# The Chinese/English Political Interpreting Corpus (CEPIC): A New Electronic Resource for Translators and Interpreters

**Jun Pan**
Hong Kong Baptist University / 224 Waterloo Road, Kowloon Tong, Hong Kong SAR, China
janicepan@hkbu.edu.hk

## Abstract

The Chinese/English Political Interpreting Corpus (CEPIC) is a new electronic and open access resource developed for translators and interpreters, especially those working with political text types. Over 6 million word tokens in size, the online corpus consists of transcripts of Chinese (Cantonese & Putonghua) / English political speeches and their translated and interpreted texts. It includes rich metadata and is POS-tagged and annotated with prosodic and paralinguistic features that are of concern to spoken language and interpreting. The online platform of the CEPIC features main functions including Keyword Search, Word Collocation and Expanded Keyword in Context, which are illustrated in the paper. The CEPIC can shed light on online translation and interpreting corpora development in the future.

## 1 Introduction

### 1.1 Rationale

The Chinese/English Political Interpreting Corpus (CEPIC) is a new electronic and open access resource developed for translators and interpreters, especially those working with political text types.

The rationale for developing the CEPIC is multifold. One of the reasons is the understudied challenges involved in political interpreting, as illustrated by Buri (2015):

> Both interpreters and translators are under continuous scrutiny in diplomatic settings. Notetakers or other members of the delegation at meetings, round tables, bilateral talks and negotiations are always ready to provide another solution claiming it is more pertinent. Moreover, interpreters and translators may be easily transformed into scapegoats especially when there are misunderstandings or friction between parties – straightforwardly attributed to misinterpretation.

Pan (2007) also identified cases of interpreters failing to capture the source text or mistakenly using the source language instead of the target language due to stress involved in interpreting for presidential speeches. Yet the cases were sporadically identified ones and could not reveal any pattern. Therefore, a corpus that collects political speeches and their interpreting, especially one that is annotated with specific interpreting and spoken language features/issues (such as code-mixing/code-switching as stated above), will benefit greatly the study of the "problematic patterns", apart from offering rich examples of interpreting and translation done by professional practitioners.

### 1.2 Related Work

Despite the significance of Corpus-based Interpreting Studies (CIS), there are still very few open access interpreting corpora (Shlesinger, 1998; Bendazzoli and Sandrelli, 2009; Setton, 2011; Straniero Sergio and Falbo, 2012; Russo et al., 2018), mainly due to the difficulties of data collection, transcription and annotation (Bendazzoli, 2018; Bernardini et al., 2018) .

Among the few number of existing (and publicly accessible) interpreting corpora, the EPTIC (European Parliament Translation and Interpreting Corpus; https://corpora.dipintra.it/eptic/) is very relevant to the CEPIC as both covered official translations and transcribed interpreted texts of speeches delivered in



Temnikova, I, Orăsan, C., Corpas Pastor, G., and Mitkov, R. (eds.) (2019) *Proceedings of the 2nd Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT 2019)*, Varna, Bulgaria, September 5 - 6, pages 82–88.
https://doi.org/10.26615/issn.2683-0078.2019_010

political settings. In particular, the EPIC (European Parliament Interpretation Corpus; http://catalog.elra.info/en-us/repository/browse/ELRA-S0323/) also included annotation of paralinguistic features, which are of interest to interpreting researchers. The EPTIC and the CEPIC are very similar to a great extent since both included simultaneous interpreting of parliamentary speeches, yet the CEPIC also collected data of consecutive interpreting, which is often employed at bilateral meetings or questions and answers at press conferences in political settings. In addition, the EPTIC only included languages translated and interpreted at the European Parliament, while a comparison with those translated and interpreted in other regions/continents would provide interesting perspectives on political translation/interpreting at large.

In this regard, the WAW corpus (http://alt.qcri.org/resources/wawcorpus/) provides a very interesting perspective by covering conference interpreting between English and Arabic in Qatar. However, the data were collected from international conferences rather than from political settings.

Many other corpora that involve the Chinese and English language interpreting in similar settings, including the CEIPPC (Chinese-English Interpreting for Premier Press Conferences, see Wang (2012); also introduced by Setton (2011) and Bendazzoli (2018)) and the CECIC (Chinese-English Conference Interpreting Corpus, see Hu (2013); also introduced by Setton (2011) and Bendazzoli (2018)), are unfortunately not open to public access. In addition, although Cantonese to Putonghua and English simultaneous interpreting has been performed at the Legislative Council (LegCo) of Hong Kong SAR for over two decades, there has seen no existing publicly available corpus designed specifically for the study of interpreting of such speeches, especially one that included paralinguistic features such as the EPIC, although part of the official transcripts are archived regularly online (on government or LegCo websites).

