**HO — Identifying**

<1 figure, 1 table>

**Identifying Classical and Written Vernacular Chinese Language: A Statistical Study Based on Texts from *New Youth* Magazine (1915–1922)**

Ho, L., Yue, C.-S. Y. and Cheng, W.

The year 2015 marks the 100th anniversary of *New Youth* magazine (1915–1922), the single most influential periodical in the history of modern Chinese literature and language. It was in *New Youth* that the progressive writers advocated to replace classical Chinese with written vernacular as the standard of modern written Chinese language. At first, just like other literati in late the Qing and the early republic, they wrote most of their arguments on language changing in classical Chinese. The dramatic shift of language happened after the May Fourth movement in 1919, and the practice of writing vernacular in every genre spread from the magazine to other publications across the country, finally changing the landscape of written Chinese for good. For a century, scholars have emphasized the importance of *New Youth* but failed to answer questions such as how long it took for the literati as well as the general public to accept the use of a new written standard and exactly how the new standard spread. Teaching computers to distinguish between classical and written vernacular Chinese may make it possible to bring in more digitized texts from that period to study and find the answers.

In this study, we adapt the concept of statistical learning theory and propose two types of quantitative analysis, supervised learning and unsupervised learning, to explore the writing style of classical Chinese and modern Chinese. These two types of methods differ in whether the researchers provide keywords. The unsupervised methods, unlike the supervised methods, are data-driven and determine keywords based on the data attribute and pattern. Since there is no standard operating procedure for applying the unsupervised methods, we will apply the idea of species diversity to determine the keywords.

Supervised learning has long been used in linguistics to differentiate authorship via keywords, and we choose 10 function words each for classical and modern Chinese as the keywords. Together with a proposed dispersion measure G\* index, we can evaluate the trend of function words used in volumes 1 and 11 of *New Youth* magazine. For example, the following graph (Figure 1) shows the trends of G\* indices for 20 function words chosen. Note that, if the G\* curve of a function word is always under the diagonal line, then the frequency of this word is increasing. On the other hand, the G\* curve of a function word over the diagonal line indicates that the word frequency is decreasing. In other words, the function words of classical Chinese become infrequent while those of modern Chinese become more common in *New Youth* magazine.

