1 Corpus-based interpreting studies in China

A critical review and future directions

Hao Yin, Han Xu, and Kanglong Liu

1.1 Introduction

Since the late 1990s, corpus-based translation research has significantly enriched translation studies, transitioning the field from a prescriptive to a descriptive approach (Baker 2019). The application of corpus linguistics enables examination of linguistic features based on extensive empirical databases of natural discourse, promoting robust and comprehensive analyses (Biber, Conrad, and Reppen 1994).

CIS initiated by Shlesinger (1998) has since illuminated interpreting research. CIS primarily aims to reveal the unique patterns and features of interpreted texts and explore the complex cognitive processes used by interpreters (Setton 2011). The European context, renowned for its multilingual environment and available data, has significantly contributed to CIS. The European Parliament Interpreting Corpus (EPIC), for example, is a valuable resource that has facilitated various studies (Bernardini, Ferraresi, and Miličević 2016; Russo, Bendazzoli, and Sandrelli 2006).

Contrasting Europe's advances, China's CIS, despite its late start, has experienced rapid growth since the late 2000s. CIS in China largely focuses on Chinese-English interpreting, involving two genetically distant languages that present unique challenges (Crystal 1987, p. 292). Understanding these complexities is vital in reviewing CIS research in China, as it highlights the distinctive difficulties faced by Chinese-English interpreters. While previous reviews exist (Wang and Tang 2020), a comprehensive examination of CIS in China's current research status is critical to further understanding the opportunities and challenges in Chinese-English interpreting studies.

1.2 Methodology

1.2.1 Data collection and description

This investigation provides a comprehensive overview of the research landscape surrounding CIS in China since the emergence of the first interpreting corpus in 2008 (Wen and Wang 2008). The focus is primarily on studies

DOI: 10.4324/9781003377931-2

conducted over the past decade, during which CIS in China experienced significant growth and development. To conduct a systematic survey, two prominent academic databases, Web of Science (WoS) and China National Knowledge Infrastructure (CNKI), were utilized. The WoS database, specifically the Social Sciences Citation Index (SSCI) edition, was queried using the title keywords "interpret*" and "corp*" to retrieve relevant articles. The search was limited to the period between January 1, 2008, and December 31, 2022, with English language restriction and document type limited to articles or early access publications. The research areas were confined to Linguistics, Language and Linguistics, and Communication. Similarly, the CNKI database, which encompasses influential Chinese scholarly journals indexed by the Chinese Social Sciences Citation Index (CSSCI), was employed. Ten key journals[1] focusing on translation studies were selected (Liu, Cheung, and Zhao 2022). The title keywords "口译" (interpreting) and "语料库" (corpus or corpora) were used in the search, with the time frame set from the same period (2008–2022) and the document type limited to articles. The search yielded 56 English articles from WoS and 41 Chinese articles from CNKI.

1.2.2 Data analysis

From the WoS database, we initially obtained a total of 56 articles. After manually screening for relevance and removing unrelated articles, we were left with 33 articles that are related to CIS in the Chinese context. Among the 33 articles published from 2013 onwards, further screening resulted in a final selection of 16 articles that met our criteria. Similarly, the CNKI database yielded a total of 41 articles. Within this collection, 17 articles were published before 2013, but only ten of them were found to be relevant after rigorous evaluation. For the articles published from 2013 onwards, a careful screening process resulted in a final selection of 22 articles that met our specific research requirements.

It is worth noting that among the ten articles published prior to 2013, researchers primarily focused on three key areas: corpus compilation (e.g., Hu and Tao 2010; Li and Wang 2012; Wang and Ye 2009), the pedagogical use of interpreting corpus (e.g., Tao 2010; Wang and Ye 2009), and the specific linguistic features of interpreted speech, including lexical bundles (Wang and Huang 2011), explicitation (e.g., Dai 2011; Hu and Tao 2009), and simplification (Li and Wang 2012). During this period, the research was in its early stages, with only a few large-scale corpora being constructed for studying interpreting. Moreover, the parameters used for investigating interpreting mainly relied on traditional measures, such as mean sentence length, mean word length, and lexical density. Statistical methods were infrequently employed, with a greater emphasis on frequency counting. As a result, our current review divides the analysis into two distinct periods: 2008–2012 and 2013–2022, with a specific focus on the latter period. The first period is referred to as the exploratory period, while the latter period is termed the maturity period. This approach allows for a focused examination of the developments in the field over the past

decade, providing valuable insights into the advancements in corpus-based interpreting studies.

Through systematic analysis and summarization of 48 articles within the domain of CIS, our primary aim is to provide a comprehensive understanding of these papers and identify emerging research trends. In order to accomplish this objective, we employed a coding scheme, adhering to the methodology delineated by Liu, Cheung, and Zhao (2022), to conduct a thorough analysis of the articles. Each article underwent meticulous examination, and relevant information pertaining to various dimensions such as article type, research framework, subject field, mode of interpreting, corpus compilation methods, corpus software employed, and linguistic features analyzed, was meticulously recorded in an Excel spreadsheet. This comprehensive analysis facilitated an in-depth exploration of the specific research content. Finally, the synthesized data and detailed analysis enabled us to discern the prevailing research trends within the field of CIS.

1.3 Overview of corpus-based interpreting studies in China

This section offers an overview of CIS in China. In Section 1.3.1, we present a comprehensive account of studies conducted from 2008 to 2012, a period marked by the emergence of a limited number of articles, predominantly focused on corpus compilation with only a few incorporating corpus-based research. As a result, this phase was characterized as the exploratory period. Moving on to Section 1.3.2, we delve into the maturity phase, spanning the years 2013 to 2022. This period witnessed a significant surge in research efforts devoted to CIS in China. The range of research topics expanded, accompanied by the introduction of comprehensive corpus software for compilation, transcription, and annotation. Furthermore, there was a notable increase in the development of self-compiled corpora, coupled with a diversification of research methodologies employed.

1.3.1 The exploratory period (2008–2012)

The period spanning from 2008 to 2012 in China can be characterized as the exploratory period. This investigation revealed that during this time, no articles were published in international journals, and only ten influential articles were identified from CNKI. Among these ten articles, three involve the general discussion of interpreting corpora, including the comprehensive exploration of the construction and development of interpreting corpora (Zhang 2009) and the general use of corpora in teaching interpreting (Tao 2010; Wang and Ye 2009).

