A Corpus-Based Study of America Media Discourse on Traditional Chinese Medicine

Yafei Li[1,a,*]

[1]*School of Foreign Languages, Qingdao University, Qingdao, 266071, China*
a. Lee_Yafei@outlook.com
*corresponding author

Abstract: Traditional Chinese Medicine (TCM) has played a significant role in the lives of the Chinese people and is gradually entering the global arena. The process of its globalization can be mirrored by the mass media, such as newspapers, due to their capacity of reflecting the changes in the culture and ideology of a certain region. This study explored the changing trends in America Media Discourse on TCM from 1860 to 1959 in the aspects of vocabulary, readability, and content, based on the analysis of a self-built corpus. The findings reveal that over a century, American media reports on TCM have been picking words that are relatively simple and close to daily life, with a low content of academic vocabulary; scored high for the Flesch Reading Ease Readability formula, decreasing the difficulty of being understood by readers; ranged from shallow to deep in the overall perspective, reflecting the historical process of TCM's transmission and development in global terms.

Keywords: corpus, Traditional Chinese Medicine (TCM), text analysis, media discourse, readability

1.   Introduction

Against the backdrop of rapid globalization and the progress of information technology, international mass media have become an important channel for people to access and understand China, which has always attracted the attention of the world because of its profound history and rich cultural heritage. Hence, international media discourses on Chinese culture have the power of shaping the world's perception of China, resulting in the accumulating attention paid by academics to the construction of China's image from the media perspective. Some scholars have analyzed and researched the China-related international media discourse in terms of text mining and co-word clustering. Yuting Lin and Xingwei Miao used Leximancer, a visual text mining software, to analyze the discourse construction and change of China's image in the U.S. mainstream media's China-related news reports over a span of 40 years since the formal establishment of diplomatic relations between China and the U.S. in 1979 [1]; Xiang Xu conducted a quantitative analysis of the content of Chinese culture reported by international media, and found that cultural symbols with Chinese characteristics were reported with low intensity [2]; Hui Wang and Bin Xin interpreted the image of China portrayed by American media through a study of the metaphorical construction of China's image in American media [3].

As an important element of traditional Chinese culture and the crystallization of the wisdom of the Chinese nation, Traditional Chinese Medicine (TCM) is unique in the world of medicine and culture, and its foreign dissemination and international influence can reflect the image of China in the international discourse. "If the clinical diagnosis and treatment is the driving force and the root of the development of Chinese medicine, then the culture of TCM is the essence and soul of the inheritance of Chinese medicine [4]." In 2015, Chinese medicine scientist Youyou Tu won the Nobel Prize in Physiology or Medicine for her great contribution to medicine by refining artemisinin. "The discovery of artemisinin is precisely thanks to the ancient Chinese medicine book The Ge Xianweng Zhouhou Beiji Fang (葛仙翁肘後備急方), or Transcendent Master Ge's Emergency-Preparedness Recipes to Keep Close at Hand [5]. It can be seen that the therapeutic efficacy of the Chinese medicine artemisinin has been recognized by the world. In addition, TCM also played an important role in China's fight against COVID-19. "Among the confirmed cases of COVID-19 in China, 74,187 people used Chinese medicine, accounting for 91.5%. Clinical efficacy observation shows that the total effective rate of TCM reaches more than 90% [6]." Traditional Chinese medicine can effectively relieve symptoms, prevent the mild forms of phenomena from escalating to severe forms, and promote the recovery of people in the recovery period. Therefore, the international media discourse on TCM has great research value in various fields.

2. Research Design

2.1. Corpus Construction

The corpus is composed of 10 sub-corpora, each of which has around 1,000 words in size, containing the American newspaper report data on TCM during periods of "1860-1869", "1870-1879", "1880-1889", "1890-1899", "1900-1909", "1910-1919", "1920-1929", "1930-1939", "1940-1949", and "1950-1959" respectively.

The primary source of this corpus is newspaper photographic images from the Library of Congress website [7]. In this study, OCR (Optical Character Recognition) technology was utilized to detect the images screened from the website, as well as extensive manual proofreading. To ensure the structural integrity of the text, the reports selected according to the randomization principle were included integrally as far as possible, which resulted in the size of each sub-corpus being slightly different. In addition, due to a variety of possible factors such as improper preservation or less coverage, the corpus resources for some periods are relatively scarce, which is also an important reason for the limited content of certain sub-corpus.

