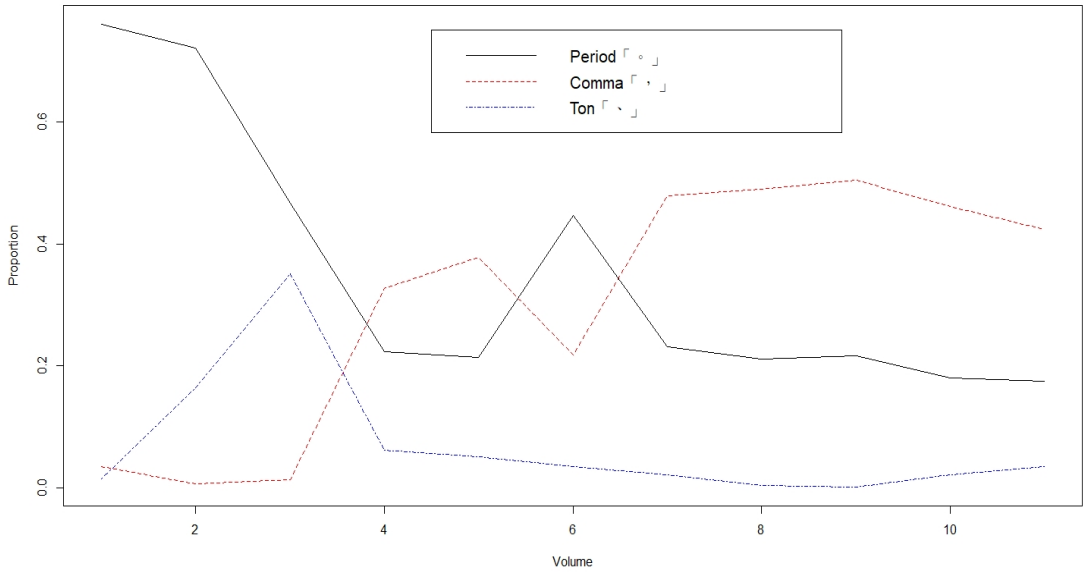

Figure 6. Proportion of Common Punctuation Compared to All Punctuation

Table 3. Number of Words in a Sentence in *the New Youth Magazine*

Vol.	1	2	3	4	5	6	7	8	9	10	11
Length	6.90	8.89	11.88	9.74	10.17	10.26	10.15	11.01	10.99	12.85	14.32

We can use the punctuation to calculate the average length of a sentence in the *New Youth Magazine*. In particular, the number of words between 5 possible punctuation marks, which are period, comma, semi-colon, question mark, and exclamation mark, is treated as the length of a sentence. The average length of a sentence is the shortest in Volume 1, and it gradually increases to more than double in Volume 11. Again, this is also evidence showing that there is a significant change in Chinese writing style in the period between 1915 and 1926.

4. Further Analysis of the *New Youth Magazine*

In the previous section, we compared the differences between classical and modern Chinese writing styles with respect to the vocabulary diversity and sentence structure, without considering the meaning of the words and phrases. The data analysis related to the meanings of words and phrases would be more complicated. In what follows we examine the top 10 words and the top 10 two-word phrases in the *New Youth Magazine* to explore the possible directions of data analysis if we include the statistical relevance) of texts Table 4 shows the top 10 words used in the 11 volumes and there are at least

two noticeable results. First, the words used most often in the 11 volumes are function words, "之" and "的", and they switch from classical Chinese to modern Chinese, matching the result in Figure 5. Another interesting result is the use of pronouns "他" (or He) and "我" (or I) starting with Volume 4. These two words are not in the top 100 words in Volumes 1 to 3, but do appear in the top 10 words in Volumes 4 to 9. (Note: They are on the top 50 words in Volumes 10 and 11.)

