Corpus-based Interpreting Studies: Challenges and Prospects

Mianmian Cai

College English Department, Xiamen University Tan Kah Kee College, Zhangzhou, Fujian 363105, China

zpfgordon@126.com

Abstract

Over the past two decades, corpus-based interpreting studies have been gaining momentum and have produced fruitful results. This paper reviews the early work in corpus-based interpreting studies at home and abroad and analyzes the obstacles and challenges which are involved in setting up interpreting corpora. This paper also attempts to explore the future prospects of corpus-based interpreting studies and proposes that corpus-based interpreting studies should consider comprehensive utilization of multiple research methodologies and focus on such key issues as interpreting discourse features, interpreting practice strategy, verification of interpreting theories and concepts, in an effort to enhance the development of interpreting studies and teaching.

Keywords

Corpus-based interpreting studies, Challenges, Prospects.

1. Introduction

Interpreting, as a special form of translation, is a much more complicated language creation behavior than written translation. It involves not only speakers, interpreters and target audience but also many variables, including interpreting scenario, context and subjective factors. Over the past decades, interpreting studies have transformed from prescriptive study to descriptive study, which is, to a great extent, attributed to the corpus-based translation studies in the 1990s. Corpus-based interpreting studies have ushered in a new paradigm of interpreting studies and quickly become a flourishing field of study. However, there are a lot of problems that need to be discussed, analyzed and dealt with. This paper briefly discusses the main methodological issues to be taken into account when creating an interpreting corpus, reviews the first pioneering attempts and aims to draw people's attention to the relevant studies on interpreting corpora establishment and application in the hope of further promoting the actual work of interpreting corpora and further enhance the development of corpus-based interpreting studies. Organization of the Text

2. Early Work in Corpus-based Interpreting Studies at Home and abroad

The year 1993 witnessed the beginning of corpus-based translation studies with the publication of Mona Baker's paper Corpus Linguistics and Translation Studies: Implication and Application. In that paper, she applied corpus linguistics methods and techniques to translation studies and elaborated on corpora's theoretical value, practical significance and research approach in translation studies [1]. Baker predicted that "the availability of large corpora of both original and translated text, together with the development of a corpus-driven methodology will enable scholars to uncover the nature of translated texts as a mediated communicative event."[2] Those contributions gave insight into some of the key aspects of corpus-based translation studies. Since then, a great number of significant and fruitful papers have been published, the majority of which focused on studies conducted on corpora made of written source and target texts, such as advantages and disadvantages of parallel corpora, the potential benefits of using corpora in translator training, etc. At present, only a few relatively mature corpora have been set up: Simultaneous Interpretation Database of Nagoya University is the largest simultaneous interpretation corpus, with altogether 182 hours of interpreting audio clips and

approximately 1 million transcribed interpreting materials. The European Parliament Interpreting Corpus (EPIC), a multilingual parallel corpus including English, Italian and Spanish, has a capacity of round 180,000 words. Other interpreting corpora under construction also include DIRSI (based on international health conferences, about 20 hours of audio clips), K2 (concerning environmental protection topic, with a capacity of 35,000 words), FOOTIE (based on the 2018 European Football championship press conference), etc.

However, as a matter of fact, only a few papers addressed the application of a corpus-driven methodology to the study of interpreting. What is interesting to point out is that since the beginning of corpus linguistics, the development of corpus-based translations studies has been more advanced than the development of corpus-based interpreting studies. There is still a considerable bap between the two, both in terms of corpus size and availability and in terms of number of studies and pedagogical applications.

Likewise, in China, corpus-based interpreting studies are less developed than studies on written language and experts have focused their attention more on interpreting process. It was not until 2007 that relevant literature on interpreting corpus establishment and studies began to be published. In 2011, Hu Kaibao divided China's corpus-based translation and interpretation studies into two stages: introduction stage (from 1999 to 2004) and rapid development stage (since 2005)[3]. In the first stage, emphasis was put on the introduction of foreign corpora, overview of research subject and methodology, etc., with a few empirical studies. Then, in the second stage, corpus-based translation and interpretation studies maintain good growth momentum, with an increasing number of empirical studies.

However, due to the fact that China has just started research in this field since the end of the 20th century, China's theoretical and empirical CIS is still at its exploratory stage. Zhang Wei [4] elaborated on the guiding principles to be upheld and the precautions to be taken in the construction of the interpreting corpus and Chinese Interpreting Learners Corpus, including linguistic information tagging. Later, CIS methodology was put forward by Liu Jian and Hu Kaibao [5]. Generally speaking, the empirical studies of interpreting mainly focus on studying the interpreting process.

