Resources and Tools for Corpus Compilation of Translated Literary Texts in Late Qing and Republican Period

Nason Anran Cao
The Chinese University of Hong Kong

Abstract

Corpus-based translation studies have increasingly attracted attention in recent years. Translated literature of the late Qing and Republican Era in China is one popular topic of research in translation studies. Building up a corpus of translated literature from this period may provide a stronger empirical basis for investigation. This article introduces online resources, including a variety of databases and tools, with a focus on the Chinese OCR software used in corpus compilation. Features of databases that contain translated literary texts of late Qing and Republican Period are compared, problems of OCR recognition for these texts are identified, and four OCR software packages are tested using sample texts. This article provides insights into large-scale digitization projects on Chinese texts as well.

Keywords

Chinese databases, corpus-based translation studies, Chinese OCR, digital humanities, literary translation