The top 10 words can also be treated as a tool to distinguish the different volumes of *the New Youth Magazine*. For example, we can count the top 10 words in common between two volumes (Table 5). For example, the numbers of words in common between Volumes 1 and 2 are 9 (2nd row and 3rd column). These numbers can be used to group the volumes and in general the number of top 10 words in common between two volumes decreases as the difference of volume number increases. Thus, it is natural to determine the grouping boundary by whether there is a big drop between two volumes. According to such a grouping system, we think that the 11 volumes can be classified into four groups with respect to the top 10 words: Volumes 1-3, 4-6, 6-9, and 10-11. The writing style of *New Youth Magazine* gradually switched from classical Chinese (Volume 1) to modern Chinese (Volume 7-11). The writing style of Volumes 4-6 can be treated as the intermediate stage between classical and modern Chinese.

Table 4. Top 10 Words in *the New Youth Magazine*

Rank	Vol.1	Vol. 2	Vol. 3	Vol. 4	Vol. 5	Vol. 6	Vol. 7	Vol. 8	Vol. 9	Vol. 10	Vol. 11
1	之	之	之	之	的	的	的	的	的	的	的
2	人	不	不	的	不	是	是	是	是	一	國
3	不	人	以	不	之	不	一	一	一	是	一
4	以	以	人	一	是	一	不	不	不	國	是
5	為	一	為	人	一	人	人	有	有	之	工
6	其	為	一	是	人	有	有	人	人	不	會
7	國	國	其	我	有	他	這	他	他	產	主
8	一	其	而	有	我	了	了	在	了	會	民
9	於	者	者	以	以	之	工	這	我	有	不
10	者	有	國	為	為	我	我	了	在	主	在

Table 5. Common Top 10 Words Between Two Volumes of *the New Youth Magazine*

Volume	1	2	3	4	5	6	7	8	9	10	11
1	10	9	9	6	6	4	3	3	3	4	3
2	9	10	9	7	7	5	4	4	4	5	3
3	9	9	10	6	6	4	3	3	3	4	3
4	6	7	6	10	10	8	7	6	7	6	4
5	6	7	6	10	10	8	7	6	7	6	4
6	4	5	4	8	8	10	8	8	9	6	4
7	3	4	3	7	7	8	10	8	8	5	5
8	3	4	3	6	6	8	8	10	9	5	5
9	3	4	3	7	7	9	8	9	10	5	5
10	4	5	4	6	6	6	5	5	5	10	7
11	3	3	3	4	4	4	5	5	5	7	10

Similar analysis can be applied to two-word phrases as well and the results are in Table 6. Again, the use of pronouns is very different for the first 3 volumes and later volumes. In the first 3 volumes, “吾人” and “吾國” are used and they are equivalent to the pronoun “we.” For the later volumes, “我們” and “他們” are used and they are the same as the pronouns “we” and “they”, respectively. The two-word phrases used to describe “now” are also different in the first 3 and later volumes. For the first 3 volumes, “今日” is among the top 30 list of two-word phrases, comparing to “現在” is among the top 30 list of two-word phrases in the later volumes. We can also apply the top 10 two-word phrases for classification, similar to that in Table 5. Of course, the results of volume classification depend on the variables chosen and it should be conducted with care.

Table 6. Top 10 Two-word Phrases in *the New Youth Magazine*

Rank	Vol.1	Vol. 2	Vol. 3	Vol. 4	Vol. 5	Vol. 6	Vol. 7	Vol. 8	Vol. 9	Vol. 10	Vol. 11
1	國家	青年	社會	先生	中國	社會	我們	他們	我們	革命	革命
2	政府	社會	文學	文學	我們	我們	他們	我們	他們	我們	我們
3	自由	世界	中國	中國	他們	他們	社會	一個	社會	社會	工人
4	社會	國家	吾人	我們	**現在**	**現在**	勞動	沒有	一個	他們	他們
5	薩稜	人類	世界	社會	先生	主義	工人	可以	沒有	經濟	中國
6	夫人	政府	先生	他們	文字	文學	**現在**	社會	所以	運動	運動

7	吾人	政治	吾國	**現在**	世界	中國	生活	**現在**	可以	對於	農民
8	政治	**今日**	政府	文字	社會	思想	時候	所以	就是	工人	經濟
9	人民	主義	道德	學生	文學	先生	人口	對於	不能	組織	英國
10	青年	國民	教育	白話	主義	經濟	問題	就是	什麼	中國	國家