As the field of interpreting research advances, there has been a noticeable surge in efforts dedicated to constructing specific interpreting corpora. A pioneering achievement in this regard was realized by Wen and Wang (2008) through the creation of the Parallel Corpus of Chinese EFL Learners–Spoken

(PACCEL-S), marking the first interpreting corpus in China. PACCEL-S encompasses transcripts from TEM 8 interpreting tests, totaling 495,231 words. This corpus is meticulously annotated with part-of-speech (POS) tags, header markup, grammatical mistake markup, paralinguistic information markup, and sentence-level alignment. While primarily designed as a learners' corpus, PACCEL-S has played a pivotal role in guiding subsequent endeavors related to the compilation, alignment, and annotation of interpreting corpora in China. Exploiting the PACCEL-S corpus, Dai (2011) conducted a study examining disfluency differences in Chinese-English interpreting, specifically exploring variations between female and male EFL learners, as well as between high-scoring and low-scoring groups. This investigation capitalized on the valuable insights provided by the PACCEL-S corpus, shedding light on the aforementioned distinctions within the context of interpreting research.

Hu and Tao (2010) contributed to this development by creating the Chinese-English Conference Interpreting Corpus (CECIC). CECIC encompasses 544,211 words and includes original Chinese texts alongside their English interpretations from press conferences held by the Chinese government between 1988 and 2008. During the compilation process, sentence alignment was conducted using ParaConc, while Chinese word segmentation was done using ICTCLAS3.0 developed by the Institute of Computing Technology of the Chinese Academy of Sciences. In addition, Hu and Tao (2010) delineated two primary research directions that can be pursued utilizing the CECIC: interpreting studies on Translation Universals and investigations into linguistic features. Building upon the CECIC, Hu and Tao (2009) conducted a study examining the explicitation of textual meaning and its underlying reasons in Chinese-English conference interpreting using ParaConc. Similarly, Wang and Huang (2011) explored the use of chunks in Chinese-English interpreting by senior students based on the same corpus. Furthermore, employing the CECIC, Hu and Tao (2012) investigated the syntactic features of press conference interpreting.

Several other corpora were compiled, each contributing to the expanding landscape of interpreting research. Wang (2012) presented the construction and processing of the "Corpus of Chinese-English Interpreting for Premier Press Conferences" (CEIPPC), a self-built corpus. This corpus comprises transcripts of original Chinese speeches and their simultaneous interpretations from the annual Chinese Premier Press Conferences held between 1998 and 2012, encompassing approximately 100,000 words. The primary corpus tool employed for this compilation was ParaConc. In a separate study, Li and Wang (2012) utilized the self-built comparable subcorpus of the "Hong Kong Bilingual Interpreting Corpus on Contemporary Social Life" (BICCSL) to investigate lexical patterns. The annotation of POS tags was carried out using the Free CLAWS WWW trial service, while WordSmith was utilized for examining word frequencies. By 2012, four significant corpora have been compiled to study Chinese-English interpreting: PACCEL-S, CECIC, CEIPPC, and BICCSL. These corpora have furnished invaluable resources for further

research and analysis within the field. However, it is important to note that during this period, there was a lack of standardized systematic techniques concerning corpus compilation strategies, with scholars employing their individual approaches to corpus construction. Furthermore, the extensive utilization of the PACCEL-S corpus in interpreting research resulted in a relatively greater emphasis on interpreting learners. It is worth noting, however, that the data contained within this corpus originated from TEM-8 interpreting tests, rather than authentic work or learning scenarios, which raises concerns regarding the extent to which analyses are based on naturally occurring interpreting discourses (Zhang 2012).

1.3.2 The maturity period (2013–2022)

The period from 2013 to 2022 stands as a significant era of growth and advancement for CIS in China, forming the primary focus of this comprehensive review. During this decade, the field of interpreting corpora research has witnessed a notable expansion, encompassing a diverse range of subjects and resulting in a substantial increase in both the size and scope of interpreting corpora compared to the previous period. Noteworthy progress has been made in various aspects of corpus development, including construction, transcription, annotation, retrieval techniques, and data analysis. This progress has led to the emergence of several exemplary corpus analysis software tools (Li and Hu 2013; Li and Halverson 2020; Wang, Li, and Li 2019). A thorough examination was carried out, encompassing a total of 38 articles published between 2013 and 2022. Among these articles, 22 are research papers, while the remaining 16 consisted of introductory, review, or summary papers that did not involve empirical studies. The focal point of the analysis centered on the 22 research papers, with a comprehensive investigation into their research approach and research object. In terms of research approach, the majority of 19 studies (86%) in the field of interpreting have predominantly focused on synchronic aspects. In these studies, researchers have directed their investigations toward specific timeframes to gain insights into interpreting practices. However, the exploration of diachronic aspects, which involves studying the evolution of interpreted text over its historical trajectory, has been relatively limited, with only one study conducted by Pan and Wang in 2021 addressing this dimension. Furthermore, two studies, namely Gu (2019) and Gu and Tipton (2020), have adopted an integrated approach by concurrently examining both synchronic and diachronic dimensions. By doing so, they have aimed to provide a comprehensive understanding of interpreting practices.

The 22 papers reviewed in this study can be broadly classified into three distinct categories based on their research object: linguistic features of interpreting language, translation universals observed in interpreting, and the interpreting process (Wang and Tang 2020). Among these categories, the interpreting process emerges as the most prevalent research subject, accounting for 59% (13 out of 22 articles). This body of work examines various aspects such

as shifts (Pan and Wang 2021; Wang and Qin 2015), non-renditions in court interpreting (Cheung 2017), the mediation of China's discourse by Chinese government-affiliated interpreters (Gu 2018, 2019; Pan and Zheng 2017), terminological preparation for simultaneous interpreting (Xu 2018a, 2018b), propositional information loss (Lu 2018), pauses in interpreting (Wang, Li, and Li 2019), uncertainty in interpreting (Shen, Lv, and Liang 2019), metaphor in interpreting (Sheng 2021), and lexical bundles in interpreting (Li and Deng 2022).

