US-China Foreign Language, July 2026, Vol. 24, No. 7, 243-251
doi:10.17265/1539-8080/2026.07.001

Research on the Construction and Application of English-Chinese Interpreting Corpus Integrated With AI Speech Recognition

CHEN Shengbai
Hunan First Normal University, Changsha, China
ZHANG Yingying
Guangxi University, Nanning, China

Compared with other types of corpus, the current research on interpretation corpus lags behind due to its particularities. This paper discusses the construction of English-Chinese (E-C) interpreting corpus integrated with artificial intelligence (AI) speech recognition technology, so as to propose its construction steps and processes. And it further studies the prospects of its applications which includes its functions from three aspects: intelligent platform for interpretation, intelligent interpretation teaching, and quality assessment, providing new objects, new perspectives, and new approaches for the study of Chinese interpreting corpus.

Keywords: corpus, interpretation teaching, intelligent platform for interpretation, quality assessment in interpreting, AI speech recognition

Introduction

Corpora are being increasingly widely applied in humanities and social sciences. The construction of various learner corpora has become a prominent research focus in foreign language learning, applied linguistics and foreign language teaching at home and abroad in recent years. Nevertheless, research on and the development of interpreting corpora still fall short of practical demands. Specialized corpora tailored for learner interpreting training and relevant applications remain scarce. Although numerous scholars, universities, and research institutions have continuously developed and innovated interpreting corpora, corpus construction in this field faces considerable obstacles due to the distinctive sources of interpreting data, and relevant research progresses relatively slowly.

Artificial intelligence (AI) technology has achieved significant progress at present, and AI-assisted translation demonstrates tremendous application potential within the language service industry (W. Wang & B. Wang, 2025). The construction of interpreting corpora and associated research will undoubtedly constitute a core dimension for the future advancement of interpreting teaching and research (Zhao, 2012). While promising

Acknowledgements: This project was supported by funds from National Social and Science Fund Project: Research on the Ecological Mechanism of Family-School Co-education of English Core Competency in Basic Education (Grant No.: 23BYY150).
CHEN Shengbai (Corresponding author), Ph.D., professor, Department of Foreign Languages, Hunan First Normal University, Changsha, China.
ZHANG Yingying, M.A., Department of Foreign Languages, Guangxi University, Nanning, China.

progress has been achieved in developing translation corpora, such resources lack flexibility and diversity in diversified practical applications. The development of interpreting corpora calls for abundant thematic investigations and broad research perspectives. Starting from the construction of interpreting corpora, this study integrates AI speech recognition technology to build a corpus supporting multifaceted applications. It aims to facilitate self-directed training for interpreting learners and offer new approaches for interpreting instruction and interpreting quality assessment. By enriching research on interpreting corpora, this study also promotes the refinement of corpus construction practices and gradually elevates the research quality in this domain. In the new era featuring rapid advances in big data and AI, in-depth research on corpus studies and the development of different corpora can reach new heights leveraging high-speed computer technologies, thereby providing new perspectives and paradigms for interpreting research.

Construction of a Multifunctionally Applicable Interpreting Corpus

Research Background

Famous overseas interpreting corpora include the English-Japanese simultaneous interpreting corpus developed by Nagoya University of Japan, and the European Parliament Interpreting Corpus (EPIC) constructed by the University of Bologna, Italy. Domestically, the Chinese-English Conference Interpreting Corpus (CECIC), developed by Shanghai Jiao Tong University, has entered the phase of research and application. The Parallel Corpus of Chinese EFL Learners (PACCEL) is also influential. Another well-known resource is the Chinese-English Consecutive Interpreting Corpus of Premier's Two Sessions Press Conferences (CEIPPC), which mainly collects simultaneous interpreting recordings from major conferences and press briefings. A review of existing domestic interpreting corpora reveals that large-scale ones remain limited in quantity with insufficient corpus size. At the current stage, these corpora are mostly deployed for narrow academic research, while widely applicable interpreting corpora are still in short supply. Research on interpreting corpora remains underdeveloped and calls for further exploration.

