Meta
Journal des traducteurs
Translators' Journal

Meta

The Chinese-English Conference Interpreting Corpus: Uses and Limitations

Kaibao Hu et Qing Tao

Volume 58, numéro 3, décembre 2013

URI : https://id.erudit.org/iderudit/1025055ar
DOI : https://doi.org/10.7202/1025055ar

Aller au sommaire du numéro

Éditeur(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (imprimé)
1492-1421 (numérique)

Découvrir la revue

Citer cet article

Hu, K. & Tao, Q. (2013). The Chinese-English Conference Interpreting Corpus: Uses and Limitations. *Meta*, *58*(3), 626–642. https://doi.org/10.7202/1025055ar

Résumé de l'article

Le présent article offre un aperçu de la compilation du corpus constitué par les interprétations des conférences sino-anglophones (CECIC ; selon l'acronyme anglais), suivi d'un exposé des résultats des recherches basées sur les données de ce corpus. Les auteurs soutiennent que les corpus d'interprétations – y compris le CECIC – sont appelés à jouer un rôle croissant dans l'étude des caractéristiques linguistiques des textes interprétés, des normes en matière d'interprétation et des processus cognitifs entourant l'interprétation. Les recherches basées sur le corpus d'interprétations des conférences sino-anglophones montrent que l'emploi de la voix passive, de la conjonction optionnelle that et de la particule infinitive to propres à l'anglais est significativement plus fréquent dans les textes interprétés que dans les textes traduits à partir de rapports provenant du gouvernement chinois et les textes anglais non traduits des conférences de presse. Généralement, les textes interprétés tendent plus fortement vers une normalisation et une explicitation que les textes traduits par écrit. Le présent article aborde également certaines limites intrinsèques aux corpus d'interprétations. Celles-ci sont en grande partie reliées à la difficulté de transcrire les aspects non verbaux de l'acte interprétatif, y compris le ton et les expressions faciales de l'orateur, ainsi que la réaction de l'auditoire. Ces éléments ont un impact visible sur le choix et l'emploi de stratégies et de méthodes par l'interprète, d'où l'intérêt qu'ils représentent dans le cadre d'études sur l'interprétation.

The Chinese-English Conference Interpreting Corpus: Uses and Limitations

KAIBAO HU
Shanghai Jiao Tong University, Shanghai, China
kbhu@sjtu.edu.cn

QING TAO
Shanghai Jiao Tong University, Shanghai, China
taoqing@sjtu.edu.cn

RÉSUMÉ

Le présent article offre un aperçu de la compilation du corpus constitué par les interprétations des conférences sino-anglophones (CECIC ; selon l'acronyme anglais), suivi d'un exposé des résultats des recherches basées sur les données de ce corpus. Les auteurs soutiennent que les corpus d'interprétations – y compris le CECIC – sont appelés à jouer un rôle croissant dans l'étude des caractéristiques linguistiques des textes interprétés, des normes en matière d'interprétation et des processus cognitifs entourant l'interprétation. Les recherches basées sur le corpus d'interprétations des conférences sino-anglophones montrent que l'emploi de la voix passive, de la conjonction optionnelle that et de la particule infinitive to propres à l'anglais est significativement plus fréquent dans les textes interprétés que dans les textes traduits à partir de rapports provenant du gouvernement chinois et les textes anglais non traduits des conférences de presse. Généralement, les textes interprétés tendent plus fortement vers une normalisation et une explicitation que les textes traduits par écrit. Le présent article aborde également certaines limites intrinsèques aux corpus d'interprétations. Celles-ci sont en grande partie reliées à la difficulté de transcrire les aspects non verbaux de l'acte interprétatif, y compris le ton et les expressions faciales de l'orateur, ainsi que la réaction de l'auditoire. Ces éléments ont un impact visible sur le choix et l'emploi de stratégies et de méthodes par l'interprète, d'où l'intérêt qu'ils représentent dans le cadre d'études sur l'interprétation.

ABSTRACT

This paper presents an overview of the compilation of the Chinese-English Conference Interpreting Corpus followed by an outline of research findings based on data obtained from the corpus. It is argued that interpreting corpora, including the Chinese-English Conference Interpreting Corpus, are called to play an increasingly important role in the study of linguistic features of interpreted texts, interpreting norms and the cognitive process of interpreting. Research based on the Chinese-English Conference Interpreting Corpus suggests that the use of English passive construction, optional connective 'that' and the infinitive particle 'to' in interpreted texts is demonstrably more frequent than in the translated English texts of the Chinese government's work reports and the non-translated English texts of press conferences. In a broader sense, interpreted texts exhibit greater tendency towards normalization and explicitation than written translated texts. This paper also touches on the limitations that have been observed while working with interpreting corpora. These limitations are in a large measure related to the difficulty in transcribing nonverbal aspects of the interpreting activity, including the speaker's tone and facial expressions, as well as the audience's facial expressions. These aspects have a clear effect on interpreter's choice/use of interpreting strategies and methods, so they merit careful consideration in interpreting studies.

