The Role of AI in Philological Research A Case Study of a Bilingual Manchu–Classical Chinese Database

Shuning Zhang
University of Tsukuba, Graduate School of Business Sciences,
Humanities and Social Sciences

This paper presents a pilot study on the use of artificial intelligence in the analysis of historical parallel texts, using as a case study the construction of a bilingual Manchu–classical Chinese research database based on the *Complete Genealogies of the Clans and Families of the Manchu Eight Banners* (八旗滿洲氏族通譜), a major Qing dynasty (17th–19th century) genealogical compilation. Using this project as a case study, the paper asks how AI can support philological research without replacing the careful reading, source criticism, and interpretive work carried out by scholars.

The project builds on the bilingual nature of the *Wuyingdian* edition, in which the Manchu and classical Chinese texts form a closely aligned parallel corpus, allowing for cross-lingual linking and comparison. Although the two versions convey equivalent content, they are used differently in this study because of their linguistic characteristics. The Manchu text, with its relatively regular grammatical structure, is more suitable for structured data extraction, while the classical Chinese text provides a more accessible basis for working with narrative and descriptive passages. This distinction also shapes how AI is applied in the project.

In this study, AI is treated as an assistive method that must remain under human control. Its use is explored across three stages of database construction: transcription, annotation, and retrieval. First, OCR-based transcription tools (such as Transkribus or eScriptorium) can greatly improve the speed of converting manuscript materials into machine-readable text. However, recognition errors are unavoidable, especially for low-resource scripts such as Manchu, and therefore require careful manual correction. Second, semi-automatic annotation of structured data depends on a sufficient amount of manually prepared training data; without this, automated methods cannot produce reliable results. Third, retrieval-based approaches can improve access to narrative and contextual information, but the outputs generated by large language models are not always accurate and must be checked carefully.

By combining these approaches, the project proposes a hybrid retrieval framework within the database that integrates structured queries with text-based search. More importantly, this paper argues that AI should be understood as a supporting tool in philological research rather than an independent analytical method. While AI can improve efficiency and make sources more accessible, it cannot replace the need for careful interpretation and human judgment. Its use therefore requires continuous attention, an awareness of its limitations, and a careful balance between technological possibilities and established research practices. In this way, this paper contributes to ongoing discussions on the responsible use of AI in the study of historical parallel corpora.