3. Challenges in Corpus-based Interpreting Studies

As mentioned in the first part, we can still see a wide gap between corpus-based translation studies and corpus-based interpreting studies in terms of corpus size and availability as well as pedagogical applications, i.e. the latter is less advanced than the former. The reasons for this considerable difference are elaborated as follows:

3.1 Corpus Collection and Transcription

Setting up electronic corpora of transcribed speech events is extremely time-consuming. It is difficult to obtain consent and collaboration from interpreters, conference organizers. Obviously, this obstacle makes it harder to collect enough large samples of interpreting data. Shlesinger points out that, given the complexity of the interpreting process, as many variables as possible must be controlled to obtain reliable results. The first variable is the type of interpreter-mediated event because interpreting in a medical conference is not the same as in court. The type of event determines participants and their roles, as well as the interpreter's role, and therefore has an impact on the interpreting service. Besides, interpreting mode should also be factored in, since consecutive, simultaneous and liaison interpreting all have their own specific characteristics. Speakers' public speaking experience, language skills, accent, etc. should also be taken into account. Last but not least, the target audience is important as well, since the expectations of a small gathering of experts are different from those of the general public in a popular science lecture. [6]

In addition, compared with written texts, the recording and transcription of interpreting data is overwhelming and highly labor intensive. Interpreting is a rather complicated social communication activity. Researchers are confronted with the problem of deciding how to transcribe the data. They

must take into account the aims of study, the data to transcribe, the conventions to use, how to encode the data to make automatic or semi-automated analysis possible, etc. [7]

Any way of transcription will inevitably only describe characteristics of certain aspects, not all. Transcription itself is also a choice. Cencini points out that the feasibility of an interpreting study depends on the characteristics presented during the transcription process. If some characteristics that the research is interested in are not included in the transcription, then the corpus will be rendered useless. [8] Due to the fact that transcription cannot include all characteristics of the interpreting behavior, researchers usually select a specific transcription system based on their interest. Consequently, the observation and description of the interpreting behavior cannot be fully objective.

So far, there are only approximately 10 sizable interpreting corpora at home and abroad. Besides, the current corpora are generally of small size, with the storage capacity of around 300,000 words, which is far from sufficient compared with translation corpora whose storage capacity is usually over one million words [9]. Therefore, the representativeness of interpreting corpora is limited.

3.2 Particularities of Corpus Annotation

The particular characteristics of interpreting corpora pose special difficulties and challenges to annotation of interpreting corpora.

3.2.1 Discourse time segmentation.

During the interpreting process, there is usually certain time lag between the speaker's speech and the interpreter's l language expression, also known as EVS (Ear Voice Span). EVS can directly reflect interpreter's familiarity of the subject, interpreter's cognitive competency, etc. In addition, examining the correspondence in language forms and semantic meanings between the source language and the target language can help determine the units of conversion of interpreted information as well as provide empirical materials for analyzing the "deverbalization" in interpreting. Therefore, to achieve the above-mentioned goals, researchers must accurately segment and annotate the source language corpus and the target language corpus and conduct alignment at different levels according to specific requirements. At present, the span of starting time between the source language and the target language still needs to be done manually, which, to a great extent, affects the accuracy of time segmentation.

3.2.2 Paralanguage tagging.

Paralanguage is the technical term for the voice cues that accompany spoken words. It is concerned with the sound of the voice and the range of meanings that people convey through their voices rather than the words they use. Among the forms of paralanguage, such vocal features as pauses, fillers, drawling, etc. are of great importance during the interpreting process. For example, pauses show that interpreters may have encountered certain obstacles. However, current corpus tagging tools fail to automatically tag these paralanguage forms. Manual work is still needed, which may affect the accuracy and consistency of corpus processing. For example, pauses can only be roughly tagged as "short pause" or "long pause", without accurate calculation of the specific duration of pauses.

