Advances in Social Science, Education and Humanities Research, volume 490
Education, Language and Inter-cultural Communication (ELIC 2020)

A Study on the Construction of English-Chinese Consecutive Interpreting Corpus

Xian Shao[1,*]

[1]Shanghai International Studies University, Shanghai, China
[*]*Corresponding author. Email: shaoxian@shisu.edu.cn*

ABSTRACT

In the late 1990s, corpus linguistics began to serve as an important research method in interpreting studies. But after twenty-year development, corpus-based interpreting studies remain at an infant stage compared with corpus-based translation studies, mainly due to the limitation of availability of interpreting corpora. This thesis begins with a brief review of corpus linguistics and the introduction of some major interpreting corpora; then it focuses on the corpus design, representativeness; at last it discusses the detailed steps of corpus construction with an example of a self-built Chinese-English consecutive interpreting corpus of professional interpreters.

Keywords: *interpreting corpus, construction, corpus design, representativeness*

I. INTRODUCTION

Interpreting studies began in the 1950s with the advent of a large number of international organizations. At that time, interpreting studies were mainly qualitatively based on researcher's individual intuition about some individual cases and subjective judgments. In their studies, the materials used are relatively few or elicited from subjects for specific purpose, thus the results gained can hardly be said to be objective or scientific. In the 1980s, the limitation and drawbacks of this kind of research began to be recognized by more and more scholars. In the late 1990s, corpus-based studies began to be adopted by scholars in the fields of interpreting studies, breaking bottlenecks of the small-scale data in empirical studies. Since then, corpus-based interpreting studies have remained at a relatively infant stage compared with corpus-based translation studies and remained in the fields of summarizing the features of some certain words or studying student's interpretation. It's mainly due to the limitation of quantity and availability of interpreting corpus (only one interpreting corpus about Chinese to English interpretation is available). In addition, because of the nature of spoken language, the construction of interpreting corpus is much more complex than translation corpus, and is also more complex than spoken corpus due to the nature of interpretation itself. Therefore, the construction of a certain interpreting corpus should be the very first key step for corpus-based interpreting studies.

*Fund: Supported by Research Foundation of China University of Petroleum-Beijing at Karamay (No.XQZX20200020).

II. LITERATURE REVIEW

A. Corpus linguistics

Corpus linguistics is a data-driven methodology for analyzing large quantities of machine-readable running text, which began at the early 1960s when the 'first-generation' of one-million words computer-readable corpora were first created (Shlesinger, 1998). Corpus design and computerized methods of corpus analysis constitute the basic methodology of corpus linguistics, which is an integral part of the definition of this discipline and an essential factor in its development. Corpus can be classified into two broad categories by the initial objectives — general corpora aimed at representing the language as fully as possible and specialized corpora designed for specialized purposes or for specific research objectives to be resolved. A specialized corpus may be designed to compile modern dictionaries, to explore features of stress in spoken English, to examine a particular language variety or dialect, or to study a particular language register, such as ESP (English for Special Purposes), EAP (English for Academic Purposes), ISP (Interpretation for Specific Purposes). It is obvious that interpreting corpus belongs to the second type.

B. Interpreting corpora

In 1998, Mariam Shlesinger analyzed the problems and benefits for applying corpus in interpreting studies. Since then, Corpus-based interpreting studies (CIS) have remained at their infant stage, whereas CTS have produced a considerable amount of research work. The delay in the development of CIS is not surprising. Making interpreting corpus electronically available for

study requires going through a number of stages, some of which are common and easy for CTS (corpus design, classification, markup, tagging, parsing), whereas others are specific and difficult to oral texts, and particularly onerous and time-consuming, such as transcription. All these aspects clearly contribute to the small present volume of interpreting corpora, such as EPIC (The European Parliament Interpreting Corpus), DIRSI (Directionality in Simultaneous Interpreting Corpus), CorIT (Italian Television Interpreting Corpus), FOOTIE (Football in Europe) and Marta Biagini's corpus on court hearings (Sergio and Falbo, 2012).