The CEPIC, therefore, aims to provide an open access corpus covering Chinese and English language political interpreting, also in the hope of offering a possible solution to future collection of interpreting corpora by providing templates of metadata collection and solutions to spoken data transcription and annotation, especially for interpreting with the language combination of Chinese (Cantonese and Putonghua) and English.

| Language subsets | Word tokens | Types |
|---|---|---|
| Chinese | 2,578,911 | 83,312 |
| *Cantonese* | *1,072,368* | *61,837* |
| *Putonghua* | *1,506,541* | *30,320* |
| English | 3,815,083 | 32,748 |
| Total | 6,393,994 | 116,060 |

Table 1: The composition of the CEPIC by language.

## 2 About the CEPIC[1]

### 2.1 General Information

The CEPIC is currently over 6 million word tokens in size. It consists of transcripts of speeches delivered by top political figures (e.g. government leaders) from Hong Kong, Beijing, Washington DC and London, as well as their translated/interpreted texts[2]. The speeches were delivered by native speakers (otherwise coded as code-mixing) and interpreted into the B language of the interpreters (usually government interpreters), a phenomenon common in political setting at which the Chinese and English languages are concerned (Pan and Wong, forthcoming). Both directions of Chinese-English and English-Chinese interpreting were covered. Table 1 shows some basic statistics of the CEPIC.

The main speech types of CEPIC include the reading of government reports such as policy addresses and budget speeches, questions and answers at press conferences, parliamentary debates, as well as remarks delivered at bilateral meetings, most of which were done and collected on a yearly basis, except for remarks at bilateral meetings when it depends on if such meetings were held in a specific year. Some of the speeches were interpreted in a consecutive mode, and some in simultaneous, which were coded in the metadata.

In particular, speeches in the Hong Kong subset were mainly interpreted from Cantonese into Putonghua and English, and those in the Beijing subset from Putonghua to English. The other two

[1] Some of the information in this section is also accessible via the CEPIC website (Pan, 2019).

[2] Speeches collected in the corpus, in particular those provided on the official government websites, are considered translations instead of interpreting, as they are translated before interpreting or revised based on the interpreted version, which, with spoken language features (e.g. spoken words and particles) deleted, read more like written language.

subsets, i.e. Washington DC and London, mainly included English speeches delivered in similar settings (which can be regarded as monolingual reference subsets to the interpreted English speeches) and whenever applicable, their interpreted versions in Chinese (usually only at bilateral meetings or joint press conferences).

### 2.2 POS Tagging

The CEPIC is POS tagged with the assistance of Stanford CoreNLP 3.9.2 (Manning et al., 2014). The English taggers used were based on the Part-of-Speech Tagging Guidelines for the Penn Treebank Project (Santorini, 1990), and the Chinese (both Putonghua and Cantonese) on the Part-Of-Speech Tagging Guide-lines for the Penn Chinese Treebank (3.0) (Xia, 2000).

A semi-automatic process was employed to enhance the accuracy rate of machine tagging, in which all taggers were checked and revised based on subsets of manually checked testing data that consisted of about 30 percent of the entire corpus. The process is documented by Pan et al. (forthcoming).

### 2.3 Speech Transcription & Annotation

Data of CEPIC were collected in two ways:

- Speech transcripts and their translations collected from government websites (Raw);
- A revised or newly transcribed version (when there are no readily available transcripts) of these speeches and their interpreted texts based on audios/videos collected from government websites and TV programme archives (Annotated). In particular, the annotated version of the CEPIC was transcribed and annotated in a way that reflects features of spoken language data.

Texts of the CEPIC were manually revised or transcribed based on audios/videos with the speeches and their interpreting, if any. Whenever possible, existing official transcripts provided on government websites and transcripts generated by voice recognition software were used as basis for transcription to help speed up the process. The transcription of CEPIC follows a standardised process and aims to represent the spoken text as close as it was delivered. In addition, all Cantonese texts were transcribed in a way to capture spoken Cantonese features (including particles that are usually omitted in official transcripts provided on government websites). Text and audio/video links were also included at the end of each text for those who may be interested in the sources of the speeches (Figure 1).