MOTS-CLÉS/KEYWORDS

interprétation de conférences sino-anglophones, corpus d'interprétations, compilation, emplois, limites

Chinese-English conference interpreting, interpreting corpora, compilation, uses, limitations

1. Introduction

The past decade has witnessed a rapid development in corpus-based translation studies, since a great number of parallel, comparable and translational corpora have been compiled and analyzed to inform research on the features that are particularly typical of translated text (Baroni and Bernardini 2006), translator's style (Baker 2000; Olohan 2003) and translator training (Bowker 2003; Bernardini and Stewart 2007). Although the need to investigate the distinctive features of interpreted texts is just as important, only a few interpreting corpora have been compiled and used in interpreting studies (Russo, Bendazolli and Sandrelli 2006; Shlesinger 2008), and corpus-based interpreting studies remain comparatively under-developed. Against this backdrop, the authors and their colleagues started to compile the Chinese-English Conference Interpreting Corpus (hereinafter referred to as CECIC) in November 2006. The primary goal of this project is to collect an adequate amount of Chinese-English conference interpreting data for the purpose of studying the linguistic features of interpreted texts, interpreting norms and the cognitive process of interpreting.

The present paper describes the design and compilation of CECIC; including its transcription and annotation, with an outline of research findings based on data obtained from the corpus followed by an analysis of the limitations that have been observed while working with interpreting corpora.

2. The Design of CECIC

CECIC is designed to include a unidirectional parallel corpus and a comparable corpus. It is made up of three sub-corpora, namely the Chinese-English Parallel Corpus of Press Conference Interpreting, the English Corpus of Press Conferences and the Chinese-English Parallel Corpus of Chinese Government's Work Report. Specifically, the corpus includes transcriptions of a number of press conferences held by the US and Chinese governments between 1998 and 2008. The government's work reports delivered by the Premier of China from 1997 to 2007 (and their English translations) were included as a sub-corpus to facilitate the comparative study of the linguistic features of translated and interpreted texts. Table 1 shows the structure of the corpus and its present size.

From Table 1, it can be seen that the corpus was designed so as to ensure as much comparability as possible between the target texts in the two parallel corpora and the original English texts. For one thing, both of the parallel sub-corpora use Chinese as the source language and English as the target language. For another, all three sub-corpora have roughly the same time span and the topics under discussion, which cover economy, politics, diplomatic policy and national defense and related issues. Besides, these texts are close enough as measured by total word count. Baker suggests (1995: 234) that a comparable corpus consists of two corpora that "should cover a similar domain, variety of language and time span and be of comparable length." CECIC therefore fits this criterion.

Table 1
The composition of CECIC

Sub-corpus	Source of data	Total word count	% of CECIC
The Chinese-English Parallel Corpus of Press Conference Interpreting	Audio and video recording of the press conferences	229,636 source texts: 133, 431 target texts: 96,205	42.2%
The English Corpus of Press Conferences	Downloaded materials from CNN websites	104,598	19.2%
The Chinese-English Parallel Corpus of Chinese Government's Work Report	Downloaded materials from *China Daily* Websites	209,987 source texts: 100,807 target texts: 109,180	38.6%
Total		**544,211**	**100%**

3. Compilation of CECIC

Admittedly, the compilation of CECIC is a challenging and time-consuming task that involves the following five steps: 1) digitizing video and tape recordings; 2) transcribing the digital video and audio files; 3) editing and word-segmenting the texts, which refers to splitting a sentence into words with one and only one blank space in between; 4) tagging and annotating the corpus; 5) aligning the texts.

3.1. Digitizing video and tape recordings

The Chinese-English conference interpreting data were stored in video and tape recordings. To facilitate the transcribing process, these recordings were converted into digital audio and video files and stored in MP3 format. These digitalized files were then saved as individual clips.

3.2. Transcribing the digital audio and video files

The digital audio and video files selected for CECIC were transcribed orthographically. The authors have tried as much as they could to accurately reproduce the information of the Chinese-English conference interpreting as it is recorded, both at the linguistic and paralinguistic levels.

On the linguistic level, all the words uttered by the speakers and interpreters were transcribed. Punctuation used to signal sentence boundaries was based on the duration of pause, intonation, syntactic function of a word and the relationship between utterance units. For example, a full stop was used after an utterance unit if a pause was long, while a question mark followed an utterance unit ending with a rising intonation. If *well* was used as a discourse marker, a comma was used after the word.

As regards the paralinguistic level, great efforts were made in transcribing truncated words, false starts, filled and unfilled pauses. Unintelligible words were also indicated. These paralinguistic features distinguish interpreting from written translation, and their transcription is quite helpful for investigating the linguistic features and norms of interpreting. However, other nonverbal aspects of the interpreting activity, including the speaker's tone and gestures, as well as the speaker's and the audience's facial expressions, were not transcribed since it was technically

challenging to transcribe them. The transcription conventions for CECIC are shown in Table 2.

TABLE 2
Transcription conventions in CECIC

Paralinguistic information	Transcription conventions
word truncation	ple---
false start	diploma---diplomat
an unfilled pause	…
a filled pause	"er", "mm", "mn," "erm" or "hm"
unintelligible words	*

(1) 李:就中国的情况来说，如果民主，发扬民主，如果进行民---，进一步推动---推进民主政，这是我们的目的。
[As far as China is concerned, it is our objective to promote democracy and to build democracy.]
(CECIC; translated by the authors)

(2) 姚: 只要按照那个去执行的话，就能做到这一点。
[As far---as long as we implement the measures, we'll certainly attain our objective.]
(CECIC; translated by the authors)

In example (1), 民主 (minzhu), the Chinese equivalent for democracy, is partially uttered as 民 (min). The transcription for the truncation is therefore 民---.

In example (2), the symbol --- is used to identify a false start, and "as long as" is the normalized version.