In recent years, rapid advances in AI and computer technologies, alongside the thriving interpreting industry, have laid relatively mature foundations for constructing interpreting corpora of considerable scale. Latest literature indicates that the West boasts no fewer than ten moderately large interpreting corpora, among which the EPIC stands as the most representative example (Wang & Fu, 2020). Against this backdrop, there are promising prospects for China to build large-scale interpreting corpora. Global progress in corpus development offers vital support for domestic research on interpreting corpus construction.

Construction Approaches

In terms of corpus design, it is first necessary to determine the overall architecture and functional modules of the corpus (Liu, 2006). Corpus construction commences with defining the object to be developed and its constituent components (Atkins, Clear, & Ostler, 1992). Based on the design of the corpus architecture, clear construction objectives, target users and expected outcomes should be specified. The priorities for the construction of interpreting corpora and relevant research lie in expanding corpus scale, diversifying corpus types, deepening the level of corpus annotation, improving the accuracy of corpus processing, so as to optimize interpreting corpus development and enhance corpus representativeness (Zhao, 2012). The interpreting corpus constructed in this study aims to facilitate the multifaceted application of interpreting corpora, boost the efficiency and accessibility of interpreting training, and offer new perspectives for interpreting teaching and

interpreting quality assessment.

Centering on interpreting training, the scale of the constructed interpreting corpus strives to maximize. In practical training sessions, AI speech translation technology is adopted: The speech produced by learners during training is recorded via AI and subsequently transcribed into text. The transcribed text is then compared with the corpus resources to identify interpreting errors. Learners often suffer from insufficient vocabulary while facing an enormous number of proper nouns. When encountering translation difficulties in highly specialized fields, such as medicine, law, and biology, the interpreting corpus can provide accurate translations of domain-specific proper nouns. Combined with AI technology, the corpus enables learners to compare their outputs against existing corpus sources to detect inappropriate word choices in training. This alleviates the cognitive load of human interpreting and allows learners to conduct self-evaluation and reflection throughout personalized training. To sum up, the general framework established in this study relies on AI to support speech input and text output, followed by comparative analysis with the interpreting corpus. At present, domestic scholars have carried out extensive explorations concerning corpus construction workflows. Drawing on mainstream construction methodologies, the development process of the interpreting corpus is divided into four stages: corpus collection, corpus annotation, corpus alignment and import, as well as data cleaning and maintenance.

Analysis of AI-assisted speech translation. In recent years, AI technologies in China have achieved rapid progress, driving substantial breakthroughs in speech input and bringing China to a world-leading position in this field. Compared with written translation, interpreting constitutes an immediate, situated communicative activity. Accordingly, multiple obstacles emerge during translation training: learners' unclear pronunciation, environmental noise, excessively fast speaking rates that hinder recognition by speech software, and lengthy speech input. One core objective of interpreting training in this study is to maximize training efficiency for learners, and accurate speech transcription of interpreting output is the primary prerequisite. Therefore, selecting reliable speech translation software constitutes the preparatory work for constructing the speech-recognition-enabled interpreting corpus.

Among mainstream free speech input tools, IFLYTEK Input Method stands out for its high recognition accuracy and widespread popularity. As early as 2010, IFLYTEK launched the world's first intelligent input method supported by cloud computing. Later, IFLYTEK and Harbin Institute of Technology (HIT) jointly released the HIT-iFLYTEK Language Cloud, the world's first Chinese natural language processing cloud service platform, to boost transcription accuracy. Deeply integrated with HIT's Language Technology Platform (LTP), the HIT-IFLYTEK Language Cloud provides a suite of natural language processing technologies, including Chinese word segmentation, part-of-speech tagging, named entity recognition, dependency parsing, and semantic role labeling (Li, 2022). According to official specifications of the IFLYTEK Input Method, it supports a transcription speed of 400 characters per minute with a general AI speech recognition accuracy rate of 98%. It can automatically insert punctuation marks during speech input and deliver real-time mutual speech translation among Chinese, English, Japanese, Korean, Russian, and other languages. In summary, the IFLYTEK speech input system serves as a suitable candidate for this research.