Linguistic features of interpreting language comprise another significant area of research, accounting for 23% (5 out of 22 articles) of the reviewed literature. Within this domain, two studies specifically examine lexical features, focusing on lexical patterns (Fu and Wang 2021) and modal verbs (Li and Hu 2013). One study delves into larger linguistic units, that is, 4-gram lexical bundles (Li and Halverson 2022). Additionally, two studies investigate discoursal features, particularly the usage of hedges (Pan and Zheng 2017; Wang and Li 2015). Furthermore, the investigation of translation universals in interpreting constitutes another prevalent research subject. In particular, explicitation (Tang and Li 2017; Li and Halverson 2020) and simplification (Lv and Liang, 2019) emerge as primary focal points within this category. It is worth noting that one article deviates from the aforementioned categories, examining how foreign media outlets reported on the press conference interpretation during the Chinese Prime Minister's meetings with both Chinese and foreign journalists by utilizing data from the Factiva Global News Database (Cheng and Xu 2020).

Regarding the utilization of corpora, a total of 29 papers in this review draw upon specific interpreting corpora, while the remaining nine papers focus on general introductions of interpreting corpora, such as review articles (Chen and Fu 2014; Wang and Fu 2020), general introductory papers of interpreting corpora (Deng 2018), and of the principles and approaches (Zhang 2013).

A comprehensive account of the methodology employed in the 29 studies, encompassing corpus design, compilation, and annotation, is presented in Section 1.4.

1.4 Corpus design

1.4.1 Sources and fields of the corpus data

It is evident that the corpora at this stage encompass a wide array of fields or disciplines, such as education (Tang and Li 2017), the legal realm (Cheung 2017), business and academia (Lu 2018), among others. The most common source of corpus data is derived from Chinese government press conferences (Fu and Wang 2021; Gu 2018, 2019; Gu and Tipton 2020; Li and Hu 2013; Pan and Wang 2021; Pan and Zheng 2017; Shen, Lv, and Liang 2019; Wang and Li 2015; Wang and Qin 2015), which cover a wide range of fields including politics, economy, military, diplomacy, and people's livelihood. Another significant

data source originates from interpreting learners (Li and Deng 2022; Wang, Li, and Li 2019; Zhang 2015a, 2015b, 2017, 2019), with the learner corpora being CILC and PACCEL-S. Among the 29 articles, 23 (79%) utilized self-constructed corpora, five (17%) employed existing corpora, and one (Sheng 2021) added new interpreting data to the existing corpus CECIC (Hu and Tao 2010) and analyzed the updated corpus.

It is noteworthy that several corpora have been consistently reused in multiple studies, reflecting their importance and value in interpreting research. For example, the self-constructed CILC has been extensively utilized by Zhang in a series of articles, covering various aspects such as corpus construction processes, paralanguage tagging standards and procedures in CILC (Zhang 2015a), interpreting strategies tagging methods and their significance in CILC (Zhang 2015b), an overview of CILC construction and research (Zhang 2017), as well as linguistic information tagging strategies and analysis in CILC (Zhang 2019). In a similar vein, Gu employed the self-made Chinese-English Political Discourses Corpus (CE-PolitDisCorp) in three articles to delve into the mediation of China's discourse on past actions and achievements by government interpreters (Gu 2018). Furthermore, Gu examined the government-affiliated interpreters' role in mediating and reconstructing China's discourse on PEOPLE (Gu 2019), and investigated how these interpreters mediate Beijing's discourse at various levels using self-referential terms (Gu and Tipton 2020). Li and Halverson utilized another self-made corpus, the Chinese-English Interpreting Corpus of Premier Press Conference (CICPPC), to illustrate the intricacies involved in determining causal factors for explicitation (Li and Halverson 2020). They also investigated the discourse functions and relationships to the source text of 4-gram lexical bundles in the interpreted text (Li and Halverson 2022). In a separate study, Cheung employed a court interpreting corpus to examine whether court interpreters actively coordinate communication while performing their interpreting duties (Cheung 2017). Furthermore, Pease, Pease, and Cheung (2018) proposed a novel method of discourse analysis based on speech act theory and formal ontology, drawing insights from the same corpus. During this phase of research, the PACCEL-S corpus, which is considered the first interpreting corpus in China, was employed on two occasions. Wang, Li, and Li (2019) investigated pauses in Chinese-English interpreting using the PACCEL-S corpus, while Li and Deng (2022) explored the frequency effects of lexical bundles in interpreting with the same corpus.

In summary, self-constructed corpora play a prominent role in corpus-based interpreting research, indicating the well-developed techniques employed in compiling English-Chinese interpreting corpora. This prevalence also underscores researchers' substantial confidence in the adaptability of such corpora. While researchers explore a wide range of interpreting fields, the available sources remain somewhat limited, primarily reliant on data derived from press conferences. This limitation likely stems from the inherent difficulty in accessing authentic interpretation corpora (Li and Li 2010). Notably, corpus data from press conferences is typically accessible online, whereas interpreting

data from other domains such as business and medicine tend to be predominantly confidential.

1.4.2 Corpus compilation and annotation

Excluding the 13 articles that did not provide details on transcription methods, manual transcription was employed for the corpus data in 14 articles. Interestingly, only two articles highlighted the utilization of relatively novel transcription methods. In one study, researchers utilized corpus transcription and alignment with the support of Tencent AI, supplemented by the manual intervention (Lu and Zhang 2022). The other article implemented compound transcription, which involved linear time alignment, sentence alignment, and information alignment (Zhang 2017). These instances suggest that the transcription of interpreting corpora predominantly relies on manual methods, while the exploration of automated transcription technology is still at an early stage.

