

## A Computational Approach for Dating Chinese Texts: Insights from the Biji Text

This study investigates how computational methods can inform the periodization of Chinese by analyzing ancient texts from the Tang to the Qing dynasties. Traditional approaches to Chinese linguistic periodization rely heavily on close reading of historical texts, a process that requires extensive domain expertise (e.g., Wang, 1958; Pan, 1989). Although recent computational studies (e.g., Tian and Kuebler, 2021; Zhou et al., 2024) have explored automated dynasty prediction, they have largely emphasized model design and benchmark construction, with limited attention to linguistic interpretation. To address this gap, we introduce a newly constructed corpus of *Biji* (笔记, “written notes”), a genre characterized by rich conversational language and diverse subject matter.

The corpus comprises approximately 17 million characters drawn from 249 works spanning the Tang through Qing dynasties. From this corpus, we derive a dynasty classification dataset covering four major dynasties, each represented by at least 30 books. The dataset is designed for both sentence-level and paragraph-level classification and employs five-fold cross-validation with manually curated splits to ensure that training and test data come from different works, thereby improving robustness.

We systematically evaluate the effectiveness of different feature representations for dating Ancient Chinese, including character- and word-based n-grams, static embeddings, and contextualized embeddings, in combination with multiple learning algorithms (SVMs and GRUs) and pre-trained language models. Overall, the task proves highly challenging. Contextualized embeddings—particularly when paired with SVM classifiers—yield the strongest performance, achieving an average F1 score of 62.11 across five folds. Furthermore, we explore linguistic similarity across dynasties through neighboring-dynasty classification. Under the assumption that greater classification difficulty reflects greater linguistic similarity, our results show that texts from the Ming and Qing dynasties are especially difficult to distinguish, suggesting that they may constitute a single linguistic period.

### Reference

Pan, Yunzhong (1989). *Hanyu Yufa Shi Gaiyao* (Summary of Chinese Grammar History). Shanghai Classics Publishing House, Shanghai.

Tian, Zuoyu and Sandra Kubler (2021). “Period classification in Chinese historical texts.” In: Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pp. 168–177.

Wang, Li (1958). *Hanyu Shigao* (Manuscript of the History of Chinese Language ). Science Press, Beijing.

Zhou, Bo, Qianglong Chen, Tianyu Wang, Xiaomi Zhong, and Yin Zhang (2023). “WYWEB: A NLP Evaluation Benchmark For Classical Chinese”. In: Findings of the Association for Computational Linguistics: ACL 2023, pp. 3294–3319.