



# On Building a Bilingual Terminology Corpus from the Perspective of Human-machine Collaborative Translation

Jiameng Wei, Yi Li\*

School of Foreign Language, Wuhan Business University, Wuhan 430118, Hubei, China.

**How to cite this paper:** Jiameng Wei, Yi Li. (2025). On Building a Bilingual Terminology Corpus from the Perspective of Human-machine Collaborative Translation. *The Educational Review, USA*, 9(7), 658-663.  
DOI: 10.26855/er.2025.07.005

**Received:** June 23, 2025

**Accepted:** July 20, 2025

**Published:** August 18, 2025

**Corresponding author:** Yi Li, School of Foreign Language, Wuhan Business University, Wuhan 430118, Hubei, China.

## Abstract

The rapid advancements in artificial intelligence and large language models have significantly improved the quality of machine translation, profoundly influencing not only professional translation workflows but also driving pedagogical innovation in translation teaching at the higher education level. However, the ongoing issue of terminological and conceptual inaccuracies in current machine translation systems highlights the need for further refinement. Given the centrality of terminology in academic discourse, ensuring accuracy in terminological translation is essential for facilitating effective scholarly communication. This paper seeks to propose a methodological framework for the systematic extraction of bilingual terminological data to support the development of specialized corpora, with a particular focus on the social sciences and humanities. The primary aim is to maintain terminological precision and conceptual consistency while ensuring strong contextual alignment. Additionally, the paper aims to design an innovative teaching model to equip translation students with the technical skills required to create discipline-specific bilingual terminology databases, addressing a critical competency gap in the contemporary language services industry.

## Keywords

Machine translation; Translation of terminology; CAT; Corpus

## 1. Introduction

In the context of informatization and digitalization, translation pedagogy has been presented with new challenges and expectations. To meet the evolving demands of the language services industry, both teachers and students must fully embrace emerging technological advancements and adapt to the shifting dynamics of the modern era. The rapid progress of artificial intelligence has led to significant enhancements in the quality of machine translation. As a result, “language service providers and clients generally maintain an optimistic outlook on the future of machine translation, recognizing its substantial potential to enhance translation efficiency” (Xing et al., 2023).

In response to these developments, teachers of colleges and universities should adopt technology-driven approaches to translation courses, equipping students with the skills needed to navigate the increasingly prevalent “machine translation + post-editing” workflow that dominates the current market.

## 2. Literature Review

The data source for this section of literature is the China National Knowledge Infrastructure (CNKI) database, where the keywords “translation technology” and “teaching” were used to search for CSSCI source journals (the last data update was on February 3, 2025), resulting in a preliminary total of 173 relevant articles. After deleting unrelated literature such as conference papers, book reviews, and interviews, the final confirmed total number of selected articles is 136.

Among them, the earliest article focusing on translation technology education appeared in 2007, titled “Integrating CAT Technology into Translation Teaching” (Li & Ma, 2007), which called for attention to teaching technological tools in translation courses to prepare competent trainees to meet the requirements of the new age. It is evident that since the 21st century, the teaching research of translation technology has gradually become a topic of concern in China’s translation academia. With the continuous iteration and upgrading of technology, academia is increasingly emphasizing research on translation technology education. This is reflected in the publication trend: from 2007 to 2018, only a handful of articles discussing the emphasis on translation technology in translation teaching, averaging only about 5 articles per year. However, since 2019, the attention in this field has significantly increased, with 14 related articles published in that year alone. Subsequently, until 2024, more than 10 papers each year have focused on this topic. This trend clearly reflects that with the vigorous development of information technology, translation education is undergoing a technology-driven transformation.

From the perspective of research content, domestic literature focusing on translation technology teaching can be divided into two main categories: first, translation technology research based on the language service industry, typically investigating the current demand in the language service market through social recruitment announcements, industry data analysis, and language service market surveys. The focus of these articles includes the gap between the curriculum or model for cultivating students of the translation major to meet the requirements of the language service industry (Yu & Wang, 2023). Second, research on translation technology teaching models and effects, mainly focusing on aspects such as the models, methods, and effects of translation technology teaching (Li & Jiang, 2024). These articles discussed curriculum design for translation technology teaching, the development of teaching resources, and the faculty. In addition, research on translation technology teaching not only focuses on the technological shift in teaching and training of students’ competence, but also discusses the related ethical issues in the background of new technologies.

In summary, research on translation technology in teaching has become a key topic in China’s translation academia. The existing literature has addressed the innovative aspects of translation technology education models, including curriculum positioning, design principles, curriculum content, and teaching methods. These research outcomes contribute to cultivating students of the translation major with technological competence.