Regarding corpus compilation, most articles did not specify the particular methods employed. However, Tang and Li (2017) indicated the use of manual alignment for aligning bilingual corpora. On the other hand, Pan and Zhang (2017) and Pan and Wang (2021) mentioned the utilization of ParaConc for aligning bilingual corpora. Extracting corpus data directly from existing resources also emerged as a convenient option. Zhang (2019) conducted a study by selecting corpora from the larger CILC corpus through stratified sampling and subsequently establishing a smaller corpus. Xu (2018a) mentioned the utilization of the Sketch Engine platform's focused web "crawler," WebBootCat, to automatically collect a network corpus. Additionally, Xu constructed a smaller English-Chinese comparable professional corpus to enhance the pre-interpreting mode. It can be inferred that there currently exists no automated tool for corpus compilation, and most of the corpus compilation work necessitates manual completion. Regarding annotation methods, it is noteworthy that 16 articles did not provide explicit information on this aspect. However, manual annotation was frequently employed, alongside the emergence of various innovative annotation techniques. Wang and Qin (2015) exemplified a combination of manual and automatic tagging, utilizing the Stanford POS Tagger for English part-of-speech tagging and ICTCLAS for Chinese part-of-speech tagging. Lv and Liang (2019) utilized the automatic part-of-speech tagging tool "Free CLAWS WWW tagger" (UCREL) to identify lexical simplification parameters. Fu and Wang (2021) employed Treetagger for part-of-speech restoration and tagging to analyze lexical features within the corpus. Additionally, Pan, Wong, and Wang (2022) demonstrated the possibility of exploring learner data through human annotation, machine-facilitated human annotation, and human-supervised/edited machine annotation, presenting three case studies to support their claims. Similarly, Lu and Zhang (2022) adopted an automatic annotation-first approach, followed by manual annotation, to annotate interpreting strategies, information equivalence, and paralanguage in

interpreting. Evidently, as research progresses in this field, annotation methods have evolved into more sophisticated and nuanced techniques.

1.4.3 Types of corpora

In terms of corpus data types, Consecutive interpreting (CI) emerges as the most prevalent mode, encompassing 55% (16 out of 29 articles) of the corpus data. Two articles made use of SI data, while eight articles (28%) incorporated both CI and SI data. One corpus data source deviated from the CI and SI categories, specifically the foreign media coverage of the Chinese Prime Minister's interactions with Chinese and foreign journalists in the Factiva Global News Database (Cheng and Xu 2020). With regard to corpus types, 23 articles employed parallel corpora, three utilized comparable corpora, and three made use of single-language corpora. The limited availability of comparable corpora strongly suggests that researchers have primarily focused on examining the linguistic features of interpreted texts or the process of interpreting, while neglecting comprehensive comparisons across various modalities, including interpreted texts, translated texts, and source languages. This noticeable lack of comparative analysis highlights a promising avenue for future research, offering an opportunity to delve further into the intricate nuances of interpreting and translation practices.

1.5 Methodological issues

This section delves into an examination of the methodological approaches employed in 22 research articles, specifically from the perspective of data collection and analysis methods.

1.5.1 Features for analysis

In terms of data collection methods, the majority of the 22 research articles (20 out of 22) employed a corpus linguistics approach. However, two studies took a different route by utilizing focus group interviews (Xu 2018b) and a combination of simulated experiments and focus group interviews (Xu 2018a). The objective of these studies was to verify the effectiveness of incorporating a corpus-based terminological preparation procedure. In the remaining 20 research articles, data analysis primarily revolved around calculating the frequency of specific linguistic indicators, which were classified into three distinct levels: lexical, syntactic, and discoursal.

Three studies delved into the lexical characteristics of interpreted texts through corpus analysis. Li and Hu (2013) examined the usage of modal verbs in Chinese-English press conference interpreting, investigating their distinctive features and underlying motivations. They conducted a frequency analysis of English modal verbs in the Chinese-English conference interpreting subcorpus and the original conference English subcorpus, comparing their distributions

based on the modal degree and modal types. Lv and Liang (2019) explored the presence of distinct lexical patterns across CI and SI outputs, examining information density, lexical repetitiveness, and lexical sophistication. Their corpus encompassed SI and CI output texts, read-out translated speeches, and non-interpreted original English speeches, with findings indicating greater simplification in CI outputs compared to SI counterparts. Fu and Wang (2021) conducted a comparative analysis of lexical features in SI, CI, and original spoken English texts. Their primary focus encompassed lexical density, type/token ratio, core vocabulary coverage, list head coverage, hapax legomena, and mean word length.

In addition to the studies focusing on lexical features, three research articles have specifically concentrated on the syntactic level by examining lexical bundles. Li and Halverson (2020) investigated the intricate task of identifying causal factors contributing to explicitation. They accomplished this by quantitatively calculating the frequency of 4-gram lexical bundles and their three structural types. In their subsequent investigation in 2022, they further explored the discourse functions and relationships between these 4-gram lexical bundles in interpreted texts and the source text. Another significant contribution was made by Li and Deng (2022), who analyzed the distribution and characteristics of lexical bundles by examining the frequency of the component words within these bundles. These studies provide valuable insights into the syntactic dimension of language through the examination of lexical bundles in interpreted texts. By exploring the frequency, distribution, and characteristics of these bundles, researchers can uncover patterns and trends that enhance our understanding of the intricate syntactic complexities inherent in interpretation.

There are 14 articles that specifically focus on the discoursal features of interpreted texts. The prevalent indicators within the category of interactional metadiscourse (Hyland 2005) include hedges, self-referential terms, and others. For example, Wang and Li (2015) examined the characteristics and motivations behind the usage of hedges by professional translators. They utilized ParaConc to identify high-frequency hedging items, such as "some," "about," "I think," "according to," and then analyzed their patterns in interpreting by finding their corresponding source text items using ParaConc. Pan and Zheng (2017) investigated gender differences in hedging in Chinese-English conference interpreting based on a transcribed parallel corpus. They employed both quantitative analysis, using AntConc to calculate the overall frequency of hedges, and qualitative analysis, which involved scrutinizing the interpreting process and documentary resources. Gu (2018) employed a corpus-based Critical Discourse Analysis (CDA) approach to critically explore how government interpreters mediate China's discourse on past actions and accomplishments in English. AntConc was used to calculate the frequency of self-referential terms, specifically the present perfect and present perfect continuous with the subjects "we," "China," and "government" in the interpreted texts. Gu (2018) further examined the mediation and (re)construction of China's discourse by

analyzing the frequency of people-related items (2019) and self-referential terms (2020) used by government-affiliated interpreters. Gu (2018) combined corpus linguistics with CDA, integrating quantitative and qualitative analysis to establish connections between linguistic features in interpreted texts and their socio-political and ideological dimensions. This approach represents a relatively new development in the field of Conference Interpreting Studies in China.