Corpus-based interpreting studies (CIS) in China began in 2007 when Hu pointed out the trend of using corpus in interpreting studies at the first international conference of corpus and translation studies at Jiao Tong University (Hu, Wu & Tao, 2007). Since then, compared with other international institutes, some interpreting corpora between Chinese and English has been built (Wang, 2012). At present, there are four major interpreting corpora in China, including finished and unfinished (the basic information of interpreting corpus seen in "Table I"). However, PACCEL is the only one available for common researchers.

TABLE I. BASIC INFORMATION OF FOUR INTERPRETING CORPUS IN CHINA

	Interpreting Corpus	**Corpus Size (words/chars)**	**Construction Year**	**Construction Institution**
1	Chinese-English Conference Interpreting Corpus (CECIC) (1988~)	1022179	2007~ Unfinished, Not available	Shanghai Jiao Tong University
2	Parallel Corpus of Chinese EFL Learners — Spoken (PACCEL-S) (2003~2007)	496177	2008 Finished, available for public	Beijing Foreign Studies University
3	Hong Kong Bilingual Interpreting Corpus on Contemporary Social Life (HKBIC)	47366	2010~ Unfinished Not available	Hong Kong Polytech University
4	The Corpus of Chinese-English Interpreting for Premier Press Conference (CEIPPC), (1998~2001)	100000	2012 Finished Not available	Guangdong University of Foreign Studies

The first corpus, Chinese-English Conference Interpreting Corpus (CECIC) was created by Hu Kaibao (2007) and his group in Shanghai Jiaotong University from 2007 (and it's still in progress), which is the largest interpreting corpus in China (about 1022179 words). CECIC is made up of three sub-corpora: the first is Chinese-English parallel sub-corpus of press conference (consecutive interpreting), including the transcription of original speeches and interpretations by professional interpreters at press conferences held by Chinese central government and the State Council from 1988 and the contents refer to the fields of politics, economy, military affairs, diplomacy, etc.; the second is an original English sub-corpus of the United States Government Press Conferences (U.S. GPCs), which includes the transcription of materials downloaded from the website of CNN and the contents are about American domestic and diplomatic policies; and the third is Chinese-English parallel sub-corpus of Chinese Government Work Report (written), whose Chinese materials is obtained from People's Publishing House and its English interpretation is from the website of China Daily. All the material in CECIC has been processed by word segmentation, tagging and sentence alignment of two parallel sub-corpus.

The second interpreting corpus, mature, completed and also the only available to the public, is Parallel Corpus of Chinese EFL learners (PACCEL) created by Wen Qiufang and her group from Beijing Foreign Studies University. Wen and Wang (2008) elaborated on the standards and process of design and construction of PACCEL, and some software for the application research of this corpus is also suggested at the end of their book. According to Wen and Wang (2008), the material of this corpus comes from the interpreting and translation test of college students from 18 universities majoring in English during their third and last year. It is a double lingual (Chinese and English) corpus comprising of two sub-corpuses: Parallel Corpus of Chinese EFL learners — Spoken (PACCEL-S, consecutive interpreting data) and Parallel Corpus of Chinese EFL learners — Written (PACCEL-W, translation data). The corpus has a total word count of 2.1 million and PACCEL-S has a size of 496177 words. All the material in PACCEL-S has been processed by sentence alignment, POS tagging and manual annotation of paralinguistic information, such as '…' is used to represent a short pause and '……' is for a long pause (Wen & Wang, 2008). And PACCEL-S's each transcript features a header containing linguistic and extra-linguistic information about the student, such as the student's sex, grade, score, etc. Li and Wang (2016) deem that PACCEL-S is a representative corpus to study student's interpreting language features.