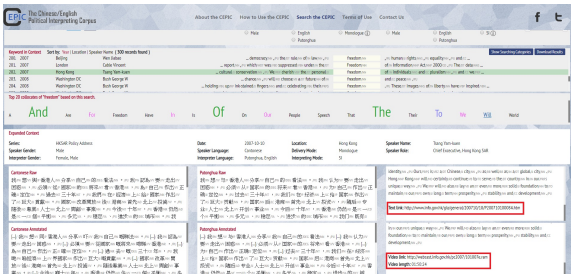

Figure 1: An image of the CEPIC texts with audio/video links and text information

The following examples shows the differences between the raw and annotated data:

- English Raw: So that is the big difference in our approach and the approach that I think might have been debated about. (Press Conference of US Budget Speech, 1997-02-06),
- English Annotated: [er] So [that] that is the big difference [er] in our approach and the approach [er] that [er] I think [er] might have been debated about. (Press Conference of US Budget Speech, 1997-02-06)

As can be seen from the above examples, the annotated version features annotations of different prosodic and paralinguistic features (e.g. fillers, repetitions and self-repair, etc.) that are of concern to the study of spoken language as well as interpreting.

## 3 Main Functions of the CEPIC[3]

The CEPIC features a user-friendly interface with three main functions.[4]

### 3.1 Keyword Search

Users can input a keyword in English or (Simplified/Traditional) Chinese in the corpus. The corpus has a lexical associative function. Therefore, when characters/letters are keyed in the search box, the associative results will automatically display beneath the search box.

[3] A full user manual including graphics of examples can be accessed from the CEPIC website (Pan, 2019).

[4] Examples and data listed in this paper were generated using the CEPIC online search engine (Pan, 2019).

| Parameters | Value |
|---|---|
| {Keyword} | Interesting |
| {Speaker Role} | Member of Parliament (UK) |
| {Time} | 1997 to 2017 |
| {Subset} | Annotated |

Table 2: Parameters used for a sample Keyword Search.

A prosodic/paralinguistic feature can also be searched when choosing the annotated version of the corpus.

Apart from choosing either the raw or annotated subset of the corpus for searching, users can adjust parameters including Part of Speech, Location, Speaker Name, Speaker Role, Speaker Gender, Speaker Language, Delivery Mode, Interpreter Gender, Interpreter Language, Interpreting Mode, and Time Span, to refine a search.

The search results can be arranged by Year, Location, or Speaker Name, and downloaded in excel format.

For instance, if the parameters listed in Table 2 are selected, a total of 8 instances can be found in the CEPIC (Figure 2).

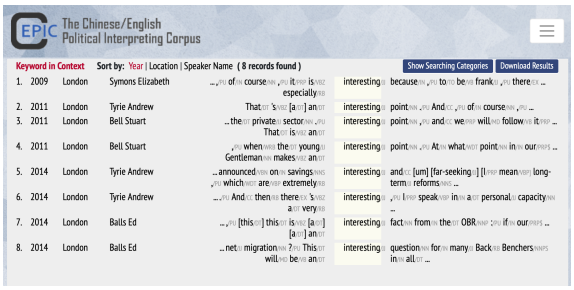


Figure 2: Results of a Keyword Search of ”interesting”(1)

Among the 8 instances, Tyrie Andrew appeared 3 times, showing a possible speaker feature in this case. In addition, all of the instances fell in the time period of 2009-2014, showing a possible trend of using the word among Members of the Parliament in the UK during this specific period of time. Such information may help interpreters and translators acquire knowledge relating to words used by certain speakers or in specific time periods.

### 3.2 Word Collocation

Users can automatically obtain a list of the top 20 collocates of the queried word token in the form of a word cloud. The collocation range is set as 7 words before and after the search term.

| Parameters | Value |
|---|---|
| {Keyword} | Interesting |
| {Location} | Hong Kong |
| {Time} | 1997 to 2017 |
| {Subset} | Raw |

Table 3: Parameters used for a sample Expanded Keyword in Context.

If users click on one of the collocates, the concordance lines that included both the search term and the collocate will appear under Keyword in Context.

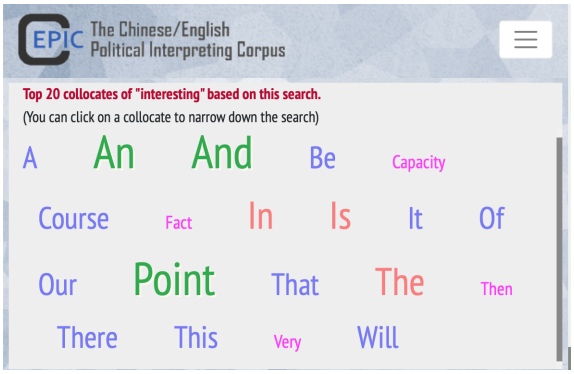


Figure 3: Word Collocation of ”interesting”

Using the same search of ”interesting” in the previous section, we can find ”an”, ”and” and ”point” as the three most frequent collocates of ”interesting” (Figure 3). Such information can benefit greatly anticipation (e.g. of linguistic structures or contextual meaning) in interpreting or translation, in particular in the case of simultaneous interpreting or when a speedy translation service is required (Gile, 1991).