## 3. Problem Formulation

The “human-machine collaboration + post-editing” model has emerged and gradually become the mainstream translation model to meet the current demands of the language service market. Given this development trend, translation teaching also needs to keep pace with the times, focusing on adapting students to the workflow of “machine translation + post-editing”. This paper discovers one of the typical problems in machine translation from a real translation project, which was carried out in a human-machine collaborative workflow.

### 3.1 Task Description

The translation project involves a book on the political, economic, military, and cultural systems of the Tangut people in Yuan Dynasty literature. This translation project was carried out on “TwinsLSP”, a translation platform designed by Transn Company (<http://pe-x.iol8.com/>). The first step was to create a translation project on the “TwinsLSP” platform. The second step was to select the machine translation engine and then use machine translation to complete the initial translation of the source text. For this translation task, the ChatGPT engine was chosen for the initial translation of the original text. The third step was post-editing, a human translator meticulously revised and polished the machine-generated target text.

### 3.2 Feedback

In the post-editing process, the accuracy and consistency of terminology translation posed a significant challenge. Specifically, this problem manifested as inconsistencies in the translation of the same term across different versions or mistranslations resulting from misunderstandings of the term's meaning. Terminologies are the specialized terms and vocabularies used to define specialized knowledge in a certain domain. Terminology has a series of distinct characteristics, including established conventions, adherence to rigorous scientific standards, substantial reasoning, and precise expression. Clearly, ensuring that terminology and its concepts are clear and unambiguous is essential. Therefore, in this project, a significant amount of effort was spent on revising the translation of terms and concepts in the target text during the post-editing process to ensure accuracy and coherence.

The source text of this translation project is a monograph on the “Tangut”, the remnants of the Western Xia during the Yuan Dynasty. The remnants, Tangut, were called “Dangxiang” in pinyin, which is a phonetic transcription of the tribe's name in Qiang or Tibetan languages by historians of the Sui and Tang Dynasties, derived from the Turkic and Mongolic terms in the Altaic language family. In the source text, “Dangxiang” was used to refer to the Western Xia remnants.

In Chinese historical texts, “Tangut” was the name for the Western Xia remnants in Chinese literature. Furthermore, the academic community generally translates the term “Tangut people” as “Tangut” in the Altaic language family. In the post-editing process, it can be observed that there are instances of mistranslation of Tangut, such as “Tangwu people”, “Dangxiang”, and “Wu people of the Tang Dynasty”. Consistency in the translation of terminology is crucial when translating academic monographs. In this project, the proofreading and unification of terminology, such as Tangut, still rely on manual post-editing.

In addition to proper nouns, this translation project also involves translating official titles and administrative institution names from ancient China, particularly from the Yuan Dynasty. The translation of such terms requires a thorough understanding of the ancient Chinese bureaucratic system and careful consideration of terms and names that are easily understood by foreign readers. Clearly, machine translation cannot correctly render official titles and administrative institution names from ancient China, which largely depend on manual post-editing, such as:

The political system of the Yuan Dynasty was characterized by a dual-institution system at the central level, namely the Central Secretariat, the Bureau of Military Affairs, and the Censorate, which held the main power of the central institutions. Additionally, the Yuan Dynasty implemented a provincial system, with provinces divided into routes, prefectures (or departments), and counties, with routes managed by the provinces (Hucker, 2008). We found that machine translation can currently correctly translate Central Secretariat (Zhong Shu Sheng), Bureau of Military Affairs (Shumi Yuan), and Censorate (Yushi Tai), but the translation of the term for the administrative office under the Central Secretariat is generally done through phonetic transliteration. Other levels of administrative units all rely on manual verification and proofreading. This indicates that terminology translation needs to accurately convey meanings and requires professional verification; currently, machine translation still relies on manual proofreading for terminology translation.

Evidently, in this translation project, the translation terms and concepts, and ensuring their consistency, depend almost entirely on manual post-editing.

### 3.3 Findings

In recent years, machine translation has achieved rapid development due to the emergence of large language models. However, at this stage, machine translation still has significant defects in translating terms and concepts, and mistranslations frequently occur. As a key carrier of knowledge inheritance, the importance of terminology translation is self-evident when it crosses the boundaries of different languages and realizes interdisciplinary communication. Translators are among the professionals who spend the most time researching and managing terminology (Nkwenti-Azeh, 2001). The terminology translation involves the transfer of ideas, where different languages, traditions, and ideas interact and collide directly, and where the transformation of meaning and the generation of new meanings also occur. Even though translators have already mastered their working languages proficiently, they may not also possess certain knowledge in a specific domain, nor does it guarantee that they are acquainted with the terms and concepts of every domain within their working languages. Traditionally, translators refer to dictionaries or professional references when they come across unfamiliar terminology to avoid errors in translation, and this “terminology

tasks can take up to 60% of the translator's time" (Garcia-Santiago & Diaz-Milon, 2024). However, this process is time-consuming. Nowadays, thanks to the advancement of machine translation technology, the efficiency of translation has been greatly improved.