The third corpus is Hong Kong Bilingual Interpreting Corpus on Contemporary Social Life (HKBIC), which is created by a research group from

Hong Kong Polytech University from 2010. The data of HKBIC is transcribed from the original speeches related to Hong Kong contemporary social life and the interpretation of some speeches. From Li & Wang's study (2012) HKBIC has three sub-corpora: the first one is a comparable corpus, which is transcribed from 17 original English speeches (each lasts about 25 minutes), and has a word account of 47,366 words; the second is a parallel corpus, which includes 10 Chinese simultaneous interpretations (each last about 30 minutes), and has a size of 40,145 words; the third is a parallel interpreting corpus of English learners, which is still on construction. All the data of its first two sub-corpora (comparable corpus and parallel corpus) is obtained from the broadcast database on the website of Hong Kong government and some social activities and discussions held by government agencies. The interpretations are completed by 12 professional interpreters working at the government for many years. And each transcript in its parallel sub-corpus features a header, containing speech subject, interpreted language, background of original speaker (such as education and profession), background of interpreters (such as graduate institutions and interpreter's qualifications), the time and place of interpretation, information about the audience and etc. (Li & Wang, 2012)

The fourth is the Corpus of Chinese-English Interpreting for Premier Press Conference (CEIPPC), which is created by Wang Binhua from Guangdong University of Foreign Studies and completed in 2012. The material of CEIPPC is transcribed from the video- and audio-records of Premier Press Conferences from 1998 to 2011, which has a total word count of about 100000. In general, fourteen records have been transcribed, corresponding to about twenty hours. CEIPPC is a bilingual parallel corpus and has been processed by POS tagging and sentence alignment.

III. CORPUS CONSTRUCTION

In this thesis, I constructed a Chinese-English consecutive interpreting corpus of professional interpreters, named the Corpus of Chinese-English Consecutive Interpreting of Press Conference (CECIPC, consisting 73,383 English words and 100,484 Chinese characters), focus will be placed particularly on the aspects of the principles of corpus design and representativeness and the different steps of construction as well.

A. Corpus design and representativeness

Due to the research purpose of this study, CECIPC aims to be designed as an open corpus, that is, this corpus is open for modification and adding material, which is not only used for this study but also used by future research. Therefore, the principles of design and representativeness of interpreting corpus will be reviewed first, and then the design and representativeness of CECIPC will be stated in detail.

1) Principles of interpreting corpus design: As Sergio and Falbo (2012, p.12) stated in their book, 'the first step when creating an interpretation corpus will be defining parameters for the selection of items, which is ultimately a careful consideration of the representativeness degree the future corpus will display'. In other words, the first phase of the corpus design process is determined by the study's objectives and the kind of value attached to its results. Determining representativeness marks of a given phenomenon requires delimiting the aspects represented, which in this case trace the profile of 'consecutive interpretation'. In an attempt to identify an interpretation 'target population', according to Halverson's (1998) definition referring to the translation field, Falbo (2001) illustrated a tentative set of criteria giving an account of the various aspects of interpretation. Every communicative event requiring the interpreter's presence could be described on the basis of five main macro-factors: interpreter, situational context, mode, language and directionality, type of interaction. Moreover, each macro-factor may be divided into categories, which in turn may contain additional sub-categories. For example, the 'interpreter' macro-factor, could be divided into three categories: professional interpreter, interpretation student or ad hoc interpreter. And each of these categories may contain different sub-categories on the basis of the subjects' age, sex and years of professional experience or training. Similarly, the 'situational context' factor may be broken down into 'real situation' – in its turn divided into 'press conference setting', 'court setting', 'medical setting', 'international organizations' setting', etc. – and 'experimental situation'. The same holds true for mode, language and directionality, and type of interaction.

The potential combination of all categories and sub-categories pertaining to each macro-factor provides a prototypical image of interpretation, as well as a snapshot of its manifestations in the real world. By selecting one combination of categories and sub-categories, it is possible to concentrate on a particular communicative situation.