3.2.3 Correspondence level between the source language and the target language.

The characteristics of interpreting, simultaneous interpreting in particular, requires interpreters to effectively convey the key information of the source language within a short period of time in the hope of facilitating the communication between the two parties. As a result, word-for-word interpreting or sentence-for-sentence interpreting is not commonly seen during the interpreting activity. Instead, information equivalence and functional equivalence are the prime factors for evaluating interpreting performance [10]. That is to say, word-for-word interpreting or sentence-for-sentence interpreting should not be practiced in interpreting. Besides, to achieve information equivalence, researchers must first decide on the division standard and quantitative indicator of interpreting information units. However, the final conclusion of these problems is yet to be reached. [11]

3.2.4 Research and Development of Corpus Concordance Tools

Corpus concordance tools have evolved into a relatively independent field of study in corpus linguistics [12]. However, as a matter of fact, current corpus concordance tools are mainly common concordance tools used by translation corpora (such as WordSmith, TACT, ACAMRIT, VocabProfile, LEXA, MicroConcord, WORDCRUNCH, WINMAXETC.[13]), which can only fulfill such regular tasks as examining the key words, high-frequency words, lexical density, etc. It is impossible to conduct effective processing of such particular interpreting information as paralanguage, time segmentation, etc. Consequently, the representativeness and objectivity of the description and deduction of interpreting text features are yet to be verified. Success in the research and development of corpus concordance tools requires collaboration of experts in relative fields, including statistics, computer linguistics.

3.3 Sharing of corpus resources

The majority of interpreting corpora are only for use by some research institutions and researchers, not open to all researchers or the public, let alone commercial management. As a result, the influence of corpora is limited and other researchers cannot conduct repetitive study or verification on specific problems, which is to the disadvantage of verifying the universality and representativeness of specific research conclusions [14].

3.4 Limitations in Research methodology

Although corpus-based interpreting studies have introduced innovative research methodology which puts emphasis on the combination of quantitative research methodology and qualitative research methodology, there are also limitations:

The interpreting corpus is the collection of interpreting products. Observation and analysis of the interpreting corpus give priority to interpreting discourse. In this case, the real communication scenario during the interpreting activity as well as interpreters may be neglected. That is to say, corpus-based interpreting studies may cause people to pay too much attention to the interpreting product while giving too little care to interpreters and the specific interpreting process. Pym points out that over-dependence on corpora may cause researchers to concentrate too much on the statistics, particularly when the capacity and type of current interpreting corpora are not satisfactory enough. Over-emphasis on corpora statistics may cause the researchers to be trapped in numbers and consequently cannot be aware of the characteristics of interpreting activity. Tymoczko warned that we must avoid blind worship of quantification. It perplexes a lot of scientific studies and renders them ridiculous and shallow. [15] Statistics itself cannot reveal the deep cognitive mechanism. Researchers need to resort to introspection, induction, interviews, surveys, experiment and other quantitative and qualitative research methods in order to understand such psychological processes as memory allocation, discourse segmentation.

4. Prospects of Corpus-Based Interpreting Studies

What needs to be pointed out is that although corpus-based interpreting studies have made some progress, there remains a lot of work to be done to deepen its interdisciplinary attribute.

4.1 Establishment and improvement of multi-type interpreting corpora

Interpreting corpora should cover a wider range of subjects (politics, science and technology, economics, education, etc.). At present, the collection of interpreting corpora is mainly based on conference interpreting data or students' interpreting practice data, without data from other social context on a large scale. Besides, the structural mode of interpreting corpora should be diversified, including parallel interpreting corpus, comparative interpreting corpus, intermodal interpreting corpus, etc.

4.2 Corpora processing level and annotation standard

As mentioned above, the tagging of paralanguage is of great significance to the corpus-based interpreting studies. Conscious application of paralanguage is an active interpreting strategy, which

not only is helpful to the effective convey of the original information but also reflects interpreters' controllability of context coordination and application ability. On the other hand, improper use of paralanguage is to the disadvantage of the fluency of interpreting expression and affects audience' evaluation of interpreters' self-confidence and performance[16]. However, current corpus tagging tools fail to automatically tag these paralanguage forms. Top priority should be given to the tagging of paralanguage at the next stage of the establishment and research of interpreting corpora. Hopefully, based on the characteristics of interpreting, researchers can develop paralanguage tagging tools and programs so as to deepen and enrich the processing level of interpreting corpora.

As for the alignment level and standard, as mentioned above, interpreters are required to effectively convey the key information of the source language within a short period of time. Due to the special quality and requirements of interpreting in terms of timeliness, the conversion unit of interpreting is inevitably different from that of translation. Therefore, information equivalence and functional equivalence, rather than word-for-word interpreting or sentence-for-sentence interpreting should be regarded as the prime factors for evaluating interpreting performance. Researchers may refer to the relevant conclusions of the information unit in linguistics so as to achieve information equivalence in interpreting corpora by solving the key issues of the division standard and quantitative indicators of the information unit.

