Lexeme-based computational dating approaches for Literary Chinese Texts

Tilman Schalmey*[1]

[1]Trier University – Germany

Abstract

Literary Chinese is often used as a term to refer to the written form of Chinese, reaching from the earliest literature such as the *Shijing* and *Shujing* until the beginning of the 20th century. Aggravated by the continuity of Chinese writing, the imperial examination and other factors, it appears that the pace of language change is somewhat slower – especially for the written language – than for languages with purely phonetic writing systems.
While observations regarding changes in grammar, vocabulary, word use and spelling are elsewhere used for linguistic dating of texts, the long-lasting and rigid tradition of Literary Chinese obstructs pinpointing in which century a perceived text was originally written, especially when relying on later or normalized digital editions.

Employing a diachronic word database created from the source quotations in the *Hanyu Da Cidian* and enriched with metadata and findings from other corpora, I experiment with methods from the field of computational linguistics to solve these issues. It is found that innovative techniques which allow us to visualize lexemes, names and temporal expressions found in a given text as a *temporal profile* may outperform mathematically more complex solutions bases on *Statistical Language Models* that have been proven to be effective and successfully employed for the precise dating of texts in many European languages. Moreover, a database-driven approach can be applied even when suitable diachronic corpora, needed for the training of statistical methods, are unavailable.

I will present my lexeme based dating approach, some conclusions and the resulting software, developed as part of my dissertation project. It aims to assist the arduous work of philologists dealing with the temporal classification of texts and may also help to address issues of forgery.

The data produced during this work also sheds new light on the history of the development of written Chinese, especially on the expansion of the lexicon and the rise of polysyllabic words.

Tilman Schalmey, M.A. studied Sinology, Economics and Scandinavian literature in Munich and Hangzhou. He taught Chinese language and linguistics at Trier university and works as a software developer and IT project manager for a Munich-based lighting design company.

*Speaker