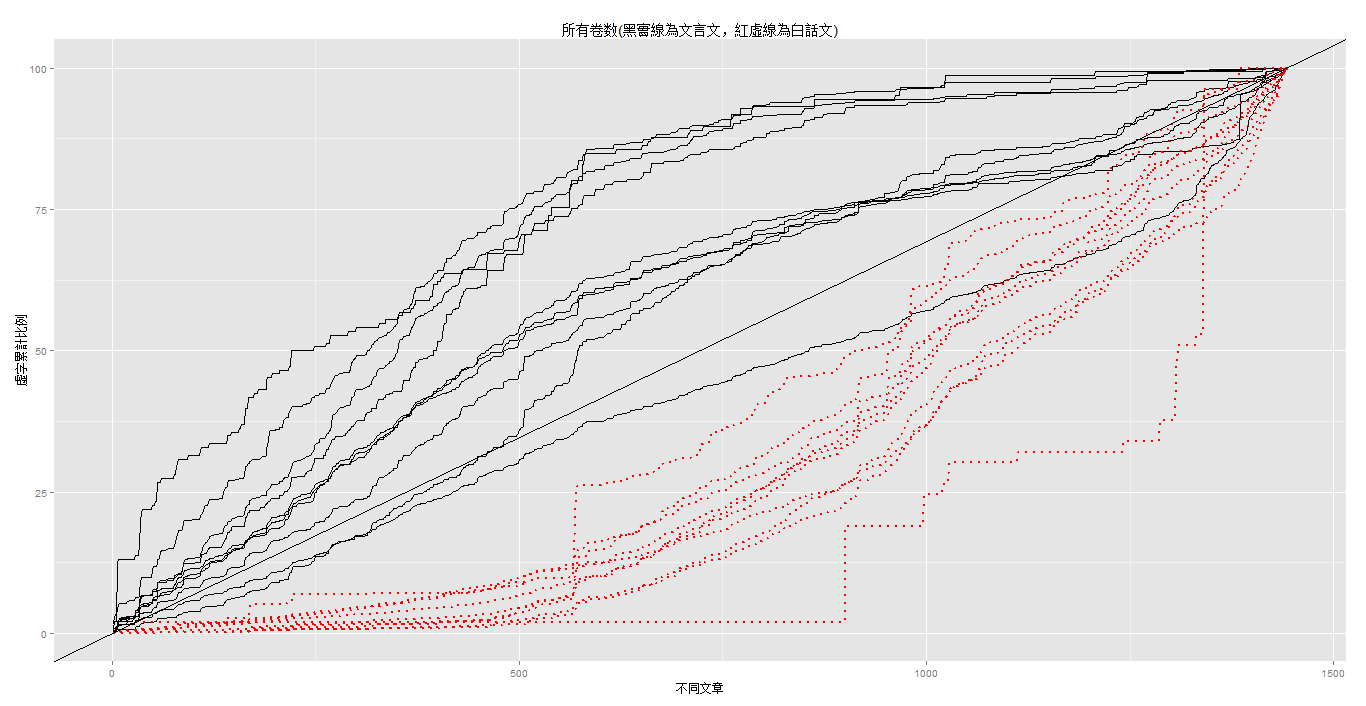


Figure 1. G\* curves of 20 function words (black line: classical Chinese; red line: modern Chinese).

For the unsupervised learning, we will focus on comparing species (or vocabulary) richness, unevenness, and sentence structure. We use numbers of words and vocabularies to measure the richness. There are quite a lot of unevenness measures, such as Gini’s index, entropy, and Simpson’s index, and they are often used for measuring the statistical dispersion. The analysis of sentence structure includes its length and the punctuation in a sentence. We will use an example to demonstrate our approach by judging the species richness and unevenness. Table 1 shows the related statistics of 11 volumes of *New Youth* magazine. The number of words is increasing from volume 1 to volume 11, but the number of vocabularies shows a decreasing pattern. This indicates that the species diversity becomes lower. The unevenness measures reveal similar information, and earlier volumes have higher species diversity. Note that a smaller value of Simpson’s index (and a larger entropy value) indicates higher species diversity. Based on these analyses, it seems that later volumes have lower species diversity and people can read the articles without recognizing many words, which matches the purpose of the May Fourth Movement.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Volume | Words Count | Vocabularies  Count | Simpson Index | Entropy |
| 1 | 248,833 | 4,379 | 0.004568 | 6.654036 |
| 2 | 291,848 | 4,344 | 0.004500 | 6.649539 |
| 3 | 290,038 | 4,227 | 0.004954 | 6.541824 |
| 4 | 305,020 | 4,298 | 0.004172 | 6.539378 |
| 5 | 343,519 | 4,125 | 0.004672 | 6.461579 |
| 6 | 389,407 | 3,848 | 0.005749 | 6.348547 |
| 7 | 586,942 | 3,850 | 0.006053 | 6.328604 |
| 8 | 461,731 | 3,753 | 0.006035 | 6.320355 |
| 9 | 437,748 | 3,745 | 0.005574 | 6.322103 |
| 10 | 342,778 | 2,980 | 0.005700 | 6.177278 |
| 11 | 489,223 | 3,093 | 0.005712 | 6.212699 |

Table 1. Descriptive statistics for *New Youth* magazine.

In addition, we will apply the measures in this study and use them to classify an article for whether it belongs to the group of classical or modern Chinese. First, the articles from all 11 volumes of *New Youth* magazine are identified manually into the groups of classical or modern Chinese. Next, we use cross-validation to determine feasible classification functions (and the best one, if possible). We separate the articles in *New Youth* magazine into two sets: training data and testing datasets. The training data are used to select important variables and construct the classification function, and this function is used to determine the labels of the testing data. We then calculate the classification error (or prediction error) of the testing data, and the function with the smallest error is treated as the best model. The process of cross-validation is often repeated several times (e.g., 1,000 times) to reduce the influence of chance. Usually we choose the classification function with the smallest average prediction error from 1,000 replications. Also, we can check the important variables in the classification function and examine possible interpretations of these variables. We expect the interpretations are helpful in differentiating the styles between classical Chinese and modern Chinese.