We may consider taking the following steps: first, strictly transcribe the source language and the target language corpora to serve as the basis for post-processing information correspondence; second, tag the new information and the old information based on the characteristics of information; third, decide on the subject of the source information and identify certain obvious paragraphs of meaning so as to help determine the integrity of information communication. What needs to be pointed out is that, within a certain period of time, a certain information unit and its content of the source language may be missing in the target language, but that doesn't count as information loss of the source information. We need to determine the importance of the lost information in a relatively long paragraph. Only in this way can we conduct a more objective analysis and evaluation of the interpreting quality.

What's more, researchers should also pay close attention to adding such information as time tags as well as achieving real-time alignment between transcribed corpora and audio and video files so as to establish a multimodal interpreting corpus. On the basis of concordance results, researchers can observe the real context of the above-mentioned interpreting characteristics and study such nonverbal characteristics as rhythm.

4.3 Research Methodology

In terms of methodology, taking into account the descriptive feature of corpora, no corpus is large enough to cover all actual language applications. Therefore, corpus-based analysis should be regarded as a supplement to other more traditional methods rather than as the only correct way. [17] Only by combining positivism and rationalism (namely intuitive judgment and rational analysis of linguistic phenomenon) can it be possible for us to reveal the essence of linguistic phenomenon. Therefore, as a innovative research tool, interpreting corpora can broaden our horizon but could never completely replace other interpreting research methods (introspection, observation, survey, experiment, etc.)

For this reason, it is recommended to conduct interdisciplinary study on corpus-based interpreting studies. We can combine corpus linguistics with not only interpreting studies but also other subjects, such as cognitive psychology, sociology, cultural study, etc. Interpreting corpus can provide language application examples, observation of the interpreting process as well as relevant statistics. At present, most studies are carried out from the perspective of linguistics. In the future, we may strive to refer to social and cultural context at a macro level or draw lessons from psychological disciplines at a micro level.

In terms of its characteristics, corpus linguistics falls into the category of descriptive linguistics. The ultimate purpose of language studies is to understand the production process and practical application of languages. As a result, on the one hand, corpus-based interpreting studies should conduct empirical

studies on interpreting theories, concepts, terms, methods and strategies, examine the specific application of relevant theories and concepts to interpreting practice as well as verify its rationality and applicability in the specific context. On the other hand, we should reasonably make use of corpus statistics and probability analysis to summarize such problems as interpreting textual features, interpreting operating mechanism and interpreting conversion strategy in the hope of exploring relevant rules and characteristics.

It is without doubt that corpus-based interpreting studies should make full use of the advantage of corpus design and application: a detailed and systematic analysis and statistics of the text linguistic features. This will provide systematic and comprehensive statistics support for objectively understand the linguistic features of interpreting activities, especially for accurately determining the interpreting effect and ultimately improve the quality of interpreting teaching and practice. For this reason, the discourse feature analysis of interpreting text may mainly consist of the following aspects: cohesion form and features of the interpreting text, semantic key words analysis of the interpreting text, repetitive verification of the universality of the interpreting text, text type features and intertextuality characteristics of interpreting.

In addition, when it comes to analysis of the interpreting practice strategy, it is of great application value for interpreting teaching and practice to use the text contrast analysis of the interpreting corpus to conduct an objective description of the frequency and distribution of such interpreting strategies as prediction, omission, explanation, fillers, etc. in interpreting practice. It can also be used for analyzing the interpreting cognitive mechanism.

Last but not least, in the future, corpus-based interpreting studies is expected to bring about profound changes in interpreter training and interpreting teaching which involve teaching and training of interpreting sills, language conversion rules and other knowledge. For a long time, interpreting teaching has mainly followed the traditional way of teaching, with old-fashioned teaching methods and content of courses. However, with the development of interpreting corpora, especially those with multimedia database and transcribed text, it is promising to innovate the traditional ways of interpreter training and interpreting teaching. Such innovation will be demonstrated not only in the selection of interpreting textbooks and training materials but also in the interpreting teaching method and evaluation of interpreting training quality. Furthermore, interpreting corpora which contain multimedia materials can also be sued by student interpreters for practicing their pronunciation in an effort to improve their pronunciation quality in future interpreting activities.