Corpus collection. The primary task of data collection is to identify corpus sources and formulate a corpus collection scheme. At present, researchers mainly acquire corpus materials from literature databases, industrial databases, professional news websites, ancient texts, book series, dictionaries, monographs, and lexical reference works. Given that sources of interpreting data are more distinctive than those for general corpora, video materials

from simultaneous interpreting conferences and press briefings can also be collected. Printed textual materials can be digitized via OCR devices and supporting software, while audio materials can be transcribed using the iFLYTEK Input Method.

Overseas, numerous studies construct corpora based on search engine outputs and conduct data collection starting from web pages. Web crawlers are adopted to retrieve corpus resources from the Internet for corpus expansion. In 2005, Gulli and Signorini estimated that there were 11.5 billion indexable web pages. In July 2008, Google stated in an official blog post that it had indexed well over one trillion web pages.

Corpus data annotation. Data annotation plays a vital role in corpus construction. The quality of annotation largely determines corpus quality, the accuracy of research findings derived from the corpus, and the extent of corpus utilization (Huang & Wang, 2021).

At present, there exist three primary data annotation strategies: manual annotation, automatic annotation, and hybrid human-machine annotation (Huang & Wang, 2021). Given the heavy workload and low efficiency of purely manual annotation, most corpora adopt automatic annotation or human-machine collaborative annotation. Part-of-speech annotation and semantic annotation can be implemented, followed by manual proofreading to rectify errors generated in automatic processing. Due to the inevitable ambiguity in phonetic feature mapping across languages in translation, lexicons used in automatic processing contain errors that require manual cleaning. This study carries out part-of-speech annotation and semantic annotation for corpus data. For part-of-speech tagging, the CLAWS English tagger developed by Lancaster University is adopted. It achieves an accuracy of 96%-97%. Its updated version, CLAWS4, was employed to perform POS tagging on the 100-million-word British National Corpus (BNC). The tagger annotates speech by assigning each word a label indicating its grammatical category. It supports three tag sets: C5, C7, and C8. The C7 tag set, consisting of 137 categories, is the most widely used and enables fine-grained linguistic differentiation.

For semantic annotation, the UCREL Semantic Analysis System (USAS), also developed by Lancaster University, is utilized. USAS constitutes a framework for automatic semantic analysis of text. It was originally developed for English by Paul Rayson in C. Subsequent versions of the multilingual semantic tagger were developed by Scott Piao in Java, and later by Andrew Moore in Python (pymusas). Adopting hierarchical semantic taxonomy, semantic lexical resources and disambiguation algorithms, such as templates and contextual rules, USAS assigns semantic categories to individual words and multi-word expressions (MWEs) within running text.

Corpus alignment and import. Huang Junhong et al. (2007) argued that the finer the unit granularity of corpus alignment, the richer linguistic information it can offer and the greater its application value. For corpus alignment, this study adopts SISUAligner 2.0.0, developed by the research team led by Professor Hu Kaibao from the Institute of Corpus Studies, Shanghai International Studies University. This tool supports parallel alignment of bilingual or multilingual texts and enables the editing and alignment of parallel text units in one-to-one, one-to-two and one-to-many correspondence modes. The exported aligned corpus is compatible with parallel corpus retrieval tools, such as ParaConc. Afterwards, the aligned corpus is saved in TXT format and imported into a database. MySQL, the most widely used relational database management system, is selected. For web applications, MySQL ranks among the premier relational database management software.

Data cleaning and maintenance. Owing to the scale of corpora, corpus construction often relies heavily on fully automatic processing. Nevertheless, automatic processing may alter raw data and consequently degrade

data quality. Accordingly, it is essential to design dedicated cleaning procedures to handle noisy data. Typical operations include removing duplicated and redundant content, as well as rectifying problematic punctuation and spelling errors.

Diversified Applications of the Interpreting Corpus

This section analyzes the application value of the corpus. The corpus can be deployed in three scenarios: an intelligent interpreting platform, intelligent teaching, and interpreting quality assessment.