2) Design and representativeness of CECIPC: The main objective of the construction of CECIPC was to collect a relatively large quantity of authentic consecutive interpreting data of professional interpreters to produce much-needed empirical research on the characteristics of interpreted texts from Chinese to English. All the material of CECIPC includes ten Press Conferences of NPC and CPPCC Sessions held by Chinese Primer or Chinese Foreign Ministry from

2013 to 2017. From those ten press conferences, all the questions and answers in Chinese and their corresponding interpreted versions were transcribed using dedicated software from the video on the website of Xinhua News Agency, and then checked and edited manually following specific conventions. The resulting corpus includes source texts in Chinese and interpreted texts in English. Due to the dialogic feature of press conference and features of consecutive interpreting, transcription of all the materials from ten press conferences were divided into two groups (each has five texts). That is two sub-corpus are set under CECIPC, including CECIPC-P (the Corpus of Chinese-English Consecutive Interpreting of Press Conference for Prime Minister) and CECIPC-F (the Corpus of Chinese-English Consecutive Interpreting of Press Conference for Foreign Minister). According to Shlesinger's (2008) classification, main features of self-built CECIPC are as follows.

- interpreter: professional interpreters working in Chinese government
- situational context: real time
- translation mode: consecutive interpreting (spoken corpus)
- language and directionality: from Chinse to English
- single-genre: dialogue (press conferences)
- parallel and comparable
- open: (Press Conferences of NPC and CPPCC Sessions)
- untagged

B. Construction of CECIPC

The methodology of CECIPC construction is based on the principle of EPIC construction provided by Russo, Bendazzoli, Sandrelli and Spinolo (2012) and the instruction of Liang, Li & Xu (2010). And the process of construction mainly includes three steps: data collection, transcription and corpus annotation.

1) Data collection: creating CECIPC multimedia archive: The original video of ten press conferences on the Xinhuanet.com were converted with Super Video to Audio Converter into ten audios files, lasting about 480 minutes of source discourse and 400 minutes of target interpretation. The recordings of the Chinese questions and answers and their corresponding English interpretation were then digitized by using Cool Edit-Pro (2.0), a sound editor. The chosen format is ".wav", which ensures good audio quality for later transcription and possible future studies of prosodic features (such as interpreting speed, distribution of pauses, hesitations, etc.). With respect to the feature of consecutive interpretation, the speaker is required to pause to allow interpreting and, typically, the speaker will pause after each complete thought. In the next part, a software called Voice Notepad will be mentioned and used for automatic transcription. But after testing the results and effect of transcription for many times, it is found that this software can work well within only 5 minutes and obviously a whole press conference is too long for it to transcribe probably. Its transcription with overtime will lead to many unreadable words. Then for the practical use of Voice Notepad (an online free software used for transcription) and also for future research, each audio file was segmented into many short audio files based on the pause for interpretation. Therefore, for each press conference we obtained one (large) audio file (the original version) and many segmented short audio files (Chinese questions, Chinese answers, and the corresponding English interpretations, respectively). Once the audio of each press conference have been segmented, all the segmentations are saved as individual clips and stored under the large audio file. The resulting CECIPC archive includes audio clips of the Chinese questions, segmented Chinese answers and corresponding English interpretations, and the transcripts of all the texts.

2) Material transcription and clean-up: Thanks to all the advances of modern technology, transcription is still a labor-intensive and arduous process, which poses a major methodological hurdle.