Other interpreting strategy indicators are also utilized, including shifts, non-renditions, uncertainty, and explicitation. In the context of interpreting, shifts refer to changes in the target speeches compared to the source speeches that result not from obligatory systemic differences between languages but from deliberate choices made by interpreters (Pan and Wang 2021). Wang and Qin (2015) investigated the communication norms in Chinese-English interpreting, specifically focusing on different types of shifts such as cohesive addition, elaboration, expansion of information, and explicitation of the intended meaning. They manually annotated these shifts and calculated the frequency of specific words or units categorized under each type. Pan and Wang (2021) narrowed their focus to target-oriented shifts and conducted a diachronic study. They compared the frequency of target-oriented shifts between different time periods to examine whether the interpretation for the Chinese government by institutional interpreters became more target-oriented during the 2010s compared to the 1990s. They used ParaConc to identify five types of target-oriented shifts through source-target comparison, followed by a qualitative analysis where examples were provided to further examine the changes in different categories of shifts between the two periods. Non-renditions are utterances in the target language that lack a corresponding counterpart in the source language (Wadensjö 1998/2014). Cheung (2017) examined the types and frequencies of non-renditions in interpreting to demonstrate how court interpreters actively coordinate communication during their interpreting duties. Shen, Lv, and Liang (2019) investigated the uncertainty experienced by expert interpreters at Chinese Premier Press Conferences by quantifying the frequency of uncertainty indicators, such as filled and silent non-juncture pauses, self-correction, repetition, and reformulation.

In contrast to the previous period, which saw only two studies focusing on the discourse level of interpreting texts (Dai 2011; Hu and Tao 2009), the current period demonstrates a notable shift in research focus. During this period, the majority of studies have dedicated their attention to investigating the discourse level of interpreting. This significant increase reflects the growing recognition of the crucial role that discourse plays in interpreting research, highlighting the emerging trend of exploring the complex interplay between language, context, and the communicative strategies employed by interpreters. This paradigm shift serves as a compelling indication of a maturing field, emphasizing the importance of further delving into discourse-level analyses to advance our understanding of interpreting processes and enhance the quality of interpreting practices.

1.5.2 Data analysis methods

Regarding data analysis, the field of CIS research in China has progressed beyond mere frequency counts and has embraced the use of statistical testing to ascertain the significance of observed differences. All 22 research articles included in the review involved quantitative analyses. Eleven articles relied on frequency calculations for data analysis, while the other 11 articles employed various statistical tests. These tests encompassed a range of methods, including t-tests (Xu 2018a; Zhang 2019), log-likelihood tests (Wang, Li, and Li 2019), chi-square tests (Li and Deng 2022; Li and Hu 2013; Pan and Wang 2021), and analysis of variance (ANOVA) tests (Lv and Liang 2019; Shen, Lv, and Liang 2019). Some articles even employed multiple statistical tests. For example, Pan and Zheng (2017) utilized both t-tests and log-likelihood tests, while Lv and Liang (2019) employed one-way ANOVA tests and post-hoc pairwise comparisons using Wilcoxon rank-sum tests (with Bonferroni correction). Fu and Wang (2021) incorporated chi-square tests, Kruskal-Wallis tests, and Mann-Whitney U tests in their study. The adoption of these statistical tests in CIS research is motivated by their ability to provide a rigorous and systematic approach to data analysis, allowing researchers to draw objective inferences and arrive at valid conclusions (Oakes 2019). Moreover, statistical tests enable researchers to assess the reliability and generalizability of their findings. By calculating p-values or effect sizes, researchers can evaluate the strength and magnitude of observed effects, determining the statistical significance and broader implications of their results (ibid.).

In comparison to the previous exploratory period, where the use of statistical tests in research articles was infrequent, the present period has witnessed a substantial increase in their application. This trend signifies the growing prevalence and adoption of statistical tests in recent years, reflecting the recognition among CIS researchers of the importance of employing robust quantitative methods. The utilization of advanced statistical techniques allows for a more comprehensive exploration of data, leading to deeper insights and more accurate conclusions. By embracing these sophisticated statistical approaches, researchers enhance the rigor and reliability of their findings, contributing to the advancement of CIS research.

1.6 Conclusion

The objective of this review is to provide a comprehensive overview of the current research landscape in CIS in China. To achieve this, we have examined influential articles from reputable sources including WoS and CNKI spanning the late 20th century to the present. The analysis reveals a noteworthy expansion in the research scope, attributed to advancements in techniques for compiling, transcribing, and annotating interpreting corpora. Furthermore, the utilization of self-generated corpora and the adoption of diverse research methodologies have significantly contributed to this expansion over the past

decade. The current state of development in CIS reveals several notable trends. First, synchronic studies dominate the research landscape, with a primary focus on investigating the interpretation process and exploring discourse-level aspects. A significant observation is the increasing utilization of statistical tests, highlighting a growing emphasis on quantitative analysis within the field. Furthermore, the expansion in the number and scope of corpora reflects a broader and more diverse landscape for interpreting corpora. However, several limitations have been identified in the field of corpus use in interpreting. To start with, there is a relative lack of diversity in the sources of corpora utilized for research, primarily focusing on the political interpreting of government press conferences. This narrow focus may restrict the generalizability of findings to other domains of interpreting. Moreover, although advancements have been made in transcription and annotation techniques, further improvements are required to cater to the diverse research purposes in the field (Wang and Tang 2020). Last, there is an imbalance in the availability of different types of corpora for research purposes, with a predominant emphasis on consecutive interpreting. Further efforts are warranted to address these limitations and promote a more comprehensive understanding of CIS practices.

Note

1 Key translation journals in China include Chinese Translators Journal (中国翻译), Shanghai Journal of Translators (上海翻译), Foreign Language Education (外语教学), Journal of Foreign Languages (外国语), Foreign Language Research (外语学刊), Foreign Languages in China (中国外语), Foreign Languages and Literature (外国语文), Technology Enhanced Foreign Language Education (外语电化教学), Foreign Language Teaching and Research (外语教学与研究), Foreign Language World (外语界).

References

Baker, Mona. 2019. "Corpus Linguistics and Translation Studies: Implications and Applications." In *Researching Translation in the Age of Technology and Global Conflict*, edited by Mona Baker, Kyung-Hye Kim, and Yifan Zhu, 9–24. London: Routledge.