Application in the Intelligent Interpreting Platform

Traditional interpreting training is mostly conducted during classroom sessions. Learners can hardly obtain effective feedback for after-class practice. In conventional interpreting classrooms, large student cohorts and limited teaching hours make one-on-one instruction barely feasible. Classroom teaching mainly focuses on interpreting methodologies and skills, yet traditional training can only bring students up to an average proficiency level, falling far short of cultivating outstanding interpreters. With the advancing computer technologies and the widespread application of AI across various fields, research on English-Chinese bilingual parallel corpora has grown increasingly sophisticated. If the interpreting corpus is integrated with AI speech recognition for interpreting training, the learning efficiency of interpreting learners can be greatly improved.

The English-Chinese bilingual interpreting corpus constructed in this study can be applied to an intelligent interpreting management platform, forming an innovative interpreting training model integrating resource integration, terminology management and interpreting context analysis. The platform aggregates abundant interpreting resources. It collects, analyzes, and categorizes videos, audios, transcripts, images, and other materials of consecutive and simultaneous interpreting conferences to present a full picture of language from multiple dimensions. Drawing on resources processed via interpreting context analysis, the platform analyzes conference backgrounds, appropriate interpreting styles, and pausing patterns. Image recognition technology is embedded into the system to realize the three-dimensional reproduction of interpreting scenarios. From three perspectives—participants, activities, and the system—the platform observes and analyzes on-site interpreting videos and extracts relevant data to construct a multimodal training environment. It enables trainees to simulate real interpreting scenarios and comprehend linguistic and cultural contexts, and supports terminology management relying on the established interpreting corpus. Upon the completion of training, the speech input system revises transcribed texts and assists users in analyzing interpreting errors. Apart from catering to interpreting learners, the platform is also accessible to the general public with interpreting demands for daily communication with native English speakers.

To intuitively demonstrate how the intelligent interpreting training platform improves learners' interpreting proficiency, this study adopts an experimental comparison method. Sophomore translation majors with comparable proficiency were recruited and divided into two groups for one-month interpreting training. The experimental group, namely, the intelligent interpreting group, practiced within the three-dimensional interpreting training environment supported by the platform, with error analysis provided by the intelligent system. The control group received conventional interpreting training without any auxiliary intervention. Details of the experiment are outlined as follows:

Selection of training materials. Both groups used identical video materials throughout the month. The materials were excerpts from key simultaneous interpreting conferences, each lasting approximately four minutes.

These clips reflect conference characteristics, core themes and pivotal terminology.

Training procedures. Members of the intelligent interpreting group logged onto the intelligent interpreting management platform to access pre-training analyses of interpreting videos, including conference backgrounds, features and styles. They practiced in the three-dimensional interpreting scenario built by the platform and reviewed evaluations generated by the system to correct mistakes after training. The control group followed conventional training methods with no additional intervention. After receiving training materials, participants in both groups completed practice, recorded videos of their performance and submitted the recordings to the researcher's mailbox.

Experimental results. To compare training outcomes after one month and control variables to the greatest extent possible, participants were selected from translation majors with similar proficiency. Prior to the experiment, all participants took a pre-experiment interpreting test graded by two professors specializing in interpreting. The average scores of the two groups are presented below: The average score of the Intelligent Interpreting Group was 71.25, while that of the Traditional Interpreting Group stood at 71.75. After one month of training, participants underwent a post-test of interpreting performance and were asked to analyze their respective translations. The test results are shown below: the average score of the Intelligent Interpreting Group was 77.5, while that of the Traditional Interpreting Group stood at 75.25.

Discussion. An analysis of the test scores reveals that both groups achieved progress, yet the improvement observed in the Intelligent Interpreting Group was more remarkable. Furthermore, according to participants' self-reflections on their final test outputs, members of the Intelligent Interpreting Group were capable of analyzing the assigned interpreting materials, adjusting lexical choices to better match the stylistic features of the source texts, and consciously immersing themselves in simulated interpreting scenarios. Findings from questionnaires distributed to the two groups indicate that participants in the Intelligent Interpreting Group acknowledged that the scenario simulation function of the intelligent interpreting platform helped them accumulate practical experience, and the automated error correction system enabled them to identify interpreting mistakes more efficiently.