One of the fundamental steps to create CECIPC is transcription, which may entail analyzing the speeches in question, since transcription is a selective process (Shlesinger 1998). Indeed, it is virtually impossible to reproduce all the characteristics of speech in writing, as there are several levels (i.e. linguistic, paralinguistic and extra-linguistic) involved in spoken communication, and each level comprises an infinite number of features, such as pauses, repetitions, prosody, body language and many more (Russo, Bendazzoli and Sndrelli, 2012). Therefore, the guiding principles when transcribing spoken material must be the nature of the material in question and the aim of the research (Armstrong, 1997). In my research, the aim of creating CECIPC was to transcribe a large quantity of source speeches and interpreted speeches to create an electronic corpus (archive) that could be analyzed automatically. Therefore, to avoid unnecessary complexities and to prevent transcription from being too time-consuming, the basic transcription is based on the method provides by Russo, Bendazzoli and Sndrelli (2012) and Liang, Li and Xu (2011) to allowing the adding of further levels of tagging when needed. Needless to say, transcribing spoken material is a

demanding task, requiring significant effort and patience (Cencini, 2002; Meyer, 1998).

In my research, there were three steps for the transcription, of which the first two significantly facilitated and speeded up the transcription process. Firstly, all the original reports of Chinese questions and answers are downloaded from the website of China National Radio. These documents, however, are not truly "verbatim" transcripts of the Chinese question and answers delivered during each press conferences, since they are stylistically revised to eliminate certain features of spoken language (e. g. repetitions, reformulations, unfinished sentences, etc.). For example, a sentence from the website "*今年我们主动加压，加大降耗力度，也就是确定能源消耗强度要下降 3.9%，而去年实际完成是下降 3.7%，这意味着要减少 2.2 亿吨煤炭消耗。*", while its corresponding verbatim form "*像今年，我们主动加压，确定降耗，也就是能源强度降耗的指标要下降 3.9%，去年实际完成是下降 3.7%，这意味着要减少用 2200 万吨煤。*" (materials from the Press Conferences of NPC and CPPCC Sessions for Prime Minister in 2014). Those reports downloaded from the website of China National Radio, nevertheless, can provide an extremely useful basis for the final transcripts. In other words, I used the original reports as a first draft for my transcripts (source texts). During the transcription, the large amount of Chinese fillers (such as 嗯，啊，呃，呢) will be omitted for the sake of convenient transcription.

Secondly, as regards the interpreted English utterances (target language), there is no written record of the consecutive interpreted versions with high quality available online. Therefore, one speech recognition software (Voice Notepad), free and available on online, was used to speed up the transcription process. Since the interpreted English utterances have already been segmented into many small parts, Voice Notepad can do the transcription with a relatively good quality that can serve as drafts for manual revise at a later stage. Thirdly, as regards all the drafts, including Chinese versions and interpreted English versions, transcription and collation are conducted manually and verbatim for three times. During this process, repeats (in Chinese drafts) were added, mistakes were corrected, pauses are inserted, and headers are compiled for each transcript. All the transcription conventions are summarized in the following "Table II" and described in detail in the following part:

TABLE II. CECIPC TRANSCRIPTION CONVENTIONS

Language Feature	Example (Utterance)	Transcription Convention
Truncated words	pu punish	*<pu-> punish*
Mispronounced words	alto also want to Minister Yang, how do you see how does China see the current international system?	*<alto> also want to* *Minister Yang, <how do you see> how does China see the current international system?*
Pauses	(filled / empty)	*ehm, …*
Numbers Figures Dates	532 4% 1997	*five hundred and thirty-two* *four per cent* *nineteen ninety-seven*
Name of organization	G20	*G-twenty*

All the transcription conventions are described in detail in the following three levels for the purpose of reproducing some characteristics of speech in writing: linguistic, paralinguistic and extra-linguistic level.

As to the linguistic level, I opted for an orthographic transcription and all the words uttered by speakers and interpreters are transcribed. Punctuation signs (such as comma and full stop) in the transcripts were added on the basis of the speaker's intonation and syntactic information available in the sentence. Repeats in the utterance will be verbatim transcribed, and specifically, figures, dates and percentages are fully spelt out.