It is important to emphasize that the work of terminology unification is an indispensable part of the pre-translation preparation phase, and its effective implementation can greatly alleviate the burden on translators during the post-editing process.

## 4. Discussion

Since OpenAI launched ChatGPT, a large language model, its impact on the translation field cannot be underestimated. In current translation practices, although machine translation relying on large language models has not yet achieved ideal accuracy in terminology translation, it is undeniable that the emergence of large language models has provided strong support for the construction of bilingual terminology corpora. However, translating terms and concepts mainly relies on post manual proofreading. Translators should maintain a high degree of self-awareness and critical thinking during the translation process to ensure the fairness and objectivity of the translation. Therefore, in the translation of humanities and social sciences terms, it is urgent to build a bilingual terminology corpus for optimization and expansion.

Therefore, this paper argues that the translation preparation phase, especially for academic texts, should first focus on constructing bilingual terminology databases and then apply them to tools such as computer-assisted translation (CAT) software to assist translators in their work.

### 4.1 Methodology for a Corpus-assisted Approach to Terminology Management

Remarkable accomplishments have been attained in research concerning the development and utilization of bilingual glossaries. For instance, during the data collection stage, data has been comprehensively amassed from diverse sources, including academic publications, specialized tomes, and industry dispatches. This endeavor aims to guarantee that the collected data is not only extensive but also highly specialized, thereby establishing a robust groundwork for the precise and efficient advancement of subsequent translation processes. Currently, terminology extraction technology is also continuously evolving, from traditional rule-based and statistical methods to the use of machine learning and deep learning algorithms, which can more accurately identify and extract terms from large-scale corpora. Advances in alignment technology have enabled bilingual terms to achieve more accurate matching at both semantic and grammatical levels, improving the quality and practicality of the glossary.

During the data collection phase, data is gathered from multiple levels and channels to ensure the comprehensiveness and accuracy of the terminology database (Lefever & Terry, 2024). The corpus can be obtained from three channels:

Firstly, the corpus should encompass bilingual abstracts sourced from journals in the social sciences as well as parallel texts drawn from seminal scholarly works. Given the authoritative status and methodological rigor of these materials, they serve as a reliable repository of high-quality, standardized terminological resources.

Secondly, the inclusion of transcripts from academic conference interpreting events and corresponding versions of research publications is essential. These texts not only enrich the diversity of the terminology database but also mirror current academic developments and emerging research trends.

Thirdly, attention should be directed toward the discussion of new concepts within scholarly social media platforms. This type of data, with its timeliness and innovativeness, allows us to identify newly coined terms in a timely manner.

Following this, the collected terminological data will undergo processing and organization via annotation techniques. In order to manage the terminology, we propose a new pattern:

First, defining the disciplinary domain of each term using information retrieved from large language models. This process supports efficient categorization and retrieval of terminology while equipping translators with relevant disciplinary background knowledge. Second, identifying the usage of terms across different contexts, such as academic and public discourses. The distinction helps translators select appropriate equivalents based on specific contexts. Third, tracing the evolution of terminology, including temporal shifts in usage over time. This allows translators to understand the historical evolution of terms and provides guidance in selecting standardized translations of terms.

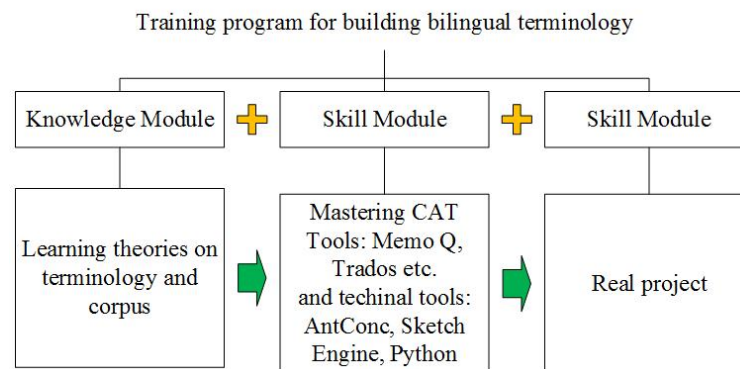
At present, large language models are capable of being harnessed for collaborative undertakings. For instance,

when annotating the terms' contextual scenarios, large language models can generate context relevant to these terms. With the contexts generated by large language models, translators can make well-informed choices.