Pauses, a common phenomenon in interpreting, are divisible into unfilled and filled pauses. The former is labeled as "…," while the latter as "er", "mm", "mn," "erm" or "hm," as is shown in Table 2. For example:

(3) 朱：我讲这个话啊，并不是想跟纽约时报那个，那两位作者啊分他的稿费，没有这个意。因为我的这个观点也没有申请专利。
[Well, by mentioning this, I do not intend to ask the co-authors of that article on the New York Times to also share with me … their fees … for that article, because I didn't ask---apply for a patent for my viewpoint in this regard.]
(CECIC; translated by the authors)

In example (3), a pause is marked by the symbol "…," while the symbol "---" is attached to "ask" to show that the word is a false start.

3.3. Editing and word-segmenting the texts

After the texts of the CECIC were converted to a machine-readable format, the next step was to edit them. First, they were to be stored in plain ASCII or UTF-8 text format. Then, EmEditor, a fast Unicode text editor, was used to remove blank lines, spaces, tables, figures, and other unnecessary symbols. This was to ensure that the texts included in the corpus contained neither any type of formatting such as bold, italics, different fonts, nor any graphic elements including figures and pictures.

Unlike an English word, a Chinese word is composed of one or more characters, instead of letters. A Chinese character may represent a syllable, but it does not necessarily constitute a word. In addition, no blank space is inserted between Chinese words. Given the differences between Chinese and English words and the need to conduct statistical analyses on the word level, such as type/token ratio and lexical density, the Chinese texts of CECIC were word-segmented by using ICTCLAS 3.0, a Chinese lexical analyzer developed by the Institute of Computing Technology of Chinese Academy of Sciences.

3.4. Annotating the corpus

The texts of CECIC are annotated in TEI format. The annotations of CECIC comprise head information mark-up, POS tags and paralinguistic information tags.

3.4.1. Head information mark-up

Head information mark-up provides metadata or extra-linguistic information concerning the guest speaker, the time when the press conferences were held, the gender of interpreters and the serial number of each text. Examples of head information mark-ups in CECIC are shown in Table 3.

Table 3
Head information mark-up

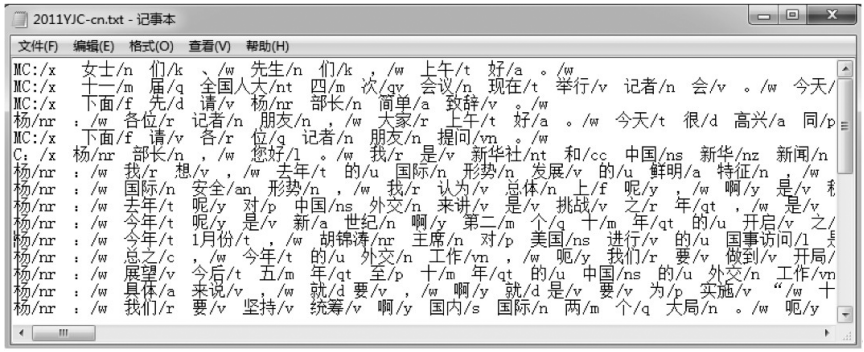
Chinese texts in CECIC	**Interpreted English texts**
<Text_head> <Speaker>Wen Jiabao</Speaker> <Interpreter>Male</Interpreter> <Time>2008</Time> <Id>ent051.txt</Id> </Text_head>	<Text_head> <Speaker>Wen Jiabao</Speaker> <Time>2008</Time> <Id>cht051.txt</Id> </Text_head>

Usually, participants in a press conference include a moderator, a guest speaker, an interpreter and the audience, including a number of correspondents. The guest speaker is expected to expound on policies and viewpoints on behalf of a government or an organization and thus plays a central role. In CECIC; the guest speaker is identified by using the mark-up <speaker>…</speaker>. As a transmitter of message across languages and cultures, the interpreter plays an instrumental role. Consequently, the audio and video recordings of press conferences typically provide no information about an interpreter's identity, except for his or her gender. To allow for comparison of the differences between the features of the texts interpreted by male and female interpreters, the mark-up <interpreter>…</interpreter> was assigned to each of the interpreted texts to indicate the interpreter's gender. In addition, mark-ups <Id>cht051.txt</Id> and <Id>ent051.txt</Id> were used to label the title and serial number of a Chinese text and its interpreted English text, in which "cht" and "ent" are the abbreviations of "Chinese texts " and "English texts," respectively.

3.4.2. POS tags

POS tags were used to mark up the parts of the speech for each token. The English texts were POS tagged by using Treetagger, a program for part-of-speech tagging and lemmatization developed by Helmut Schmid at the University of Stuttgart. For the Chinese texts, ICTCLAS 3.0, which has the dual functions of word-segmenting and POS tagging, was used. The word-segmented and POS tagged Chinese texts are shown in Figure 1.

FIGURE 1
The word-segmented and POS tagged Chinese texts of CECIC



3.4.3. Paralinguistic information tags

Paralinguistic information tags include information about paralinguistic features specific to spoken communication. They primarily involve the tags for pause, word truncation, repetition and revision (Table 4).

TABLE 4
Paralinguistic information tagging for CECIC

Paralinguistic features	Tags for paralinguistic features
Pause	<pause>…</pause>
Word truncation	<truncated>…</truncated>
Repetition	<repetition>…</repetition>
Revision	<revision>…</revision>

As one might expect, tagging of this kind is labor-intensive as it has to be undertaken and checked manually.