2.2. Data Processing

This study analyzed the American media discourse on the topic of TCM quantitatively and qualitatively in terms of vocabulary, readability, and content, with the help of tools such as Python programming and Voyant Tools.

2.2.1. Study on Vocabulary Based on AWL

In order to further study the discourse style of American media reports on TCM from 1860 to 1959, this study adopted the Academic Word List (AWL) developed by Averil Coxhead of School of Linguistics and Applied Language Studies at the Victoria University of Willington as the reference data [8]. Sticking to the sampling criteria of academic words, Averil Coxhead filtered the 2000 words with the highest usage rate in English, and then filtered out 570 word families of academic words, which amounted to 2,577 words in total. Based on the Academic Word List, this study used Python coding to count the total number of academic vocabulary words as well as the total number

of words in the text of 10 sub-corpora, and then calculate the average percentage of academic vocabulary words. Part of the code is attached below:

```
import os
path1 = r'G:\text\1950-1960'
files = os.listdir(path1)
for i in range(0,6):
    a = 0
    domain = os.path.abspath(r'G:\text\1950-1960')
    file = os.path.join(str(domain),files[i])
    fn = open(str(file),encoding='utf8')
f = fn.readlines()
    #print(f)
    for line in f:
        line = line.rstrip('\n')
        for j in awl:
            if j in line:
                print(j)
                a=a+1
    print(a)
```

2.2.2. Study on Readability Based on the Flesch Reading Ease Readability Formula

Since the very beginning of the 20th century, text readability has been one of the difficulties and focuses in the field of linguistics [9]. The most common standard for examining the legibility of certain text is the Flesch Reading Ease Readability Formula, which takes word length (WL), which means the number of syllables per 100 words, and sentence length (SL), which means the average number of words per sentence, as two significant indicators [10]. Under normal circumstances, the more polysyllabic words and the longer the sentences in an English text, the lower reading readability is scored and the more difficult it is to be understood.

$$0.39 \left(\frac{\text{total words}}{\text{total sentences}} \right) + 11.8 \left(\frac{\text{total syllables}}{\text{total words}} \right) - 15.59 \tag{1}$$

Thanks to the Flesch readability measurement tool embedded in Microsoft Word [11], word length, sentence length, and readability score can be obtained.

2.2.3. Study on Content Based on Voyant Tools

For the content of media reports, this paper used Voyant Tools to analyze the themes of the texts as well as the focus of American newspaper reports on traditional Chinese medicine in different periods through the functions of Cirrus (word frequency analysis) and Links (topic word analysis). Specifically, the words like "TCM", "medicine", "Chinese", and "China", which were used as search terms to collect text data, have been attached to the stop words list in Voyant Tools, so as not to interfere with the word frequency statistics of the text.

3. Results and Discussion

3.1. Vocabulary

The statistical results show that the average proportion of academic vocabulary in the 10 sub-corpora, although fluctuating, is consistently low, with a maximum of only about 3%, which to

a certain extent reflects the discourse characteristics of American media reports on the topic of TCM over the span of 100 years. On the one hand, the use of plain and everyday vocabulary can promote the development of an easy-to-read style, assisting the media in achieving its goal of wide dissemination; on the other hand, because TCM has a cross-cultural character in the American media, related reports are conducted in a simpler manner to help readers perceive things from a new culture.

Table 1: Statistics on the count and percentage of academic terms.