Bernardini, Silvia, Adriano Ferraresi, and Maja Miličević. 2016. "From EPIC to EPTIC—Exploring Simplification in Interpreting and Translation from an Intermodal Perspective." *Target* 28(1): 61–86.

Biber, Douglas, Susan Conrad, and Randi Reppen. 1994. "Corpus-based Approaches to Issues in Applied Linguistics." *Applied Linguistics* 15(2): 169–89.

Chen, Jing 陈菁 and Fu Rongbo 符荣波. 2014. "Guoneiwai yuliaoku kouyi yanjiu jinzhan (1998–2012)—yi xiang ji yu xiangguan wenxian de jiliang fenxi" 国内外语料库口译研究进展 (1998–2012)—项基于相关文献的计量分析 [New Developments in Corpus-based Interpreting Studies: A Bibliometric Analysis of Relevant Chinese and Overseas Literature]. *Zhongguo fanyi*中国翻译 [Chinese Translators Journal] 35(01): 36–42, 126.

Cheng, Lulu 程璐璐 and Xu Wensheng 许文胜. 2020. "Jiyu yuliaoku de jizhe zhaodaihui kouyi yanjiu—waimei baodao de quxiaoxingwei shijiao" 基于语料库的记者招待会

口译研究 – 外媒报道的取效行为视角 [A Corpus-based Study on Press Conference Interpretation from the Perspective of Foreign Media's Selective Coverage]. *Shanghai fanyi* 上海翻译 [Shanghai Journal of Translators] (154): 42–7, 94.

Cheung, Andrew K. F. 2017. "Non-renditions in Court Interpreting: A Corpus-based Study." *Babel* 63(2): 174–99.

Crystal, David. 1987. *The Cambridge Encyclopedia of Language*. Cambridge: Cambridge University Press.

Dai, Chaohui 戴朝晖. 2011. "Zhongguo daxuesheng Hanying kouyi fei liuli xianxiang yanjiu" 中国大学生汉英口译非流利现象研究 [A Study on Disfluency in Chinese to English Interpretations of Chinese EFL Learners]. *Shanghai fanyi* 上海翻译 [Shanghai Translation] (1): 38–43.

Deng, Juntao 邓军涛. 2018. "Kouyi jiaoxue yuliaoku: neihan, jizhi yu zhanwang" 口译教学语料库:内涵、机制与展望 [Speech Repository for Interpreter Training: Concepts, Mechanisms and Prospects]. *Waiyujie* 外语界 [Foreign Language World] (186): 46–54.

Fu, Rongbo 符荣波 and Wang Kefei 王克非. 2021. "Jiyu kuamoshi leibi yuliaoku de hanying kouyi cihui tezheng yanjiu" 基于跨模式类比语料库的汉英口译词汇特征研究 [Lexical Patterns in Interpreted and Spontaneous English Speeches: A Comparable, Intermodal and Corpus-based Study]. *Waiyu jiaoxue yu yanjiu* 外语教学与研究 [Foreign Language Teaching and Research] 53(6): 912–23, 961.

Gu, Chonglong. 2018. "Forging a Glorious Past via the 'Present Perfect': A Corpus-based CDA Analysis of China's Past Accomplishments Discourse Mediat (is) ed at China's Interpreted Political Press Conferences." *Discourse, Context and Media* 24: 137–49.

Gu, Chonglong. 2019. "(Re)Manufacturing Consent in English: A Corpus-based Critical Discourse Analysis of Government Interpreters' Mediation of China's Discourse on PEOPLE at Televised Political Press Conferences." *Target* 31(3): 465–99.

Gu, Chonglong and Rebecca Tipton. 2020. "(Re-)Voicing Beijing's Discourse through Self-referentiality: A Corpus-based CDA Analysis of Government Interpreters' Discursive Mediation at China's Political Press Conferences (1998–2017)." *Perspectives* 28(3): 406–23.

Hu, Kaibao 胡开宝 and Tao Qing陶庆. 2009. "Hanying huiyi kouyi zhong yupian yiyi xianhua jiqi dongyin yanjiu – yi xiang ji yu pingxing yuliaoku de yanjiu" 汉英会议口译中语篇意义显化及其动因研究——一项基于平行语料库的研究 [Explicitation in the Chinese-English Conference Interpreting and Its Motivation – A Study Based on Parallel Corpus]. *Jiefangjun waiguoyu xueyuan xuebao* 解放军外国语学院学报 [Journal of PLA University of Foreign Languages] 32(4): 67–73.

Hu, Kaibao 胡开宝 and Tao Qing陶庆. 2010. "Hanying huiyi kouyi yuliaoku de chuangjian yu yingyong yanjiu" 汉英会议口译语料库的创建与应用研究 [The Compilation and Application of Chinese-English Conference Interpreting Corpus]. *Zhongguo fanyi* 中国翻译 [Chinese Translators Journal] 31(5): 49–56, 95.

Hu, Kaibao 胡开宝 and Tao Qing陶庆. 2012. "Jizhe zhaodaihui Hanying kouyi jufa caozuo guifan yanjiu" 记者招待会汉英口译句法操作规范研究 [Syntactic Operational Norms of Press Conference Interpreting (Chinese-English)]. *Waiyu jiaoxue yu yanjiu* 外语教学与研究 [Foreign Language Teaching and Research] 44(5): 738–50, 801.

Hyland, Ken. 2005. *Metadiscourse: Exploring Writing in Interaction*. London: Continuum.

Li, Dechao 李德超 and Wang Kefei 王克非. 2012. "Hanying tongchuan zhong cihui moshi de yuliaoku kaocha" 汉英同传中词汇模式的语料库考察 [A Corpus-based

Study on Lexical Patterns in Simultaneous Interpreting from Chinese into English]. *Xiandai waiyu* 现代外语 [Modern Foreign Languages] 35(4): 409–15, 438.

Li, Jing 李婧 and Li Dechao 李德超. 2010. "Jiyu yuliaoku de kouyi yanjiu: huigu yu zhanwang" 基于语料库的口译研究：回顾与展望 [Corpus-based Interpreting Studies: The State of the Art]. *Zhongguo waiyu* 中国外语 [Foreign Languages in China] 7(05): 100–5, 111.