And the second is paralinguistic level. Based on my research purpose, only a small number of features are annotated, namely truncated words and mispronounced words. Truncated words (i.e. unfinished words) are transcribed with a hyphen (-) at the end and marked between angular brackets (e. g. *<Pre-> President it is a pleasure to be here…*). Mispronounced words and those with an internal truncation are first 'normalized' and then transcribed as they were actually uttered between angular brackets (e.g. *Minister Yang, <how do you see> how does China see the current international system?*).

Pauses are also included in the transcripts, but they are annotated on the basis of personal perception only, that is, they have not been measured by using appropriate electronic tools. Only obvious silent (…) and filled pauses (ehm) are transcribed, but no additional information is provided on their duration (e.g. *Ehm <as> ehm in this process we hope that by cutting overcapacity in those heavy industries, we will ehm bring about a sustained and sound growth of these sectors*). These annotations only serve the purpose of producing user-friendly transcripts that reflect oral data as closely as possible.

The third extra-linguistic level provides the information about file, speaker and interpreter. All of this is recorded in the appropriate fields of a specially-designed header, which comes at the beginning of each transcript and was used to set the parameters to carry out future studies for some specific purpose. And the design of header is open and easy for adding materials to CECIPC in the future. The CECIPC header is made up of a number of fields, including information about file classification, speech, speakers, interpreters and etc. The information contained in the header is a sort of ID card of each transcript and can be useful when querying the corpus. Figure 1 gives an example of the header template that I use for the press conference of NPC & CPPCC sessions for Foreign Minister in 2014:

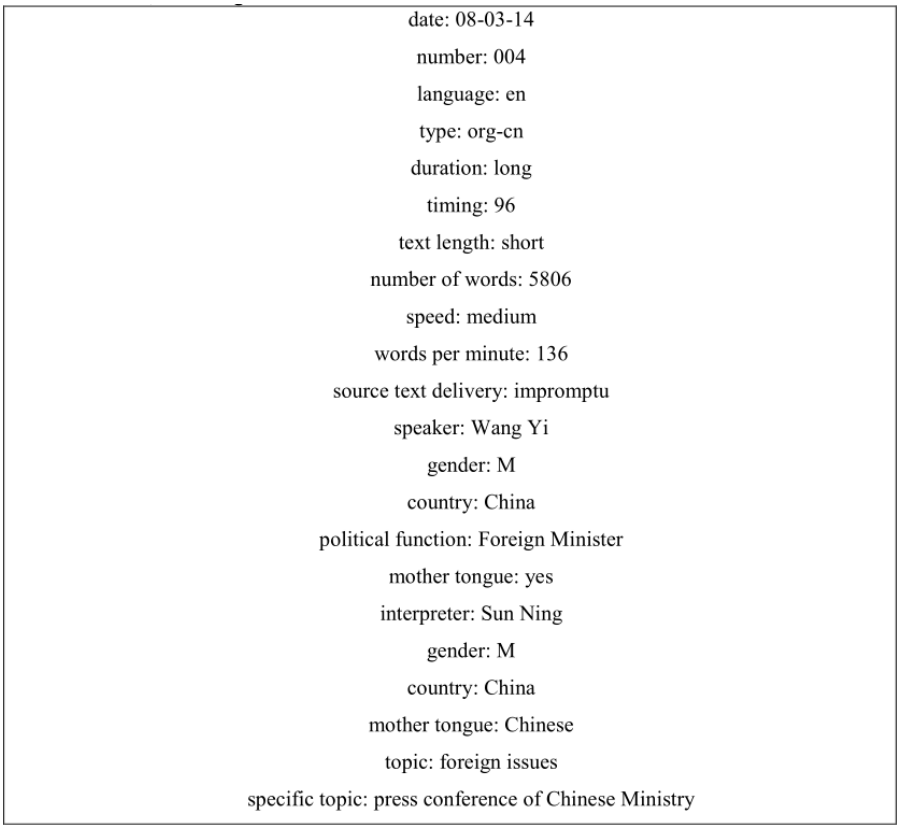

Fig. 1. Header of the press conference of NPC & CPPCC sessions for Foreign Minister in 2014.