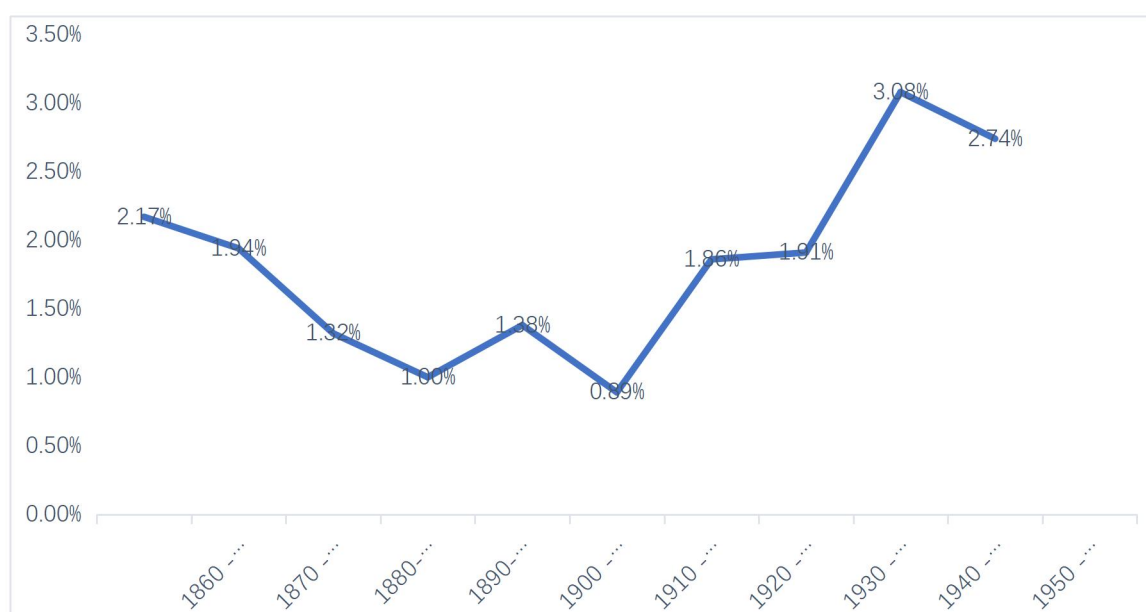| Time	The Count of Academic Words	The Count of Words	Academic Words/Words
1860 - 1869	18	826	2.17%
1870 - 1879	21	1,078	1.94%
1880- 1889	13	981	1.32%
1890- 1899	7	699	1.00%
1900 - 1909	16	1,157	1.38%
1910 - 1919	10	1,125	0.89%
1920 - 1929	20	1,076	1.86%
1930 - 1939	17	890	1.91%
1940 - 1949	31	1,008	3.08%
1950 - 1959	23	840	2.74%

Figure 1: Line graph of the percentage of academic vocabulary in the reports.
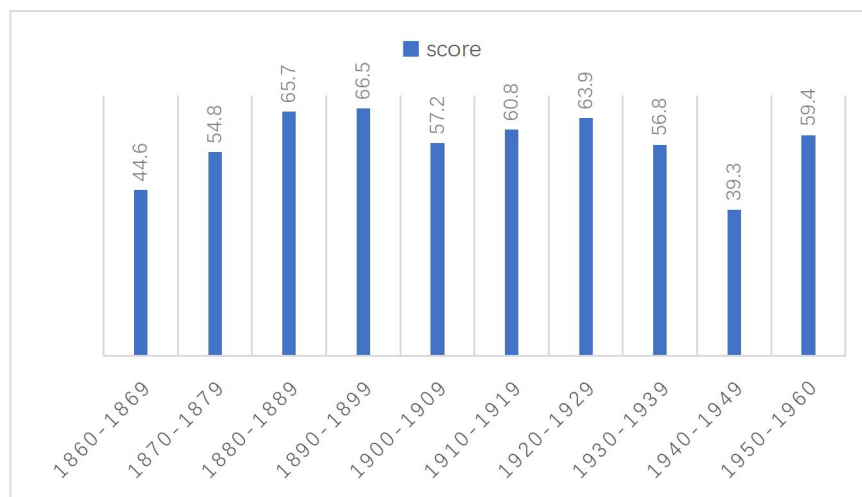
3.2. Readability

Figure 2: The Statistical Graph of the Flesch Reading Ease of 10 Sub-Corpora.

On the basis of the Flesch Reading Ease Readability Formula, this study calculates the corpus covered by the 10 sub-corpora separately, and the results are shown in the following graph.

According to the results, 80% of the relevant reports have a Reading Ease Score that exceeds 50, proving that the American media reports on TCM are basically easy to read, the language used is concise and direct, and the opinions are clearly expressed, in alliance with the social functions of newspapers as one kind of mass media. To an extent, American media discourse has played a role in promoting people's understanding of TCM.

However, the reports before the 1870s lacked strong relevance to TCM and were relatively difficult to read, but they were able to show some traces of TCM entering the eyesight of people in the Western world. Additionally, there was less coverage of TCM in the American media during the 1940s, and the readability score for texts of this period is 39.3, which is the lowest among the total ten sub-corpora.

3.3. Content

Combining the results shown in Table 2 and textual content, it is found that the intensity and focus of American media coverage of TCM have undergone many changes from 1860 to 1959.

