Abstract

The chronological classification of texts can be crucial for clarifying authenticity and interpretation. For several Western languages, statistical language models (*SLMs*), amongst other methods, have been proven useful for automatically assigning timespan (*chronon*) labels to texts. This is made possible by changes in style, grammar, vocabulary, and phonology. When it comes to Classical and Literary Chinese sources, dating can be complicated not only by the existence of forgeries, complex textual lineages, and obscure authorship, but also the fact that many genres attempted to remain faithful to ancient rhetorical and linguistic patterns. Additionally, major phonological changes are not often reflected in Chinese script. The present study assesses both new and established computational dating methods and examines related issues in computational processing of Literary Chinese.

The history of the Chinese written language is the starting point for this study. On the basis of official dynastic histories (*zhengshi* 正史), it is shown that both grammatical and lexical changes can give clues to the time of production, even within a corpus of stylistically homogenous texts. Tied to the appearance of new concepts, lexical innovation is found to be a key indicator of the time of a text's creation.

Readers are then introduced to the current state of research on computational and philological methods of textual dating, emphasizing *SLMs*. The problems of Chinese word segmentation, the lack of diachronic Chinese corpora, as well as recognition of named entities and temporal expressions are discussed. A diachronic Chinese lexeme database for making lexical changes useful for dating is constructed from the earliest word use attestations (*loci classici*) sourced from a digital version of the *Hanyu da cidian* 漢語大詞典.

The main part of the study is dedicated to the development, testing, and comparison of dating methods for Literary Chinese texts. *SLMs* are adapted to be used with Chinese n-gram and plain text corpora. Text neologism profiles, generated from the lexeme database, are introduced as an innovative approach to emphasize lexicalization in automated dating. Accuracy of dating is increased by the usage of dated proper names and temporal expressions. In an attempt to treat time as a continuous variable, the average year of lexicalization (*AYL*) of words in a given text is also tested as a dating indicator.

It is found that *SLMs* can be successfully employed for assigning chronological categories to Literary Chinese texts. However, neologism profiles prove more robust against the rigidness of the written language, require less specific training, and can easily be combined with and aid the work of a philologist. Nevertheless, some Classical texts remain resistant to a linguistic analysis. All three evaluated methods can be tested through a ready-to-use online tool, *VisualTime*, developed by the author as part of this study (https://visualtime.schalmey.de).