3.5. Aligning the texts

Aligning texts of a corpus involves aligning two sets of texts at discourse, paragraph and sentence levels. The former two can be undertaken automatically, but sentential alignment has to be performed partly automatically and partly manually. For the alignment of CECIC; the authors used ParaConc, yielding sentence-level alignment in accordance with the following criteria:

a) An alignment unit in CECIC is one orthographic sentence in the source text and its corresponding version in the target text.
b) The corpus texts are aligned directionally from the source text to the target text, allowing researchers to better understand the interpreter's use of particular strategies and analyze various translations of the same word or expression.
c) Efforts are made to achieve one-to-one correspondence between the sentences in the source and target texts, although one-to-two and one-to-many correspondences are also accepted.
d) A full stop, a question mark, an exclamation mark or a dash constitutes the mark of a sentence.
e) A semi-colon, used to separate longer sentence components, is regarded as the mark of a sentence if one-to-one correspondence is achieved.

4. Uses of CECIC

Electronic corpora have long been awaited in interpreting studies in order to validate the many hypotheses and theories suggested by scholars on interpreters' strategies and the interpreting process (Shlesinger 1998). Most interpreting studies have been qualitative or case studies based on a small amount of data. Few scholars have conducted corpus-based interpreting studies using a wealth of data and statistical measures. But in the past couple of years, a number of interpreting corpora have been compiled to investigate the features of interpreted texts, interpreting norms and interpreting strategies. Based on the data obtained from English-Spanish Interpreting Corpus, Lindquist's study (2004) analyzes interpreting strategies, such as conversion, omission and addition. Russo, Bendazzoli and Sandrelli (2006) outline the compilation of European Parliament Interpreting Corpus (hereinafter referred to as EPIC) and its use in investigating the lexical patterns of interpreted speeches. Using CIAIR Simultaneous Interpreting Corpus, Tohyama and Matsubara (2006) discuss the syntactic operational norms of Japanese-English interpreting based on a comparative analysis of 4,578 Japanese sentences and their English translations.

With the compilation of CECIC, the corpus has been queried to examine the features that are arguably typical of English texts interpreted from Chinese, and investigate whether and how two features of translation, i.e., normalization and explicitation, are reflected in the interpreted texts. Normalization refers to "the tendency to conform to patterns and practices which are typical of the target language, even to the point of exaggerating them" (Baker 1996: 176-177). Explicitation is defined as "the overall tendency to spell things out rather than leave them implicit in translation" (Baker 1996: 180) or to make implicit information explicit and precise. The authors made a statistical analysis of the use of passive construction – a typical grammatical structure in English – in the interpreted texts, the translated ones and their non-translated English counterparts, in an attempt to find out whether and how normalization may occur in the interpreted texts. It is noteworthy that, if explicitation is an inherent feature of translation, translated texts will manifest a higher frequency of the use of optional syntactic elements than non-translated texts. To investigate the tendency towards explicitation in interpreted texts, the authors conducted a comparative analysis of the use of the optional connective "that" and infinitive particle "to" in the aforementioned three texts.

4.1. Passive construction in interpreted texts of press conference interpreting

The passive construction is a syntactic structure in which the subject of a sentence denotes the recipient of an activity rather than the agent. It can be further classified into short and long passive constructions. The short one consists of an auxiliary and a past participle of a transitive verb or a past participle of a transitive verb only. The long one contains an expression introduced by "by," while the short one does not.

Using WordSmith tools, the authors have retrieved the concordances of passive constructions in CECIC; and have conducted a statistical analysis of the occurrences and frequencies of such constructions in the three sub-corpora. Findings indicate that English passive constructions occur with much higher frequency in the interpreted texts than in either the translated ones or the non-translated English counterparts.

As shown in Table 5, the percentage of passive constructions is 1.18% in the interpreted English texts, 2.74 times more than that in the original English texts of press conferences and 1.66 times more than in the translated English texts. It is noteworthy that passive construction is far less frequent than active construction in the Chinese language, since its use is generally linked to something undesirable or unfortunate. It would be logical to assume that passive construction appears less frequently in the interpreted English texts than in the original English texts, as the interpreted English texts are the reproduction of the Chinese outputs in the press conferences. However, contrary to the authors' expectations, the frequency of passive constructions in the interpreted texts is much higher than that in the other two types of texts. Clearly, the interpreted texts manifest a remarkable tendency towards normalization.

TABLE 5

Frequency of passive constructions in CECIC

Passive constructions	**Chinese-English Parallel Corpus of Press Conference Interpreting**	**English Corpus of Press Conferences**	**Chinese-English Parallel Corpus of Chinese Government's Work Report**
Long passive constructions	234	109	123
Short passive constructions	908	341	651
Total passive constructions	1142	450	774
% of passive constructions	**1.18%**	**0.43%**	**0.71%**

To find out what types of Chinese syntactic structures correspond to passive construction in interpreted English texts, the authors began with an analysis of the "to be + past participle" construction, a typical type of passive construction, and its corresponding Chinese constructions. A total of 890 occurrences of the "to be + past participle" construction in the interpreted texts were identified, to which 9 types of Chinese syntactic structures correspond.