Intelligent Teaching

Further research is still required regarding corpus processing, teaching models and other relevant issues in existing studies. Interpreting teaching supported by computer-aided technology faces the challenge of insufficient integration between teaching methodologies and technical tools (Huang & Wu, 2021).

Traditional interpreting teaching in China generally adopts the "teacher plus textbook" model, which features rigid content and dampens students' learning motivation. Professional interpreting instruction should not be a lecture-based model focusing on disciplinary theories. Instead, it ought to prioritize learners' practical capabilities—namely their core practical interpreting skills—and adopt a learner-centred practical teaching model. Aimed at fostering students' interpreting competence and even professional literacy, interpreting teaching is characterized by skill-orientation, practicality and simulation (Wang & Ye, 2009). Accordingly, the single "teacher plus textbook" model cannot fully unlock learners' potential, leaving insufficient training for their comprehensive capabilities.

The emergence of corpora has exerted a profound impact on teaching and may break the existing deadlock. Interpreting corpora feature authentic contexts and natural language, providing abundant references for classroom instruction and learners' autonomous practice. Based on the development of interpreting corpora, a shareable platform for teaching and practical training can be constructed. By applying the interpreting corpus built in this

research to interpreting instruction and carrying out corpus-based classroom teaching, interpreting teaching can be organically integrated with the interpreting corpus. The shift from conventional instruction to intelligent teaching can be realized via intelligent pedagogy, which is expected to greatly optimize interpreting teaching design and improve teaching outcomes, learners' learning efficiency and practical competence.

Intelligent interpreting teaching needs to transform the traditional passive model of "teachers lecturing and students listening." A new teaching framework should be established, in which teachers act as facilitators and students occupy the central position, encouraging students to participate autonomously, collaborate with peers and engage in innovative practice. Teachers conduct instruction around case studies with the same theme in each session and guide students to practise with diverse cases. Each theme carries distinctive features: For instance, business conference themes involve specialized business terminology, while environmental protection themes cover the latest environmental policies and initiatives. Major procedures of intelligent interpreting teaching are outlined below.

First, teachers select case themes and interpreting practice videos with appropriate durations. Key vocabulary of the selected themes shall be covered within the interpreting corpus.

Second, teachers share additional interpreting video resources on relevant themes. Students acquire contextual information from these materials and retrieve thematic vocabulary from the corpus according to source texts to reduce time spent on background research and facilitate pre-interpreting preparation.

Third, students make predictions based on the assigned training themes, getting acquainted with relevant terminology and concepts in advance to enhance interpreting accuracy and flexibility.

Fourth, the interpreting process is recorded on video. After completing practice, students log into the intelligent interpreting training platform and upload recordings. Teachers can access the platform's analytical feedback on learners' interpreting accuracy, fluency, and completeness via the backend, followed by targeted comments on students' performance. Teachers may then organize group discussions for students to share the functions of the interpreting corpus and insights gained from the training.

The intelligent interpreting teaching model combining the interpreting corpus and the intelligent interpreting training platform conforms to teaching demands in the information era. This model not only supplies teachers with ample interpreting examples and provides learners with after-class learning materials. Supported by information technology, it leverages integrated resources to construct interpreting scenarios and reproduce real interpreting settings. Classroom activities can therefore simulate authentic interpreting tasks, realizing scenario-based classroom instruction.

Interpreting Quality Assessment

Interpreting quality assessment has attracted growing attention among relevant researchers. With the advancement of modern technologies and the improvement of corpus construction, applying interpreting corpora to interpreting quality assessment can boost the development of relevant research. At the 9th National Conference and International Symposium on Interpreting Studies, Professor Maurizio Viezzi proposed four determinants of quality: equivalence, accuracy, appropriateness, and usability (Liu & Xu, 2012). A review of literature on interpreting quality assessment reveals that researchers generally converge on four measurable evaluation dimensions: information equivalence, accuracy, fluency, and completeness. These four quantifiable indicators play a vital role in interpreting quality evaluation, and an interpreting corpus can be adopted to assess these dimensions and provide partial references for evaluating interpreters' performance.

