Academic Education Publishing House

Philosophy and Social Science Vol. 1 No. 10, 2024

A Study on the Construction of a Large-Scale Multimodal Parallel Corpus of Typical Official Documents of China

Xia Liang, Hua Zhang*

Public Education Department, Jinan Vocational College of Nursing, Jinan, Shandong, China

**Corresponding Author.*

Abstract: Corpora are essential resources for language research, playing a significant role in linguistics, translation studies, and interdisciplinary research. However, current international corpus construction primarily focuses on European and American languages, with insufficient attention to Chinese. Despite notable progress domestically, gaps remain in the development of multimodal bilingual parallel corpora. This study proposes a research plan guided by to construct a large-scale diachronic multimodal bilingual parallel corpus. The corpus will include key official speeches, governance documents, and multimodal data (texts, audio, videos, etc.), with detailed metadata annotation and high-quality bilingual alignment. It aims to support complex retrieval functions and interdisciplinary applications, providing theoretical and data support for linguistics, translation teaching, natural language processing, and international communication. By integrating theories of parallel and multimodal corpora, the study emphasizes corpus balance and representativeness and leverages advanced retrieval platforms for diversified functionality. The corpus demonstrates extensive application value in language teaching, translation practice, and international communication, promoting the standardization and global dissemination of Chinese-specific terms and supporting cultural research. As data expands and technologies improve, the corpus will further strengthen China's cultural confidence and international discourse power in the new era.

Keywords: Multimodal Corpus; Typical Official Texts of China; Language Comparison Studies; Translation Studies; International Communication

1. Introduction

Corpora are the cornerstone of language research. Since the mid-20th century, they have played an increasingly important role in linguistics, translation studies, computational linguistics, and related disciplines. However, research on international parallel corpora has long been concentrated on European and American languages, with relatively little focus on non-Western languages, particularly Chinese. In contrast, while significant progress has been made in corpus construction in China, there are still clear deficiencies in the development of multimodal and bilingual parallel corpora. [1] This situation has to some extent hindered the construction of China's discourse power in international language research, falling short of the country's status and needs as a cultural powerhouse.

To better respond to the development demands of typical political texts of China, enhance China's cultural confidence, and promote the construction of a great power discourse system, it is imperative to develop a large-scale, systematic, and era-specific multimodal bilingual parallel corpus. The authors propose a plan to construct a diachronic multimodal bilingual parallel corpus of typical political texts of China containing tens of millions of words. This corpus will not only include traditional text resources but will also integrate multimodal data such as images, audio, and video, achieving comprehensive bilingual alignment and multidimensional retrieval functions. [2] Its construction will provide foundational support for natural language processing, translation teaching and practice, bilingual dictionary compilation, and the study and dissemination of China's distinctive discourse in the new era.

2. Current Status and Development Trends of Domestic and International Research

In the field of international corpus research,

European and American scholars have taken the lead and achieved significant results, especially in the construction of parallel corpora. Landmark corpora, such as the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA), serve as classic resources that provide rich data and robust technical support for studies in linguistics, semantics, discourse analysis, and translation studies. Parallel corpus research has also expanded into cross-linguistic comparison, with resources like the English-Norwegian and English-Swedish parallel corpora developed by the University of Oslo, advancing comparative studies across different linguistic systems. However, international corpus research still primarily focuses on European and American languages, with scarce resources available for Chinese parallel corpora. The existing corpora are also relatively limited in scope, which restricts the depth and breadth of bilingual comparative research. [3]

Domestically, although corpus research started later, it has developed rapidly in recent years. Numerous universities and research institutions have invested substantial resources in constructing corpora. For example, the CCL Corpus from Beijing Foreign Studies University and the BBC Corpus from Beijing Language and Culture University have significant applications in the study of modern Chinese and English. Furthermore, specialized parallel corpora have emerged for specific research domains, such as the Shakespearean English-Chinese Parallel Corpus from Shanghai Jiao Tong University, the Dream of the Red Chamber Parallel Corpus from Yanshan University, and the Lu Xun Novels Chinese-English Parallel Corpus from Shaoxing University. [4] These corpora provide valuable data support for linguistic research, literary translation practice, and education. Nevertheless, compared to the international advanced level, domestic corpora still lag in integrating multimodal data, bilingual alignment capabilities, and dynamic updating mechanisms. Notably, there is a marked gap in the availability of comprehensive corpora reflecting the distinctive features of typical political texts of China.