Table 6

Chinese syntactic structures translated into the 'to be + past participle' construction

Type	Chinese syntactic structures rendered into the'to be + past participle' construction	Number of instances
A	verb + object (zero subject)	240
B	subject + verb (subject as the recipient)	187
C	subject + verb + object (the subject is neither the recipient nor the agent)	174
D	BEI construction, SHOU construction and YOU construction	88
E	Subject + subjective complement	73
F	DUI construction	73
G	adverbial phrase or verbal phrase	52
H	verbal phrase functioning as attribute	33
I	BA construction, XIANG construction and JIANG construction	26
Total		**890**

As shown in Table 6, types A, B and C structures are likelier than the other types to be translated into the "to be + past participle" construction. The number of instances of the "to be + past participle" constructions translated from type A is the largest, accounting for 27% of all the occurrences of this type of construction in the interpreted texts. 21% and 19.6% of the constructions are correspondent to types B and C structures respectively, whereas the smallest number of the occurrences of the construction are correspondent to type I.

(4) 据初步统计，"十五"期间累计完成通用航空作业飞行33.6万小时，比"九五"期间增长59%，五年平均增长率为11%左右。
[According to preliminary statistics, during the tenth Five-Year Plan period, a total of 336,000 flight hours of general aviation were operated, up 59 percent over the ninth Five-Year Plan period, with an average growth rate of 11 percent during the five years.]

(CECIC; translated by the authors)

(5) 五年中与42个国家签署了新的双边航空运输协定或航权安排，2005年末中国与他国航运输协定达98个。
[In the five years, new bilateral air services arrangements or air traffic rights arrangements have been concluded with 42 countries, and by the end of 2005 a total of 98 bilateral air transport arrangements have been concluded between China and other countries.]

(CECIC; translated by the authors)

In examples (4) and (5), the Chinese sentences are zero-subject sentences or sentences without a subject. Such sentences are quite common in the Chinese language, since the subject of a sentence is often omitted when it is self-evident or uncertain. But in English, every sentence but an imperative sentence must contain a subject.

In translating these zero-subject sentences, an interpreter into English must choose between active and passive constructions. In the case of the former, the interpreter must instantly determine what the subjects are, even when s/he has no information on which to base the choice. Although it sometimes seems fair and safe for the interpreter to add "we" or "our country" as the subject in the translated texts, that would make the translations personal and subjective in violation of interpreter neutrality. If s/he chooses the passive, the interpreter will simply render the object

of the Chinese sentence as the subject in the English translation, which saves more effort than translating it into English active construction. Moreover, this approach leaves the new or important information in the Chinese original more pronounced in the translated English text. Generally speaking, the given information conveyed by a Chinese sentence precedes the new information, as it is transmitted by the subject and predicate (including the object) of a sentence. However, if the object of a Chinese zero-subject is rendered as the English subject, the new information will be placed at the beginning in the English translation, thus highlighted. That may explain why an interpreter is likelier to interpret a Chinese zero-subject sentence into English passive construction in the course of conference interpreting.

In addition, unlike the English subject that represents either the agent or the recipient of an activity, the subject in the Chinese language not only denotes the agent or recipient of an activity, but also introduces the time or place of an activity or the scope affected it. When it represents the time, place or scope, the Chinese subject cannot be translated into an English subject. On the other hand, an object in Chinese is sometimes rendered as a subject in English, with the Chinese sentence being interpreted into the 'to be + past participle' construction.

(6) 当时邓小平同志还在世，在他的支持下，以江泽民同志为核心的党中央决定加强宏观调控，采取了16条措施，其中13条是经济措施。
[At that time Deng Xiaoping was still alive. With his support, and also under the leadership of the CPC Central Committee with Comrade Jiang Zemin at the core, the decision was made to strengthen macro regulation and control. Sixteen measures were adopted, of which 13 were economic measures.]

(CECIC; translated by the authors)

(7) 全行业五年固定资产总投资947亿元。共新增机场21个，改建了一大批机场。
[In the five years, a total investment of 94.7 billion yuan was made in fixed assets in the whole industry. 21 new airports were added and a large number of airports were modified and expanded.]

(CECIC; translated by the authors)

In example (6), the Chinese sentence represents an active construction with the subject denoting the scope of an activity. The lengthy Chinese subject (underlined) conveys given information, i.e., the fact that Comrade Jiang Zemin was at the core of the leadership of the CPC Central Committee. The predicate (italicized) provides new information, i.e., information about the strengthening of macro regulation and the adoption of 16 measures. If the sentence were translated literally into an English active construction, the translation would be lengthy and somewhat awkward. To give prominence to the new information, the interpreter renders the Chinese sentence into two passive constructions, changing the object in Chinese into the subject in English and the subject in Chinese into an adverbial in English accordingly.

In example (7), the subject of the first sentence denotes the scope affected by an activity. The second sentence is a zero-subject sentence featuring a "verb + object" structure. Considering the difference between the Chinese subject and the English subject, the interpreter renders both Chinese sentences into English a passive construction that highlights the important information and allows for a better understanding of his outputs by the audience.

(8) 共有28名省、地（市）、县（区、市）和乡镇党政负责人因此受到党纪处分。
[18 country and township leaders and 8 municipal government leaders in charge of work safety were punished with Party disciplinary and administrative sanctions.]
(CECIC; translated by the authors)

In example (8), the Chinese sentence is an instance of SHOU construction, in which the subject is the recipient of an activity. This construction, as well as the DUI construction, BA construction, XIANG construction and JIANG construction, can be translated into an English passive construction, since the subjects or objects of these constructions represent the recipients of an activity. However, the objects introduced by such prepositions as SHOU, DUI, BA, XIANG and JIANG in these constructions tend to be long and complicated. This makes it difficult for an interpreter to determine whether these objects denote the time or place of an activity or the recipients of an activity. That appears to account for our findings whereby these constructions are not often rendered into the English passive construction in conference interpreting. As shown in Table 6, only 21% of the total occurrences of the "to be + past participle" construction are translations of the above Chinese constructions.