Practical procedures. First, interpreters log into the intelligent interpreting management platform supported by the interpreting corpus and upload field interpreting recordings. The platform transcribes recorded interpreting output into textual transcripts. Subsequently, the transcribed interpreter output is compared against reference translations (generated via corpus-based machine translation and revised manually). The platform analyzes the professionalism of interpreters' lexical choices at the word level, calculates the rate of information equivalence and accuracy between target texts and source texts at the sentence level, evaluates the completeness of interpretations at the discourse level, and measures fluency based on pause intervals extracted from uploaded audio recordings. Through monitoring the whole interpreting process, the system also analyzes the pronunciation, intonation and speech rate of the source language.

Effect analysis. Compared with conventional manual interpreting quality evaluation, quantitative assessment supported by the interpreting corpus delivers clear scores and analytical results, which objectifies the vague descriptive comments commonly adopted in manual evaluation. It should be noted that machine-based assessment only enables partial quantification, as factors, such as reaction speed, application of interpreting strategies, interpreters' posture and appearance cannot be analyzed automatically. This method improves evaluation efficiency. Meanwhile, interpreters can clearly identify their weaknesses in terms of information equivalence, accuracy, fluency, and completeness, and conduct targeted improvements down to specific indicators. This approach also brings substantial benefits for students: It stimulates exploratory learning, enhances learners' autonomous learning ability, and helps them detect deficiencies in a timely manner.

Conclusion

In summary, based on the construction of interpreting corpora, this paper explores the development and diversified applications of an interpreting corpus integrated with AI speech recognition technology. It aims to innovate interpreting teaching and improve the efficiency of interpreting quality assessment. Drawing on this research framework, a self-directed interpreting learning platform can be established to facilitate online interactive interpreting practice and evaluation for learners, so as to satisfy personalized demands in interpreting learning. The integrated application of speech input and interpreting corpora realizes innovations in specialized interpreting corpus development and broadens relevant research methodologies. It is expected that continuous exploration and innovation will contribute to the establishment of an interpreter training system with Chinese characteristics.

References

Atkins, S., Clear, J., & Ostler, N. (1992). Corpus design criteria. *Literary and Linguistic Computing, 7*(1), 1-16.

Huang, J. H., Fan, Y., & Huang, P. (2007). A review of alignment techniques for bilingual parallel corpora. *Computer-Assisted Foreign Language Education, 12*(118), 21-25.

Huang, L. H., & Wu, Y. (2021). Construction and application of a multimodal interpreting teaching corpus based on immersive experience and modeling. *Foreign Language Learning Theory and Practice, 176*(4), 127-136.

Huang, S. Q., & Wang, D. B. (2021). A review of domestic corpus research. *Journal of Information Resources Management*, 3, 4-17.

Li, X. L. (2022). Promoting foreign language discipline construction via digital humanities. *Chinese Social Sciences Today,* 2022-11-08(003). DOI:10.28131/n.cnki.ncshk.2022.005026.

Liu, H. (2006). Construction of ultra-large classified corpus. *New Technology of Library and Information Service*, 1, 70-73.

Liu, H. P., & Xu, M. (2012). Exploring interpreter training models in the era of globalization: Review of the 9th National Conference and International Symposium on Interpreting Studies. *Chinese Translators Journal*, 5, 53-59.

Wang, B. H., & Ye, L. (2009). Construction of teaching-oriented interpreting corpora: Theory and practice. *Foreign Language World*, 2, 23-32.

Wang, K. F., & Fu, R. B. (2020). Corpus-based interpreting research: Progress and prospects. *Chinese Translators Journal*, 6, 13-20.

Wang, W. W., & Wang, B. H. (2025). A survey on the current status of applying AI-assisted machine translation tools in China's language service industry. *Foreign Language Educational Technology*, 2, 25-30.

Zhang W. (2012). Research status and development trends of interpreting corpus studies in the past decade. *Journal of Zhejiang University (Humanities and Social Sciences)*, 2, 193-205.