3. Research Objectives and Theoretical Framework

In response to the aforementioned gaps, the research team aims to construct a large-scale diachronic multimodal bilingual parallel corpus of typical political texts of China. This corpus focuses on three core tasks: resource collection, platform development, and interdisciplinary application. [5]

For resource collection, the corpus will include key speeches, governance documents, and diachronic materials. The data will encompass diverse formats—text, audio, images, and videos—across multiple languages to enhance representativeness and diversity.

In terms of platform development, the corpus will leverage advanced online technologies to design sophisticated retrieval features, supporting tools such as translation assistance and keyword analysis to provide researchers with convenient data services.

For interdisciplinary applications, the corpus will serve as a critical data platform for studying China's international discourse power in the new era, offering theoretical and empirical support for fields like linguistics, translation studies, political science, and sociology.

From a theoretical perspective, the research team integrates theories from parallel and multimodal corpora, emphasizing balance and representativeness to ensure extensive coverage in registers, timelines, and languages. Detailed metadata annotation—covering language, era, stylistic features, and translation direction—will enhance retrieval and application efficiency. Additionally, the corpus will incorporate linguistic annotation theories, providing fine-grained part-of-speech and grammatical annotations to support linguistic research and natural language processing technologies.

This corpus will not only provide robust data support for translation teaching, practice, and bilingual dictionary compilation but will also play a pivotal role in developing a Chinese-specific international communication capacity and discourse system.

4. Construction Process of the Bilingual Parallel Corpus

Building a bilingual multimodal parallel corpus is a complex and systematic project encompassing key stages such as resource collection, data processing, and platform development. This corpus will offer rich

empirical evidence for language research and support translation practice and cross-cultural communication.

(1) Resource Collection

Resource collection is fundamental to corpus construction, directly determining the diversity and quality of data. The research team adopts a diachronic collection approach, beginning with the 2012 speech made by the new Politburo Standing Committee to domestic and foreign media. The corpus will systematically include key speeches by the General Secretary, governance documents, and official work reports, covering various languages and registers to ensure diversity and representativeness. [6]

To reflect the multimodal nature of the corpus, the team collects corresponding audio, video, and image data. Non-textual data is converted into structured textual data through speech transcription and optical character recognition (OCR) technologies. This process enriches the corpus content and format, enabling broader research possibilities.

(2) Data Processing

In the data processing stage, the team employs a metadata design approach to meticulously annotate the corpus. Metadata includes language type, timeframe, stylistic features, and translation direction, providing precise reference dimensions for classification, retrieval, and subsequent analysis.

For bilingual data alignment, sentence alignment techniques combining algorithms and manual proofreading are employed to ensure high-quality alignment. For non-textual data, the team further refines speech transcription and OCR techniques to systematically transform multimodal information into structured data. This process enhances the compatibility and usability of the corpus, making multimodal data applications more intuitive and efficient. [7]

To meet diverse storage and retrieval needs, the data is saved in XML and JSON formats, facilitating database import and adaptability to various retrieval scenarios.

(3) Platform Development

Platform development is the core of the corpus's practical application, aiming to provide users with efficient and convenient retrieval and analysis tools. The team uses PHP and MySQL technologies to build the retrieval platform, ensuring scalability and stability for handling large-scale data and user interactions.

The platform's functional modules include word list generation, frequency statistics, keyword indexing, and full-text search.

Word List Generation: Automatically extracts and generates bilingual word lists, offering direct language learning resources.

Frequency Statistics: Reflects vocabulary usage frequency and distribution patterns, providing quantitative insights for language use research.

Keyword Indexing: Pinpoints keyword occurrences within the corpus, enabling users to locate critical information quickly.

Full-Text Search: Allows users to conduct in-depth searches across the entire corpus using keywords or phrases.

The platform also integrates translation assistance tools, such as automatic translation suggestions and term equivalence table exports, improving the efficiency and accuracy of translation tasks.