4.2. Optional connective "that" in interpreted texts of press conferences

The optional connective "that" is used to introduce noun clauses or adverbial clauses. However, "that" is often omitted when it is used to introduce noun clauses functioning as the object or adverbial clauses. According to Quirk *et al.* (1985: 1049), "that" is often omitted when used to introduce clauses functioning as the object and adverbial clauses of cause and effect. Rohdenburg (1996) argues that "that" is frequently used to signal the relationship between the main clause and subordinate clause. Olohan and Baker's study (2000) indicates that the optional "that" is far more frequent in the translated English texts of the Translational English Corpus than in the original English texts of the British National Corpus.

To investigate the use of the optional "that," the authors made an analysis of all the concordances of optional 'that' extracted from CECIC; and found that the frequency of optional "that" in English interpreted texts is 13% higher than that of the original English texts and three times that of the translated English texts.

Table 7
Frequency of the optional connective "that"

	Occurrence	Frequency (per 10,000 words)
Interpreted English texts of CECIC	276	60.2
Original English texts of press conferences	170	53.5
English translations of Chinese government's work report	42	20.3

An analysis of the collocations of "verb + that" reveals that "say (says, saying, said) that" and "believe that" are used most frequently in interpreted English texts, while the collocations "make sure that," "say (says, saying, said) that" and "believe that" occur with high frequency in the original English texts. The collocation "ensure that" is the highest in terms of frequency per 10,000 words in the translated English texts, as shown in Table 8.

It is clear that the interpreters tend to use optional connective "that," particularly the collocations of "say (says, saying, said) that" and "believe that" more often. As a matter of fact, the connective "that" functions as a cue as to the syntactic structure and semantic information of what an interpreter is going to talk about, thus facilitating their understanding of his outputs. Therefore, compared to the translated English texts and the original English texts, the interpreted English texts exhibit a more noticeable tendency towards explicitation in the use of the optional connective "that."

Table 8
Frequency of the collocations of 'verb + that'

Collocation	Source of data	Occurrence	Frequency (per 10,000 words)
say (said, saying, says) that	interpreted English texts	40	10.4
believe that	interpreted English text	30	6.6
make sure that	original English texts	28	8.8
say (said, saying, says) that	original English texts	21	6.6
believe that	original English texts	19	5.98
ensure that	translated English texts	25	12.1

4.3. Infinitive particle "to" in the interpreted English texts of press conferences

The infinitive particle "to" is used to introduce infinitive construction. It is usually omitted in the second of two coordinate infinitive constructions. However, for the purpose of explicitating the coordinate relationship between two infinitive constructions, the particle "to" is preserved in the second infinitive construction. For example:

(9) I don't want to take a position on one key player's alleged position and compare it to how somebody else in the administration feels.

(CECIC)

(10) I wondered if you reconsidered the wisdom of placing nominees at the disposal of White House handlers whose jobs seems to be to shave all the rough edges off their positions and to prevent them from saying anything that might be controversial?

(CECIC)

Examples 9 and 10 are retrieved from the English Corpus of Press Conferences of CECIC. To illustrate the point, the authors examined the concordances of the infinitive particle "to" and identified the instances of the particle "to" used to introduce the second infinitive construction. The findings are shown in Table 9.

As shown in Table 9, the frequency of the particle "to" introducing the second infinitive construction in the interpreted texts is 9.19 per 10,000 words, i.e., 2.65 times more than that in the original English texts, and 3.8 times more than that in the translated texts. In the three types of texts, the number of coordinate infinitive constructions is 120, 49, and 47 respectively, and the infinitive particle "to" is used to introduce the second infinitive in 35%, 22% and 11% of these constructions. So it is fair to conclude that the interpreted English texts exhibit heavier use of the infinitive particle "to" to render explicit the coordinate relationship between infinitive constructions.

Table 9
The infinitive particle "to" used to introduce the second infinitive construction in CECIC

Text	Total occurrences of coordinate infinitive constructions	Occurrences of the infinitive particle "to" in the second coordinate infinitive construction	Frequencies of the infinitive particle "to" in the second coordinate infinitive construction (per 10,000 words)
Interpreted English texts	120	42	9.19
Original English texts	49	11	3.46
Translated English texts	47	5	2.42

(11) 我们天天都在看人民来信，怎么满足他们的愿望，实现他们的要求呢？
[We are reading letters from our people every day and we are doing our best to satisfy their needs and to meet their demands.]

(CECIC; translated by the authors)

(12) 现在台湾也有些人是挟洋天子来保护自己，其目的是在于拖延统一，继续分裂祖国，是这么一个问题。
[That is, they are trying to rely on the foreigners so as to protect themselves. And I think their true purpose is to delay the reunification of the country and to continue to perpetuate the state of a division of the motherland.]

(CECIC; translated by the authors)

The infinitive particle "to" in the above two instances is used to introduce the second infinitive constructions, signaling what is going to be discussed next and thus helping the audience understand the interpreter's outputs better. It appears that an interpreter often resorts to the use of the particle to increase explicitness. In contrast, the particle "to" is usually omitted in the second of two coordinate infinitive constructions in written translation.