All variables showing significant differences in the preceding analysis can be used to construct a classification model, if our goal is to differentiate where the articles come from. We used Volume 1 vs. Volume 7 and Volume 7 vs. Volume 11 to demonstrate the classification results. There are 162, 132, and 56 articles in Volumes 1, 7, and 11, respectively. Also, we considered cross-validation (CV) and divided the articles of each volume into 90% to the training set and 10% to the testing data. We considered all the variables in Sections 3 and 4, such as vocabulary diversity and top 10 words. In this study, we considered 47 independent variables and plugged them into the logistic regression. We repeated the CV process and randomly separated the data into training and testing sets 100 times, and the average accuracy (properly classifying articles) of classification is shown in Table 7. The classification results are fairly accurate, indicating that the variables chosen can effectively distinguish the differences between volumes.

Table 7. Classification Accuracy of *the New Youth Magazine*

	Volumes 1 vs. 7	Volumes 7 vs. 11
Training Data	96.10%	93.20%
Testing Data	95.95%	92.17%

5. Conclusion and Discussion

Big data not only makes our life more convenient but also changes how we behave. Through massive information accumulation and rapid data processing, we can instantly acquire facts, a task that used to take a lot of time. For example, maps and tour guides were necessary if we travelled to foreign countries or unfamiliar places in the past, but apps like Google Maps make things easier. Big data also changes how we study and do research. Take the famous American author Stephen King as an example. He started to use the pen name Richard Bachman in 1977 but the alias was never discovered by the

readers or publishers until he revealed this fact to the press later. J.K. Rowling, the author of Harry Potter, also used the pen name Robert Galbraith to publish novels in 2013 but this fact was soon identified by computer experts (Archer and Jockers, 2016). Text mining was the key in discovering that J.K. Rowling and Robert Galbraith were the same person, and recognizing an author's writing style is only one of the possible applications of text mining.

In order to proceed with text mining, first we need to transform texts into meaningful data which can be analysed by computers. This process is also called structurization and there are no unified standards for text structurization as of yet. In this study, we propose an approach to retrieve information from Chinese texts, by treating words as species and applying the notion of species diversity in data analysis. In specific, we introduce Tukey's Exploratory Data Analysis (EDA) to extract important information from texts. We apply the EDA tools to explore the change of writing style from classical Chinese to modern Chinese in the early 20th century. It is believed that the May Fourth Movement in 1919 was central in the change of writing style and *the New Youth Magazine* (11 volumes), published in 1915-1926, is a good candidate for observing this style change.

EDA can achieve at least three objectives: (1) Data cleaning and identifying unusual observations or variables; (2) Evaluation of findings from past studies; (3) Exploring and discovering new information. Humanities scholars can obtain useful information, especially from the last two objectives. For verifying the findings from past studies, we can use two examples as a demonstration. Classical Chinese is believed to be more diversified than modern Chinese, and this matches to the results of Table 2 and Figure 2 (TTR), where there are more vocabularies used in the early volumes of *New Youth Magazine*. Moreover, it is expected that the length of a sentence is shorter in classical Chinese, and this is confirmed from Table 3.

On the other hand, the EDA results can often provide precise information for future study. For example, the use of Chinese punctuations was not unified before Chinese government announced the regulation in 1920. However, it is not clear the process how the general public learned to use the punctuation. The results in Figure 6 suggest that people might mistake slight-pause mark "、" for comma "，" in the first three volumes before the use of punctuation was regularized. Also, the use of comma is at least twice that of period "。" in modern Chinese writing (Volumes 7~11), but it was not the case in Volumes 1~6 of *New Youth Magazine*. The analysis results of punctuation open new possibilities of further study.

The EDA results of Chinese writing style can also be separated into three parts: vocabulary diversity, function words, and punctuations, and we found many differences between volumes from the proposed approach. It is believed that the early volumes

belong to classical Chinese and later volumes are modern Chinese. The significant differences in all three parts of the analysis indicate that the proposed approach can be used to describe the changes in Chinese writing. These analysis results are also plugged into the classification model and achieve high accuracy in identifying where the articles are from. It seems that the variables detected by EDA can lead us to further study of writing style.