In "Fig. 1", the first group of four fields (date, speech number, language and type) is a reference code used to classify the speeches. The first number (08) indicates the day, the second item (03) indicates the month (in this case, February), followed by the year (14, that is, 2014). The letters (m) or (p) tell us whether the speech was delivered during a morning or afternoon sitting (in this particular case, in the morning). The number that follows (in our example 004) is a progressive number we assign to speeches. The abbreviations “en”, “ch” indicate a speech in English or Chinese respectively; “org” and “int” indicate whether it is an original speech (i.e. a source text) or an interpretation (i.e. a target text). If it is an interpreted speech, we indicate both source and target languages, for example “ch-en” means that the speech was interpreted from Chinese into English. This reference code is followed by a number of fields containing information on the speech, namely duration, text length and speed. We have recorded the exact figures indicating duration in seconds (timing), the number of words in the speech and the words per minute (calculated by dividing number of words by duration). We have also classified the duration of speeches as short, medium or long (short: < 120 secs; medium 121-360 secs; long: >360 secs). The same applies to text length, classified as short, medium or long (short: < 300 words; medium 301-1000 words; long > 1000). Speed was classified as low, medium or high (low: < 130 w/m; medium: 131-160 w/m; high: > 160 w/m). For example, as to the material from the press conference of NPC & CPPCC sessions for Foreign Minister, the average speed of interpreter is 136 word/minute, a medium speed which is usually regarded as a proper speed for formal setting. It must be pointed out that these values were calculated on the basis of the present corpus of speeches, and therefore can only be considered representative of this type of material, that is, speeches delivered during a specific group of Prime Minister (or Foreign Minister), each different journalist

and interpreter. At each conference, questions, answers and interpretations are delivered in turn. The utmost length of each utterance in turn is less than 5 minutes, which would be considered short, since questions and answers at press conference normally are not very long compared with speeches at the conferences.

Other speech-related information includes the source text delivery, that is, the utterance mode, classified as read, impromptu, or mixed: the labels refer to whether the speaker is seen to be reading a script, or only reading portions of it, or improvising his/her utterance. The information is recorded in the transcripts of interpreted utterance as well, since it is important to know whether the source utterance was read or delivered "off-the-cuff" when analyzing the interpretation into another language.

Speaker-related fields in the header include: name, gender, country of origin, mother tongue, political function and political group. As this corpus is a parallel corpus, it's very important that the interpreter's information needs to be added, the values are assigned to the fields of name, gender, country and mother tongue (the mother tongue is usually an important element for researching in interpreting).

The labels "Prime Minister" and "Foreign Minister" indicate that the speaker is either a Prime Minister or a Foreign Minister. The last field is the space reserved for comments. As was mentioned above, this space is used to add information. In CECIPC, the types of the setting is provided in the last field (e.g. *specific topic: press conference of Chinese Ministry*)

C. Tools used in corpus construction

All the tools used in this study could be divided into two categories. The first kind of tools are used to construct CECIPC, including Super Video to Audio Converter, Cool Edit-Pro (2.0) and Voice Notepad**:**

1) Super Video to Audio Converter: This software is used to convert the video of press conferences online to audio files needed for my research. It's a software that has an extremely rich set of Output Containers, Video Codecs and Audio Codecs. It encodes and concerts any multimedia file into many different containers using its various internally implemented Video and Audio Codecs.

2) Cool Edit-Pro (2.0): This software is used to cut and edit the audio files obtained from Super Video to Audio Converter. Cool Edit-Pro (2.0) is a digital audio workstation from Adobe systems featuring both a multitrack, non-destructive mix/edit environment and a destructive approach waveform editing view. This software is particularly useful during the process of manual transcription and collation. When the utterance is not clear, no matter in source language or target language, this software can extremely slow down the language for better identification.

3) Voice Notepad: It's a free online software used for automatic transcription of the audio files. Due to its limitations, the audio files are segmented during the stage of material collection into small files with short duration.