In addition to the uses discussed above, CECIC was also used to investigate the explicitation of textual meaning (Hu and Tao 2009) and syntactic operational norms in interpreting (Hu 2010).

5. The Limitations of CECIC

As a parallel corpus and a comparable corpus, CECIC; with its advantages in automatic extraction of concordances and statistical analysis, has been used to identify the features typical of interpreted English texts, as well as the tendency towards normalization and explicitation in interpreting based on interpreters' outputs of 96,205 words. This represents a major breakthrough compared to the interpreting research based primarily on the sparse, anecdotal data or the output of students and trainees in the field. However, the drawbacks of CECIC with regard to its size and transcription cannot be disregarded.

5.1. Limitation in size

The size of a corpus affects its representativeness and validity. As it stands, CECIC comprises more than 500,000 tokens, about 100,000 tokens for each of the five text categories, i.e., the source texts and the target texts of the two parallel corpora and

the transcripts of original English press conferences. Given the difficulty in obtaining and transcribing interpreting data, this should be considered an acceptable and reasonable size for current research. In comparison with other corpora of spoken language or corpora of written translations, though, the size of CECIC is not adequate.

What, then, is an adequate size for description of a language or a language variety? Sinclair (1991: 20) suggested that 10-20 million words might constitute "a useful small general corpus," but "will not be adequate for a reliable description of the language as a whole." Kennedy (2000: 68) contends that a corpus of 100,000 words is adequate for the study of prosody, and "a robustly reliable analysis of the use of verb-form morphology can be undertaken on a corpus of half a million words." It is argued that the adequacy of a corpus depends on the purpose to which a corpus is put. "A bigger corpus is not necessarily more useful than a smaller one, particularly when studying high frequency words."(Olohan 2004: 46). According to Zipf's law (Zipf 1949), the tokens of high-frequency word types generally account for a very high percentage of the tokens in a corpus, which is evidenced by the analyses of the data of LOB corpora and BNC. In the 1-million-word LOB Corpora, 100 word types occur more than 1,000 times, whereas 8,000 word types occur more than 1,000 times in the 100-million-word BNC, and they take up about 95 percent of the tokens in the corpus (Kennedy 2000: 68). Thus, for the study of low frequency phenomena such as unusual collocations and hapax legomena, word forms that occur only once in a corpus, very large corpora are necessary. If a corpus is not large enough, some low-frequency words or unusual collocations will be unlikely to occur. But for the investigation of high-frequency phenomena, a corpus of half a million words is needed, since high frequency words or syntactic structures can be well represented in such a corpus. Kennedy (2000: 68) points out that "studies of many syntactic processes and high frequency vocabulary generally require corpora of between half a million and one million words." As mentioned above, CECIC was designed to study the typical features of interpreted English texts, interpreting norms and strategies, which requires an analysis of high frequency words or syntactic structures. Therefore, a size of half a million tokens seems to be a reasonable threshold for an interpreting corpus.

Related to the size of CECIC is the number of the individual samples or texts that make up the corpus. It is argued that the higher the number of individual samples or texts, the greater the reliability of the analysis based on the corpus data, provided that the selection of texts included in the corpus is adequate. The Chinese-English Parallel Corpus of Press Conference Interpreting, one of the three sub-corpora of CECIC; contains 30 Chinese texts and their interpreted English texts, each of about 7,600 words. The corpus is quite satisfactory in size and number of samples compared with the sub-corpora of the EPIC corpus, which consists of nine sub-corpora with a total of 177,295 words. The size of each sub-corpora ranges from 6,000 to 42,000 words, and six sub-corpora include no more than 21 samples each except the other three, which contain 81 samples each. However, greater efforts have to be made to enlarge the corpus size before the validity and reliability of the interpreting research based on CECIC can be improved. Thus, we have been trying to obtain a significant amount of conference interpreting data and increase the size of the Chinese-English Parallel Corpus of Press Conference, which will hopefully grow to 450,000 words in late 2012.

5.2. Limitation in transcription

Transcribing conference interpreting files involves reproducing in plain texts multimodal information conveyed by audio and video files of press conferences, which inevitably implies the loss of certain kinds of paralinguistic information, such as tones and facial expressions, for it is technically challenging to transcribe these nonverbal aspects of the interpreting activity. The difficulty lies not only in the act of transcription, per se, but in the fact that certain elements of spoken communication are both so subtle and so subjective as to defy description (Cook 1995: 51-52; O'Connell *et al*. 1993). Shlesinger (1998) points out that "[w]hile transcription, however laborious, can provide us with a representation of the interpreter's linguistic output, its failure to reflect the concomitant paralinguistic dimensions is a major drawback."

The authors have endeavored to transcribe some of the paralinguistic and kinesic features characteristic of interpreting as a special kind of spoken communication, such as word truncations, false starts and pauses, but other paralinguistic features, including intonation and facial expressions, are not reflected in the transcripts of CECIC. As a result, the application of CECIC to interpreting research has been limited to the features of interpreted English texts that lend themselves to the transcription. To some extent, the failure to represent some of the paralinguistic features affects the validity and reliability of the interpreting studies based on the Chinese-English Parallel Corpus of Press Conference Interpreting.