Note that the purpose of this study is to introduce EDA tools, as well as standard operating procedures of data analysis, in order to identify key features which can differentiate the change in Chinese writing style. Of course, we can definitely apply these features to Confirmatory Data Analysis (CDA) for constructing accurate statistical and machine learning models as shown in Table 7. We think that these key features can also provide insightful suggestions for future research. For example, the proportion of the top 500 words is more than 80% for Volumes 7-11 (modern Chinese) and it is much higher than the texts used in Taiwan today (around 70%). It would be interesting to trace the writing style of modern Chinese in Taiwan since 1949.

The proposed EDA approach do not include the meaning of words and phrases, and we only focused on analysing the information related to vocabulary diversity and sentence structure. This will surely restrict the scope of EDA and including the meanings of words and phrases can provide more useful information. For instance, Tables 4 and 5 show that the use of pronouns are very different in classical and modern Chinese, and this suggests that the pronouns are also an important feature in the study of Chinese writing style. We will continue exploring the EDA techniques and try to modify the proposed approach to cover the meaning of words and phrases.

References

Primary sources:

Secondary sources

Archer, J. and Jockers M.L. (2016). *The Bestseller Code: Anatomy of the Blockbuster Novel*. New York: St. Martin's Press.

Ash, A. (2009). "China's New New Youth." DigitalCommons@University of Nebraska-Lincoln. https://digitalcommons.unl.edu/chinabeatarchive/626/.

Ash, A. (2019). "New Youth in China". Dissent.

https://www.dissentmagazine.org/online_articles/new-youth-in-china-may-fourth-anniversary.

Brillinger, D. (2011), "Data Analysis, Exploratory", *International Encyclopedia of Political Science*, pp. 530-537, Badie, Berg-Schlosser & Morlino ed., SAGE.

Efron, B. and Thisted, R. (1976). "Estimating the Number of Unseen Species: How many Words did Shakespeare Know?" *Biometrika* 63, pp. 435-447.

El Morr C., Ali-Hassan H. (2019). "Descriptive, Predictive, and Prescriptive Analytics," In: Analytics in Healthcare: A Practical Introduction. SpringerBriefs in Health Care Management and Economics. Springer, Cham.

Forcina, A. and Giorgi, G.M. (2005). "Early Gini's Contributions to Inequality Measurement and Statistical Inference," *Handbook on Income Inequality Measurement*, pp. 245-260, Silber ed., Springer.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition, Springer Series in Statistics.

Ho, L., Yue, C.J., and Cheng, W. (2014), "From Classical Chinese to Modern Chinese: A Study of Function Words from New Youth Magazine", *Journal of the History of Ideas in East Asia* 7, pp. 427-454. (in Chinese)

Li, K. and Dew, J.E. (2009). *Classical Chinese: A Functional Approach*, Cheng & Tsui Co.

Manschreck, T.C., Maher, B.A. and Ader, D.N. (1981). "Formal Thought Disorder, the Type-Token Ratio, and Disturbed Voluntary Motor Movement in Schizophrenia," *The British Journal of Psychiatry* 139(1), pp. 7-15.

Praveen, S. (2017). "Influence of Structured, Semi-Structured, Unstructured Data on Various Data Models," *International Journal of Scientific & Engineering Research* 8(12), pp. 67-69.

Tukey, J.W. (1962). "The Future of Data Analysis," *Annals of Mathematical Statistics* 33(1), pp. 1-67.

Tukey, J.W. (1977). *Exploratory Data Analysis*, Addison-Wesley.

Yue, C.J., Clayton, M., and Lin, F. (2001), "A Nonparametric Estimator of Species Overlap," *Biometrics* 57(3), pp. 743-749.

Yue, C.J. and Clayton, M. (2005). "A Similarity Measure based on Species Proportions," *Communications in Statistics: Theory and Methods* 34, pp. 2123-2131.

Yue, C.J., Ho, L., Pan, Y. and Cheng, W. (2016). "A Quantitative Study of Chinese Writing Style, Based on New Youth," *Concepts & Context in East Asia* 5, pp. 87-102.