5. Conclusion

This paper reviews the progress of corpus-based interpreting studies made in the fields of corpora establishment, interpreter's style, interpreting teaching, interpreter training, etc. It is undeniable that over the past two decades, corpus-based interpreting studies have produced substantial research findings. While acknowledging these efforts, this paper also points out the existing problems and obstacles in current corpus-based interpreting studies, including problems at the technical level and limitations in research methodology. Despite that, corpus-based interpreting studies have propelled the innovation of interpreting research methodologies as well as widened and deepened interpreting studies. Top priorities should be given to the following aspects: On the one hand, we must enlarge the capacity of interpreting corpora, enrich the types of corpora, deepen the corpora processing level and improve the accuracy of corpora processing so as to enhance the representativeness of corpora; on the other hand, we should also broaden our horizons of corpus-based interpreting studies, apply multiple research methodologies for comprehensive examination and analysis, reinforce interdisciplinary cooperation and communication in an effort to improve the quality of corpus-based interpreting studies. Only in this way can we improve the quality of teaching and research of interpreting corpora. It should be acknowledged that a diversified theoretical system, an open research horizon and comprehensive research fields are the main features of translation studies. [18] As a result, it should be given more attention and support to combine corpus linguistics and interpreting activities to create interpreting corpora with multiple usages (teaching and studies). This not only complies

with the major development trend of translation studies but also should be considered as a new issue in corpora establishment as well as application.

All in all, as far as the future work is concerned, corpus-based interpreting studies should be able to develop towards the direction of establishing a universal interpreting corpus, conduct effective interdisciplinary research as well as a combination of theoretical exploration and practice. It is safe to say that corpus-based interpreting studies boasts promising prospects but it also needs collaboration and cooperation of researchers from other disciplines and fields.

References

[1] Xu Mingwu, Zhao Chunlong, The Name and Nature of China's Corpus-based Translation Studies [J]. Shanghai Journal of Translation, 2018 (4): 3-8.
[2] Mona Baker, Corpora in Translation Studies. An Overview and Suggestions for Future Research [J]. Target, 1995(Vol. 7:2): 223–243.
[3] Hu Kaibao. Introduction to Corpus-based Translation Studies[M]. Shanghai: Shanghai Jiaotong University Publishing House, 2011.
[4] Zhang Wei. Interpreting Corpus: Some Theoretical and Practical Issues [J]. Chinese Translators Journal, 2009(3): 54-59.
[5] Liu Jian, Hu Kaibao. The Compilation and Use of Multimodal Interpreting Corpora [J]. Foreign Languages in China, 2015(5): 77-85.
[6] Shlesinger, M. Corpus-based Interpreting Studies as an offshoot of Corpus-based Translation Studies [J]. Meta, 1998(4): 43.
[7] Armstrong, S. Corpus-based Methods for NLP and Translation Studies [J]. Interpreting, 1997 (2): 1-2.
[8] Cencini, M. On the Importance of an Encoding Standard for Corpus-based Interpreting Studies [J.]. inTRAlinea, Special Issue: CULT2K. <available: http://www. intralinea. org/specials /article/1678>
[9] Pan Feng, Hu Kaibao. Corpus-based Interpreting Studies: Problems and Prospects [J]. Language and Translation. 1995(2):57-58.
[10] Liu Heping, Theoretical Thinking on a Unified Teaching Syllabus of Interpreting Teaching [J]. China Translation, 2002(3):56-58.
[12] Yang Huizhong, An Introduction to Corpus Linguistics [M]. Shanghai: Shanghai Foreign Language Education Press, 2002.
[13] He Anping, Corpus Linguistics and English Teaching [M]. Beijing: Foreign Language Teaching and Research Press, 2004.
[14] Zhang Wei, Interpreting Corpus and Relevant Researches in the Last Decade: Present Conditions and Oncoming Trends [J], Journal of Zhejiang University (Humanities and Social sciences). 2012 (42/2): 197-198.
[15] Tymoczko, M., Computerized Corpora and the Future of Translation Studies [J]. Meta. 1998 (43/4):652-660.
[16] Zhang Wei, Tagging of Paralanguage in CILC: Standard and Procedure [J]. Technology Enhanced Foreign Language Education. 2015(161): 27-28.
[17] Biber, Douglas, Conrad, Susan and Reppen, Randi. Corpus Linguistics [M]. Beijing: Foreign Language Teaching and Research Press, 2000.
[18] Yang Ping, Reflection on Translation Studies in Contemporary China [J]. Chinese Translators Journal. 2004 (1): 1-3