6. Conclusions

In this paper, the compilation of CECIC is described and explained with an analysis of the study of the use of the passive construction, the optional connective "that" and the infinitive particle "to" based on data obtained from the corpus. Research shows that the passive construction, the optional connective "that" and the infinitive particle "to" all occur with higher frequency in the interpreted English texts than in both the translated English texts and the non-translated English texts of press conferences. Therefore, the interpreted texts exhibit noticeable tendencies towards normalization and explicitation. It is argued that the different roles of Chinese and English subjects in a sentence as well as the interpreter's inclination to highlight new information contribute to the higher frequency of the passive construction used in the interpreted texts. In addition, the optional connective "that" and the infinitive particle "to" are often used in interpreting as discourse markers that facilitate the audience's understanding of what an interpreter is saying and give the interpreter extra time to move on. Notwithstanding the size limitations of CECIC and the difficulty in representing some of the nonverbal aspects, we hope that the findings yielded so far may shed light on some of the salient features of interpreters' outputs and may pave the way for further analyses of the features that distinguish interpreting from other forms of linguistic outputs.

REFERENCES

Baker, Mona (1995): Corpora in Translation Studies: An Overview and Some Suggestions for Future Research. *Target*. 7(2):223-243.

Baker, Mona (1996): Corpora in translation studies: The challenges that lie ahead. *In*: Harold Somers, ed. *Terminology, LSP and Translation: Studies in Language Engineering, in Honour of Juan C. Sager*. Amsterdam/Philadelphia: John Benjamins Publishing Company, 175-186.

Baker, Mona (2000): Towards a Methodology for Investigating the Style of a Literary Translator. *Target*. 12(2):241-266.

Baroni, Macro and Bernardini, Silvia (2006): A New Approach to the Study of Translationese: Machine Learning the Difference between Original and Translated Text. *Literary and Linguistic Computing*. 21(3):259-274.

Bernardini, Silvia and Stewart, Dominic (2007): Corpora in translator education: an introduction. *In*: Federico Zanettin, Silvia Bernardini and Dominic Stewart, eds. *Corpora in Translator Education*. Beijing: Foreign Language Teaching and Research Press, 1-14.

Bowker, Lynne. (2003): Corpus-based applications for translator training: exploring the possibilities. *In*: Sylviane Granger, Jacques Lerot and Stephanie Petch-Tyson, eds. *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*. Beijing: Foreign Language Teaching and Research Press, 169-183

Cook, Guy (1995): Theoretical issues: transcribing the untranscribable. *In*: Geoffrey Leech, ed. *Spoken English on Computer: Transcription, Mark-up and Application*. New York: Longman, 35-53.

Hu, Kaibao and Tao, Qing (2009): A Corpus-based Study of Explicitation of Textual Meaning in Chinese-English Conference Interpreting. *PLA International Studies University Journal*. 32(5):67-73.

Hu, Yiyue (2010): *A Corpus-based study of syntactic operational norms in Chinese-English conference interpreting*. M.A. thesis, unpublished. Shanghai: Shanghai Jiao Tong University.

Kennedy, Graeme (2000): *An Introduction to Corpus Linguistics*. Beijing: Foreign Language Teaching and Research Press.

Lindquist, Peter (2004): Challenging Conventional Wisdom: A Corpus-based Model for Interpreter Performance Evaluation. *The ATA Chronicle*. 38(1):74-82.

O'Connell, Daniel C., Kowal, Sabine J. and Kowal, Sabine (1993): Some Sources of Error in the Transcription of Real Time in Spoken Discourse. *The Jerome Quarterly*. 8(3):3-11.

Olohan, Maeve (2003): How Frequent are the Contractions? A Study of Contracted Forms in the Translational English Corpus. *Target*. 15(1):59-89.

Olohan, Maeve (2004): *Introducing Corpora in Translation Studies*. London/New York: Routledge.

Olohan, Maeve and Baker, Mona (2000): Reporting 'that' in Translated English: Evidence from Subconscious Processes of Explicitation? *Across Languages and Cultures*. 1(2):141-158.

Quirk, Randolph, Greenbaum, Sidney, Leech, Geoffrey, Svartvik, Jan, *et al*. (1985): *A Comprehensive Grammar of the English Language*. London: Longman.

Rohdenburg, Gunter (1996): Cognitive Complexity and Increased Grammatical Explicitness in English. *Cognitive Linguistics*. 7(2):149-182.

Russo, Maria, Bendazzoli, Claudio and Sandrelli, Annalisa (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):221-249.

Shlesinger, Mariam (1998): Corpus-based Interpreting Studies as an Offshoot of Corpus-based Translation Studies. *Meta*. 43(4):486-493.

Shlesinger, Mariam (2008): Towards a definition of Interpretese: an intermodal, corpus-based study. *In*: Gyde Hansen, Andrew Chesterman and Heidrum Gerzynisch-Arbogast, eds. (2009): *Efforts and Models in Interpreting and Translation Research*. Amsterdam/Philadelphia: John Benjamins, 237-253.

Sinclair, John (1991): *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Tohyama, Hitomi and matsubara, Shigeki (2006): Collections of simultaneous interpreting patterns by using bilingual spoken monologue corpus. *In*: *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06).* (Language Resources and Evaluation Conference, Genoa, 22-28 May 2006). *European Language Resources Association (ELRA)*, 2564-2569.

Zipf, George K. (1949): *Human Behaviour and the Principle of Least Effort*. Cambridge MA: Addison-Wesley.