Re-navigating the Vernacular Language Movement and Chinese Translation Literature, 1898-1938: An Examination of Prefaces Using Topic Modeling

Chen, Sixing
chensx31@connect.hku.hk
The University of Hong Kong, Hong Kong S.A.R. (China)

Du, Keli
duk@uni-trier.de
Trier Center for Digital Humanities, Germany

Li, Jin
13671123112@163.com
Renmin University of China, China

Introduction

The second wave of Western learning initiated the prosperity of translation literature in modern China. Lydia Liu proposed the so-called “translated modernity” to understand the transition of Chinese society through the import, distortion, and indigenization of new words. (Liu, 1995) The first question Chinese translators had to address was which Chinese language would be suitable for translating foreign works. There are two general styles of written Chinese, Classical Chinese and Written Vernacular Chinese. The latter includes Mandarin, other dialects (such as Cantonese, Fuzhou dialect, etc.), and Romanization. There was a call for vernacular Chinese since the 1860s. The cultural status of varieties of the Chinese language changed along with the Vernacular Language Movement (VLM). This project aims to re-examine the VLM by applying topic modeling to the preface collection. Only the preface is examined because it denotes the translator’s purpose and translation strategy.

Approach

Our corpus includes 2314 prefaces of translated works in Chinese from 1894 to 1938, covering around 1.08 million words. As the length of each preface varies significantly, we first arranged the prefaces by year and then combined those in the same year into one non-spaced text. Next, we split the texts into 1000-word chunks to train the topic models. Following Schofield’s instruction (Schofield et al., 2017), the most frequent function words were considered stop words and should be removed. Several topic models with different numbers of topics (20, 30, 40) have been trained by using Mallet (McCallum, 2002) with or without hyperparameter optimization. We notice that several topics always appear in different topic models trained with different parameter settings. We choose the topic model with 20 topics, as the distinction between the attributes of the topics is more distinct. Topic 16 [1] always contains words related to the VLM. Its top words are visualized in Fig.1.

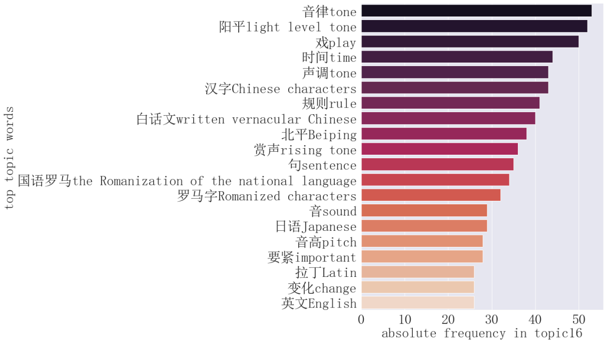


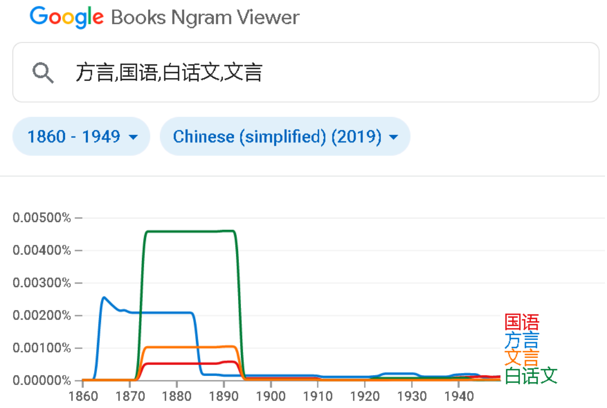
Fig.1 Topic16: top topic words

Results

Topic16 primarily addresses the VLM issues, indicating the relation between language and translation literature. Three points are observed in Topic 16.

First, Written Vernacular Chinese gradually replaced Classical Chinese as the mainstream written language. The words “白话文 Written Vernacular Chinese(40) [2],” “国语罗马 Romanization of the national language(34),” “罗马字 Romanized characters(32),” and “国语 the national language(24)” are related to Written Vernacular Chinese, which showed a high frequency over “文言Classical Chinese(11)” in the topic. Notably, Romanization had been highlighted along the VLM, which is embodied in the high-frequency words such as “国语罗马(34),” “罗马字(32),” “拉丁Latin(26).” It shows the crisis of the Chinese written system: whether Chinese characters were a hindrance to enlightening Chinese people.

Second, Mandarin distinguished itself from the VLM as it was decreed as the national language in 1909. Although other dialects were also deemed as vernacular Chinese, the frequency of “方言dialect(5)” is much lower than that of “国语the national lan-guage(24)” in the topic. The frequency of “方言dialect” in Google Ngram Viewer was higher than that of “国语the national language.” In future work, we intend to apply sentiment analysis to examine the emotion expression toward dialect and the national language to investigate whether other Han dialects were marginalized by the national language. If the answer is yes, the sentiment related to “dialect” should be negative.



Google Books Ngram Viewer

方言,国语,白话文,文言

1860 - 1949

Chinese (simplified) (2019)

Fig.2 Google Ngram Viewer search results of “方言dialect,” “国语the national language,” “白话文Written Vernacular Chinese,” and “文言Classical Chinese”

Third, “ 北平Beiping,” the name of Beijing before 1949, appears in this topic. It suggests that Beijing Mandarin surpassed Nanjing Mandarin to become the basis of the national language. To verify this assumption, we plan to apply Word Embedding Models to compare the distance between “ 北京Beijing” and “ 国语the national language” as well as that between “ 南京Nanjing” and “ 国语the na-tional language” in future research.

Notes

1. The topic16 comes from a topic model with 20 topics, all of which can be found here: https://drive.google.com/file/d/1-RzEPaKQkY1UKuab6NripNvO3K5vC1oT/view?usp=sharing
2. The number here represents the absolute frequency of the word in topic16.

Bibliography

Liu, L. H. (1995). *Translingual Practice: Literature, National Culture, and Translated Modernity–China, 1900-1937*. Stanford University Press.

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit.

Schofield, A., Magnusson, M. ans and Mimno, D. (2017). Pulling out the stops: Rethinking stopword removal for topic models. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. pp. 432–36.

Tang, X. and Su, Q. (2022). That Slepen Al the Nyght with Open Ye! Cross-era Sequence Segmentation with Switch-memory. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 7830–40 doi: 10.18653/v1/2022.acl-long.540. https://aclanthology.org/2022.acl-long.540 (accessed 21 October 2022).

Zhang, X., Wu, P., Cai, J. and Wang, K. (2019). A Contrastive Study of Chinese Text Segmentation Tools in Marketing Notification Texts. *Journal of Physics: Conference Series*, 1302(2): 022010 doi: 10.1088/1742-6596/1302/2/022010. https://iopscience.iop.org/article/10.1088/1742-6596/1302/2/022010 (accessed 21 October 2022).