Li, Xin 李鑫 and Hu Kaibao 胡开宝. 2013. "Jiyu yuliaoku de jizhe zhaodaihui hanying kouyi zhong qingtai dongci de yingyong yanjiu" 基于语料库的记者招待会汉英口译中情态动词的应用研究 [A Corpus-based Study of Modal Verbs in Chinese-English Government Press Conference Interpretation]. *Waiyu dianhua jiaoxue* 外语电化教学 [Technology Enhanced Foreign Language Education] (151): 26–32, 74.

Li, Yang李洋 and Deng Yi邓轶. 2022. "Kouyi zhong yukuai pinlv xiaoying de yuliaoku yanjiu" 口译中语块频率效应的语料库研究 [A Corpus-based Exploration of Lexical Bundles' Frequency Effects on Interpreting]. *Zhongguo fanyi*中国翻译 [Chinese Translators Journal] 43(4): 147–55, 192.

Li, Yang and Sandra L. Halverson. 2020. "A Corpus-based Exploration into Lexical Bundles in Interpreting." *Across Languages and Cultures* 21(1): 1–22.

Li, Yang and Sandra L. Halverson. 2022. "Lexical Bundles in Formulaic Interpreting: A Corpus-based Descriptive Exploration." *Translation and Interpreting Studies*. doi: 10.1075/tis.19037.li.

Liu, Kanglong, Joyce Oiwun Cheung, and Nan Zhao. 2022. "Learner Corpus Research in Hong Kong: Past, Present and Future." *Corpora* 17: 79–97.

Lu, Wei 路玮 and Zhang Wei张威. 2022. "Daxing zhongying lianxian kouyi yuliaoku gongxiang pingtai jianshe: gongneng yu caozuo" 大型中英连线口译语料库共享平台建设:功能与操作 [Construction of the Large-scale Chinese-English Connected Interpreting Corpus Sharing Platform: Function and Operation]. *Zhongguo fanyi* 中国翻译 [Chinese Translators Journal] 43(5): 108–17.

Lu, Xinchao. 2018. "Propositional Information Loss in English-to-Chinese Simultaneous Conference Interpreting: A Corpus-based Study." *Babel* 64(5–6): 792–818.

Lv, Qianxi and Junying Liang. 2019. "Is Consecutive Interpreting Easier than Simultaneous Interpreting? A Corpus-Based Study of Lexical Simplification in Interpretation." *Perspectives* 27 (1): 91–106.

Oakes, Michael. 2019. *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Pan, Feng and Binhua Wang. 2021. "Is Interpreting of China's Political Discourse Becoming More Target-oriented? A Corpus-based Diachronic Comparison between the 1990s and the 2010s." *Babel* 67(2): 222–44.

Pan, Feng and Binghan Zheng. 2017. "Gender Difference of Hedging in Interpreting for Chinese Government Press Conferences: A Corpus-Based Study." *Across Languages and Cultures* 18(2): 171–93.

Pan, Jun, Billy Tak-Ming Wong, and Honghua Wang. 2022. "Navigating Learner Data in Translator and Interpreter Training." *Babel* 68(2): 236–66.

Pease, Adam, Jennifer Cheung Pease, and Andrew K. F. Cheung. 2018. "Formal Ontology for Discourse Analysis of a Corpus of Court Interpreting." *Babel* 64(4): 594–618.

Russo, Mariachiara, Claudio Bendazzoli, and Annalisa Sandrelli. 2006. "Looking for Lexical Patterns in a Trilingual Corpus of Source and Interpreted Speeches: Extended Analysis of EPIC (European Parliament Interpreting Corpus)." *FORUM* 4(1): 221–54.

Setton, Robin. 2011. "Corpus-based Interpreting Studies (CIS): Overview and Prospects." In *Corpus-based Translation Studies: Research and Applications*, edited by Alet Kruger, Kim Wallmach, and Jeremy Munday, 33–75. London: Continuums.

Shen, Mingxia, Qianxi Lv, and Junying Liang. 2019. "A Corpus-driven Analysis of Uncertainty and Uncertainty Management in Chinese Premier Press Conference Interpreting." *Translation and Interpreting Studies* 14(1): 135–58.

Sheng, Dandan 盛丹丹. 2021. "Hanying huiyi kouyi zhong de lvcheng yinyu—jiyu yuliaoku de yanjiu" 汉英会议口译中的旅程隐喻 – 基于语料库的研究 [A Corpus-based Study on Journey Metaphors in Chinese-English Conference Interpreting]. *Shanghai fanyi* 上海翻译 [Shanghai Journal of Translators] (156): 65–70.

Shlesinger, Miriam. 1998. "Corpus-based Interpreting Studies as an Offshoot of Corpus-Based Translation Studies." *Meta* 43(4): 486.

Tang, Fang and Dechao Li. 2017. "A Corpus-based Investigation of Explicitation Patterns between Professional and Student Interpreters in Chinese-English Consecutive Interpreting." *The Interpreter and Translator Trainer* 11(4): 373–95.

Tao, Youlan 陶友兰. 2010. "Jiyu yuliaoku de fanyi zhuanye kouyi jiaocai jianshe" 基于语料库的翻译专业口译教材建设 [On the Making of Corpus-based Interpretation Textbooks for Translation Majors]. *Waiyujie* 外语界 [Foreign Language World] (4): 2–8.

Wadensjö, Cecilia. 1998/2014. *Interpreting as Interaction.* London: Routledge.

Wang, Binhua 王斌华. 2012. "Yuliaoku kouyi yanjiu – kouyi chanpin yanjiu fangfa de tupo" 语料库口译研究 – 口译产品研究方法的突破 [Corpus-based Interpreting Studies – A Breakthrough in the Research of Interpreting Products]. *Zhongguo waiyu* 中国外语 [*Foreign Languages in China*] 9(3): 94.

Wang, Binhua 王斌华, and Qin Hongwu 秦洪武. 2015. "Hanying kouyi mubiaoyu jiaoji guifan de miaoshu yanjiu – jiyu xianchang kouyi yuliaoku zhong zengbuxing pianyi de fenxi" 汉英口译目标语交际规范的描写研究 – 基于现场口译语料库中增补性偏移的分析 [Describing the Target-language Communication Norms in Chinese-English Interpreting]. *Waiyu jiaoxue yu yanjiu* 外语教学与研究 [Foreign Language Teaching and Research] 47(04): 597–610+641.