Finally, after completing the collection of the data for the corpus, the next step involves text cleaning (Garcia-Santiago & Diaz-Milon, 2024). Based on the aforementioned corpus collection procedures, which encompass techniques like online gathering, manual entry, and scanning recognition, the retrieved text is likely to include non-standard symbols, inconsistent formats, and the like. To prevent any adverse impact on retrieval outcomes, this text necessitates cleaning.

## 4.2 Framework for Training Students' Competence in Building a Bilingual Terminology Corpus

The core of building a bilingual terminology corpus lies in training students who possess both linguistic and computer literacy. The students should not only be proficient in relevant technologies but also enhance their terminology management capabilities comprehensively with the assistance of artificial intelligence tools. This includes a deep understanding of the basic concepts of terminology management, mastery of the terminology management process, and proficient use of common terminology management tools. In light of this, this paper will propose a training program for a bilingual terminology corpus, divided into three parts: knowledge module, skills module, and practice module, as illustrated in Figure 1.



**Figure 1. Training program for building bilingual terminology.**

The knowledge module primarily involves terminology knowledge and corpus linguistics knowledge. The skill module includes familiarity with technical tools such as AntConc, Sketch Engine, Python, and CAT tools like MemoQ and Trados. In the practice module, it will meet industry needs, undertake real terminology standardization, patent document terminology alignment, and other projects, or develop a teaching case library to cover typical terminology translation application scenarios. Based on these three modules, compound terminology engineering talents with both linguistic literacy and technical implementation capabilities can be cultivated. Through this path, we aim to cultivate professionals who have high-level terminology processing capabilities, are familiar with the application of technical tools, and can effectively integrate into the human-machine collaborative translation process.

## 5. Conclusion

The technology-empowered training model for students of the translation major has gained a consensus within the language service industry. Through translation project experiences, it is revealed that current machine translation still falls short in the accuracy of translating terms and concepts, especially in humanities and social science texts. This paper proposes a translation preparation work centered on the building of a terminology corpus and actively investigates methodologies for building a bilingual terminology corpus. Furthermore, this paper also envisions that with the gradual accumulation of the terminology corpus, future exploration will further refine the approach of machine translation to elevate the quality of terminology translation. This approach can effectively tackle problems such as inaccurate terminology translation and insufficient contextual understanding encountered by machine translation in specific fields.

---

## Funding

This paper was supported by the research project from Wuhan Business University (Grant No. 2023N013) and the College Students' Innovative Entrepreneurial Training Plan Program of Wuhan Business University (Grant No. 202411654162).

## References

- Garcia-Santiago, L., & Diaz-Milon, M. (2024). Pedagogical and communicative resilience before Industry 4.0 in higher education in translation and interpreting in the twenty-first century. *Education and Information Technologies*, 29, 495-515.
- Hucker, C. O. (2008). A dictionary of official titles in imperial China. Beijing University Press.
- Knight, D., Fitzpatrick, T., et al. (2023). Corpus to curriculum: Developing wordlists for adult learners of Welsh. *Applied Corpus Linguistics*, 3, 1-9.
- Lefever, E., & Terryn, A. R. (2024). Computational terminology. In Y. Pan, H. Huihui, & D. Li (Eds.), *New advances in translation technology: Applications and pedagogy* (pp. 141-159). Springer.
- Li, S., & Ma, L. (2007). Integrating CAT technology into translation teaching. *Foreign Language World*, 3, 35-43.
- Li, X., Pu, L., & Jiang, H. (2024). Study on the construction of a human-machine collaborative translation teaching model empowered by technology. *Journal of the Foreign Language World*, 3, 43-50.
- Nkweni-Azeh, B. (2001). Term banks. In M. Baker (Ed.), *Routledge encyclopedia of translation studies* (pp. 249-251). Routledge.
- Xing, Y., Zhang, Y. M., Du, N., et al. (2023). 2023 China translation and language service industry development report. Translators Association of China. Retrieved July 17, 2025, from [http://www.tac-online.org.cn/2023-10/18/content\\_42940430.html](http://www.tac-online.org.cn/2023-10/18/content_42940430.html)
- Yu, Y., & Wang, X. (2023). An analysis on machine translation post-editing competence and its cultivation from the perspective of categorization theory. *Foreign Language Education*, 1, 90-96.