5. Application Research Based on the Corpus

The bilingual parallel multimodal corpus plays a crucial role in studying and applying the discourse power of typical political texts of China. By systematically collecting key speeches by the General Secretary, governance documents, and official work reports, the corpus integrates a variety of languages, registers, and multimodal data such as audio, video, and images. This comprehensive and multidimensional perspective enables the corpus to accurately and timely reflect the developmental trajectory of typical political texts of China, providing robust data support and theoretical foundations for research in linguistics, translation studies, political science, sociology, and other disciplines. [8]

At the practical level, the corpus significantly advances the study of core vocabulary and terms associated with typical political texts of China. By systematically analyzing and standardizing these terms, the corpus establishes a standardized linguistic framework for bilingual translation and international communication, improving the accuracy and consistency of international discourse. Additionally, the corpus serves as a foundation for synchronic and diachronic studies of language phenomena in the new era.

Synchronic studies reveal the linguistic features of different languages within a specific timeframe, while diachronic studies trace language changes over time in response to social development. For instance, researchers can use the corpus to analyze the frequency and semantic shifts of specific terms across historical periods, uncovering patterns of linguistic evolution and the societal impacts on language. [9]

The corpus also demonstrates unique advantages in policy research and sociolinguistics. It provides data support for national policy-making and studies on language dissemination patterns, particularly in constructing an international communication framework. By exploring the connotations and values of Chinese cultural traditions, the corpus contributes to building a discourse system. For example, local media have creatively integrated traditional cultural elements into innovative programming formats, providing successful cases for promoting Chinese culture internationally. These practices underscore the corpus's dual utility for academic research and as a resource for integrating culture and language in international communication. [10]

In the fields of language education and translation practice, the corpus offers extensive support. It provides language learners with authentic contexts for understanding linguistic structures and expressions while serving as a comprehensive reference for translation teaching and practice. Initiatives like "ideological and political education in the classroom" have facilitated the integration of ideological education into language instruction. The corpus's translation assistance tools further enhance the efficiency and quality of translation work, injecting new vitality into multilingual education and training. As a vital component of national think tanks, the corpus provides scientific backing and data support for policy-making, strategic planning, and international communication capacity-building, highlighting its strategic value in national development.

6. Conclusion

The corpus of typical political texts of China is not only a critical tool for academic research but also a vital resource for international communication and cultural exchange. Its multimodal features and dynamic updating mechanism offer comprehensive support for linguistics, cultural dissemination, policy-making, and international collaboration.

With advancements in technology and the enrichment of data resources, the application potential of the corpus will continue to expand, providing stronger support for constructing and disseminating a discourse system. Through the collection and analysis of multimodal data, the development of advanced retrieval platforms, and interdisciplinary applications, the corpus will become a cornerstone infrastructure for studying discourse in the new era. As the corpus evolves and optimizes, its application value in academic research, language teaching, and policy-making will further enhance, making significant contributions to China’s cultural confidence and international discourse power.

Acknowledgments

This paper is supported by the National College Foreign Language Teaching and Research Project (No. 2023SD0031).

References

[1] Construction and Application of the Multilingual Parallel Corpus for "The Governance of China" by Li Xiaoqian, Hu Kaibao. Foreign Language Audio-Visual Teaching, 2021.

[2] Reform of Translation Teaching in the Era of Technology: Exploration of Translation Professional Corpus Construction by Chai Mingjing, Wang Jing. Foreign Language Audio-Visual Teaching, 2017.

[3] Research on the Construction of Bilingual Parallel Corpus-Based Translation Teaching Platform and Its Pedagogical Model by Ge Lingling, Li Guangwei, Wang Yiming. Foreign Language World, 2015.

[4] Research on English-Chinese Bilingual Parallel Corpus-Based Translation Teaching Model by Xiong Bing. Foreign Language World, 2015.

[5] The Maturing of Constructivist Instructional Design: Some Basic Principles That Can Guide Practice. Educational Technology, 2000.

[6] Language Contact Through Translation by Shuangzi Pang, Kefei Wang. Target: International Journal of Translation

Studies, 2020.
[7] Corpus and Bilingual Comparative Studies by Qin Hongwu, Kong Lei. Foreign Language Teaching and Research Press, 2019.
[8] Introduction to Corpus Translation Studies by Hu Kaibao. Shanghai Jiao Tong University Press, 2011.
[9] Corpus Application Tutorial by Liang Maocheng, Li Wenzhong, Xu Jiajin. Foreign Language Teaching and Research Press, 2010.
[10] Text and Technology: In Honor of John Sinclair by Mona Baker. Amsterdam-Philadelphia, 1993.