Wang, Binhua and Fang Tang. 2020. "Corpus-based Interpreting Studies in China: Overview and Prospects." In *Corpus-based Translation and Interpreting Studies in Chinese Contexts: Present and Future*, edited by Kaibao Hu and Kyung Hye Kim, 61–87. Londonn: Palgrave Macmillan.

Wang, Binhua 王斌华, and Ye Liang 叶亮. 2009. "Mianxiang jiaoxue de kouyi yuliaoku jianshe: lilun yu shijian" 面向教学的口译语料库建设:理论与实践 [Constructing a Corpus for Interpreting Teaching: Theory and Practice]. *Waiyujie* 外语界 [Foreign Language World] (2): 23–32.

Wang, Jiayi 王家义, Defeng Li 李德凤, and Liqing Li 李丽青. 2019. "Xuexizhe kouyi chanchu zhong de tingdun – yixiang jiyu zhongguo daxuesheng kouyi yuliaoku de yanjiu" 学习者口译产出中的停顿 – 项基于中国大学生口译语料库的研究 [A Study of Pauses in EFL Learner's Interpreting Based on PACCEL-S Corpus]. *Waiyu jiaoxue* 外语教学 [Foreign Language Education] 40(5): 78–83.

Wang, Kefei 王克非 and Fu Rongbo 符荣波. 2020. "Yuliaoku kouyi yanjiu: jinzhan yu zouxiang" 语料库口译研究：进展与走向 [Corpus-based Interpreting Studies: A Methodological Review and Preview]. *Zhongguo fanyi* 中国翻译 [Chinese Translators Journal] 41(6): 13–20, 190.

Wang, Li 王丽 and Li Tao 李桃. 2015. "Jiyu yuliaoku de hanying huiyi mohu xianzhiyu kouyi yanjiu" 基于语料库的汉英会议模糊限制语口译研究 [A Corpus-based Study

on Hedges in Chinese-English Conference Interpreting]. *Zhongguo fanyi* 中国翻译 [Chinese Translators Journal] 36(5): 96–100.

Wang, Wenyu 王文宇 and Huang Ye黄燕. 2011. "Yingyu zhuanye gaonianji xuesheng hanying kouyi zhong de yukuai shiyong yanjiu" 英语专业高年级学生汉英口译中的语块使用研究 [Investigating the Use of Chunks in Chinese-English Interpretation by College English Majors]. *Waiyu yu waiyu jiaoxue* 外语与外语教学 [Foreign Languages and Their Teaching] (5): 73–7, 82.

Wen, Qiufang 文秋芳 and Wang Jinquan 王金铨. 2008. *Zhongguo daxuesheng yinghan hanying koubiyi yuliaoku* 中国大学生英汉汉英口笔译语料库 [Parallel Corpus of Chinese EFL Learners]. Beijing 北京: Waiyu jiaoxue yu yanjiu chuban she 外语教学与研究出版社 [Foreign Language Teaching and Research Press].

Xu, Ran. 2018a. "Corpus-Based Terminological Preparation for Simultaneous Interpreting." *Interpreting* 20(1): 29–58.

Xu, Ran. 2018b. "Jiyu yuliaoku jishu de kouyi yiqian zhunbei moshi jiangou" 基于语料库技术的口译译前准备模式建构 [Construction of Interpreting Pre-translation Preparation Model Based on Corpus Technology]. *Zhongguo fanyi* 中国翻译 [Chinese Translators Journal] 39(3): 53–9.

Zhang, Wei 张威. 2009. "Kouyi yuliaoku de kaifa yu jianshe: lilun yu shijian de ruogan wenti" 口译语料库的开发与建设:理论与实践的若干问题 [Interpreting Corpus: Some Theoretical and Practical Issues]. *Zhongguo fanyi* 中国翻译 [Chinese Translators Journal] 30(3): 54–9, 96.

Zhang, Wei 张威. 2012. "Jin shinian lai kouyi yuliaoku yanjiu xianzhuang ji fazhan qushi"近十年来口译语料库研究现状及发展趋势 [Interpreting Corpus and Relevant Researches in the Last Decade: Present Conditions and Oncoming Trends]. *Zhejiang daxue xuebao (Renwen shehui kexue ban)* 浙江大学学报(人文社会科学版) [Journal of Zhejiang University (Humanities and Social Sciences)] 42(02): 193–205.

Zhang, Wei 张威. 2013. "Kouyi yuliaoku de yuanze yu fangfa" 口译语料库研究的原则与方法 [Corpus-related Interpreting Studies: Principles and Approaches]. *Waiyu dianhua jiaoxue* 外语电化教学 [Technology Enhanced Foreign Language Education] (149): 63–8.

Zhang, Wei 张威. 2015a. "Zhongguo kouyi xuexizhe yuliaoku de fuyuyan bioazhu: biaozhun yu chengxu" 中国口译学习者语料库的副语言标注：标准与程序 [Tagging of Paralanguage in CILC: Standard and Procedure]. *Waiyu dianhua jiaoxue* 外语电化教学 [Technology Enhanced Foreign Language Education] (161): 23–30.

Zhang, Wei 张威. 2015b. "Zhongguo kouyi xuexizhe yuliaoku de kouyi celue biaozhu: fangfa yu yiyi" 中国口译学习者语料库的口译策略标注：方法与意义 [Tagging of Interpreting Strategies in CILC: Method and Significance]. *Waiguoyu* 外国语 [Journal of Foreign Languages] 38(5): 63–73.

Zhang, Wei 张威. 2017. "Zhongguo kouyi xuexizhe yuliaoku jianshe yu yanjiu: lilun yu Shijian de ruogan sikao" 中国口译学习者语料库建设与研究：理论与实践的若干思考 [Construction and Research of Chinese Interpreting Learner Corpus: Some Reflections on Theory and Practice]. *Zhongguo fanyi* 中国翻译 [Chinese Translators Journal] 38(1): 53–60.

Zhang, Wei 张威. 2019. "Zhongguo kouyi xuexizhe yuliaoku de yuyan xinxi biaozhu: celue ji fenxi" 中国口译学习者语料库的语言信息标注：策略及分析 [Linguistic Information Tagging in CILC: Strategies and Analysis]. *Waiguoyu* 外国语 [Journal of Foreign Languages] 42(1): 83–93.