IV. CONCLUSION

This thesis starts from reviewing the literature of corpus linguistics and brief introduction of some major interpreting corpora at home and abroad, then if focuses on the details of corpus construction, including corpus design, representativeness, different steps of construction, with an example – CECIPC, which is designed to be a specialized and open corpus. With regards to the different steps of corpus construction, the transcription is still a labor-intensive, arduously demanding task, which poses a major methodological hurdle. However, thanks to all the advances of modern technology and the principles and recommendations provided by the previous studies, the construction of CECIPC has been completed at the present stage and also open for the future modification and material addition. Knowing exactly what is in the corpus, in what proportions, and being able to read whole data and texts is important in providing insights for further corpus exploration, and at the very least, reminds the user that they are looking at real language taken out of its original context. Although the size (more than 70000 words) is relatively small compared with some foreign large interpreting corpora, this specialized corpus can serve as a real and open platform not only for this current study but also for future research with the addition of more real data. Meanwhile, this thesis provides an integrated process of constructing interpreting corpus (multimedia archive).

References

[1] Armstrong, Susan. (1997). Corpus-based methods for NLP and translation studies, Interpreting 2/1-2, 141-162.

[2] Cencini, Marco (2002). On the importance of an encoding standard for corpus-based Interpreting Studies, in CULT2K, Special Issue of inTRAlinea, <http://www.intralinea.it/specials/cult2k/eng_open.php?id=P107>.

[3] Falbo, Caterina. (2001). "Un corpus orale per l'interpretazione", in Margarito, Mariagrazia / Galazzi, Enrica / Lebhar Politi, Monique (eds), Oralità nella parola e nella scrittura / Oralité dans la parole et dans l'écriture, Torino, Edizioni Libreria Cortina, 319-335.

[4] Halverson, Sandra (1998). Translation studies and representative corpora: establishing links between translation corpora, theoretical/descriptive categories and a conception of the object of study, Meta 43/4, 494-514.

[5] Hu Kaibao, Wu Yong & Tao Qin (2007), Trends and Problems in Corpus- based Translation Studies:A Review of 2007

International Conference and Workshop on Corpora and Translation Studies. Journal of Foreign Languages, (5): 64-69.

[6] Liang Maocheng, Li Wenzhong & Xu Jiajin (2010), Using Corpora: A practical Coursebook, Foreign Language Teaching and Research Press.

[7] Li Dechao & Wang Kefei (2012). A corpus-based study on lexical patterns in simultaneous interpreting from Chinese into English. Modern foreign languages (4): 409~415.

[8] Li Yang & Wang Shaoshuang (2016) The Bibliometric Studies on China's Corpus-Based Interpreting Studies. Journal of Beijing International Studies University, (5): 71-83.

[9] Meyer, Bernd. (1998). What transcriptions of authentic discourse can reveal about interpreting, Interpreting 3/1, 65-83.

[10] Russo, Mariachiara / Bendazzoli, Claudio / Sandrelli, Annalisa (2012). Looking for lexical patterns in a trilingual corpus of source and interpreted speeches: extended analysis of EPIC (European Parliament Interpreting Corpus), Forum 4/1, 221-254.

[11] Sergio, Francesco Straniero and Falbo, Caterina (eds) (2012). Breaking Ground in Corpus-based Interpreting Studies, © Peter Lang AG, International Academic Publishers, Bern

[12] Shlesinger, Miriam (1998). Corpus-based Interpreting Studies as an offshoot of corpus-based Translation Studies, Meta 43/4, 486-493.

[13] Wang Jianhua (2012). A Study on Chinese College Students' Use of Lexical Chunks and Quality of Interpreting. Chinese Translators Journal. (2): 47-51.

[14] Wen Qiufang & Wang Jinquan (2008). Parallel Corpus of Chinese EFL Learners. Foreign Language Teaching and Research Press.