In the 1860s, American newspapers seldom reported on traditional Chinese medicine as a separate topic. TCM usually appeared together with the American media's reports on other East Asian countries, such as Japan, etc. Between 1870 and 1900, the number of American newspaper reports on TCM gradually increased. Most of these reports tended to focus on Chinese doctors and hospitals, and some of the reports' descriptions of traditional Chinese medicines had a certain literary coloring. In the reports after 1900, "rattlesnake", "salve", "lizard", "alcohol", and other traditional Chinese medicine materials were mentioned more and more, and the expression of the comparison between Chinese and Western medicines was also improved. Some articles gave amazing praise to the efficacy of TCM, while some of them held a different opinion. All indications are that the American media discourse on TCM went from shallow to deep, conveying the curiosity of the American public about Chinese medicine in that era, as well as a relatively objective appraisal of Chinese medicine's ability to cure illnesses without knowing the mechanism.

Table 2: Top 5 high-frequency words in the 10 periods.

Period	Most Frequent Words
1860-1869	Japan (9); government (8); port (7); foreign (7); Kanagawa (15)
1870-1879	practice(8); pulse (5); physicians (5); doctors (4); disease (4)
1880-1889	year(9); women (6); death (5); cases (5); hospital (4)
1890-1899	boot (5); tree(4); day (4); patient (3); use (3)
1900-1909	oil (12); rattlesnake (11); Washington (5); snake (7); salve (5)
1910-1919	York (11); doctor (10); men (9); foreign (8); treatment (6)
1920-1929	bear (9); practice (4); mouth (4); killed (4); teeth (3)
1930-1939	lizard (7); alcohol (7); industrial (5); use (4); turtle (4)
1940-1949	year (7); county (5); training (4); united (3); surgeon (3)
1950-1959	western (9); hospital (8); patient (5); use (4); traditional (4)

4.   Conclusion

Based on a self-constructed small-scale specialized corpus, this study focused on American media discourse related to TCM and conducted textual research and thematic content exploration at three levels: vocabulary, readability, and content, providing a new perspective and data basis for the study of the dissemination and development of TCM in the global context. Due to the relatively small size of the corpus, it is not able to present a more particular picture of the changes in the American media discourse on TCM, and more factual evidence is still required to corroborate the data discovered in this study mutually. However, in general, the texts of American media reports on TCM between 1860 and 1959 are easy to read, and the angles of the reports vary from shallow to deep with the development of time, depicting the historical process of the dissemination of TCM to the broader world.

References

[1]   Yuting Lin, Xingwei Miao. Discourse Construction of China's National Image in American Mainstream Media--An Ephemeral Study Based on Text Mining [J]. Journal of PLA College of Foreign Languages ,2023,46(03):87-95.

[2]   Xiang Xu. Chinese Culture and Its Typological Characteristics in International Social Media Communication: An Empirical Study Based on Online Text Mining [J]. Journal of Harbin Institute of Technology(Social Science Edition),2017,19(02):59-67+2.

[3]   Hui Wang, Bin Xin. A Study on the Metaphorical Construction of China's Image in American Media--Taking the Report on "U.S. Withdrawal from the TPP" as an Example [J]. Foreign Language Teaching,2019,40(03):32-38.

[4]   Qicheng Zhang. Culture of Traditional Chinese Medicine [M]. Beijing: People's Health Press, 2017:107.

[5]   Lili Wang, Jingshu Liu. Exploration of the Development path of Chinese Medicine Education in Russia under the "Belt and Road" War [J]. China Medicine Guide, 2017, 15(26):287 28.

[6]   Hao Lan, Linxiao Xiong, Feng Chen. Status Quo, Problems and Solutions for Promoting Traditional Chinese Medicine Culture in Russia under "the Belt and Road Initiative" [J/OL]. World TCM: 1-5[2023-07-30].

[7]   Home: Library of Congress (2023).https://www.loc.gov.

[8]   The Academic Word List (2023).https://www.wgtn.ac.nz/lals/resources/academicwordlist.

[9]   Yanwei Wang. A Comparative Study on the Readability of Professional English Textbooks and Test Reading Texts[J]. Foreign Language Research,2021,38(02):70-75.

[10]   Flesch R (1948). "A new readability yardstick". Journal of Applied Psychology. 32 (3): 221-233. doi:10.1037/h0057532. PMID 18867058.

[11]   Qinfang Zhuang. Discussion on the Readability of English Text Dissemination[J]. Modern Communication,2017,39(12):163-165.