### 3.3 Expanded Keyword in Context

Users can further click a keyword to obtain an Expanded Context, with the respective sub-corpora aligned at the paragraph-level.

The Expanded Context includes the detailed information about the selected Keyword, which also features six windows that display the same speech segment in different languages and versions at paragraph level. For every paragraph, there is a link that redirects to the original text (for the Raw version) or its audio/video (for the Annotated version; including information of the audio/video length).

Again, using ”interesting” as a keyword with the parameters set in Table 3, 2 instances can be found (Figure 4).

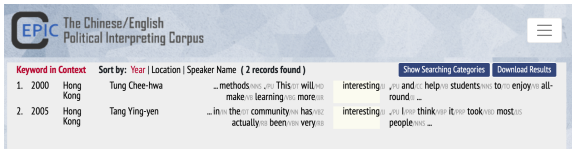


Figure 4: Results of a Keyword Search of "interesting" (2)

The corresponding words of the first "interesting" in the source text in Cantonese and the interpreted/translated versions in Putonghua are the same nouns, i.e. "hing3ceoi3" (in Cantonese Raw) and "xing4qu4" (in both Putonghua Raw and Annotated; both meaning "interest") (Figure 5). The correspondences of the second "interesting" are, however, "jau5ceoi3" in Cantonese Annotated and "you3yi4si1" in Putonghua Annotated (both meaning "interesting", though the former refers to something funnier), but "qiang2lie4" in Putonghua Raw (meaning "intensive") (Figure 6). These renditions indicate certain strategies employed by the speaker or interpreter/translator, i.e. normalisation (in the cases of "hing3ceoi3" and "xing4qu4") and explicitation (in the case of "qiang2lie4").

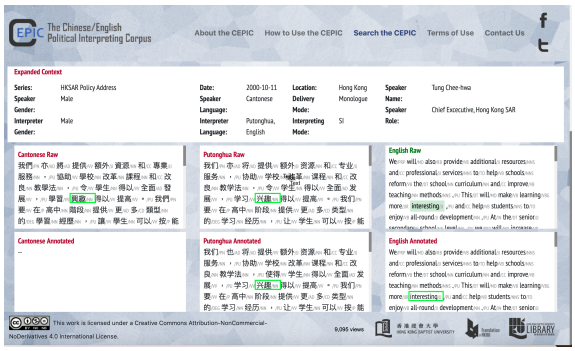


Figure 5: Expanded Keyword in Context of "interesting" (1)

With the help of the detailed information of the Expanded Context, translators/interpreters can then find out how a term is translated/interpreted among Cantonese, Putonghua and English. They can study in detail how the words and their contexts were rendered in spoken and written contexts, or even find out how self-corrections were rendered in a different language, especially in the case of simultaneous interpreting.

Since users can search the CEPIC easily online, interpreters and translators can get timely support not only at the preparation stage, but also during


Figure 6: Expanded Keyword in Context of "interesting" (2)

the process of translation and interpreting. In addition, the CEPIC can benefit language learners, who can make use of the video links to study the pronunciation of certain terms.

## 4 The Way Forward

The CEPIC, as discussed in the previous sections, offers a new online and open access resource for translators and interpreters, with its collection of rich annotated corpora data. It can, as illustrated in the previous section, be used for the preparation of translation and interpreting tasks, and provide online support to interpreters and translators during interpreting/translation. Apart from acquiring knowledge about the use of certain words in political language and interpretation, users can benefit much from exploring the CEPIC in different ways, including finding possible solutions for certain words that are difficult to translate and/or do not have a one-to-one equivalence in the target language.

The CEPIC will provide a good basis for further research on many different topics in interpreting research. The corpus itself will be further expanded and the online platform continuously enhanced to meet various research and education purposes.

In addition, the CEPIC can shed light on future collection and annotation of translation and interpreting corpora, especially the latter, with its systematic annotation scheme, rich metadata information, and unique display and alignment of different language versions.

With its large amount of transcribed interpreting and spoken data of political texts, the CEPIC will also lead to the development of possible tools for computer-assisted interpreting, semi-automatic transcription and alignment, and semi-

automatic POS enhancement (especially for Cantonese). In particular, its data can be used to train machine translation systems (for political texts) or automatic speech recognition and speech-to-text transcription systems (of English, Cantonese and Putonghua).

## Acknowledgements

The CEPIC is developed with the funding and support of the Early Career Scheme (ECS) of Hong Kong SAR's Research Grants Council (Project No.: 22608716), and the Digital Scholarship Grant and the Faculty Research Grant of the Hong Kong Baptist University (Project No.: FRG2/17-18/046).

I would like to thank my colleagues Dr. Billy Tak Ming WONG and Ms. Rebekah WONG for their support and advice, and all the research assistants, student helpers and library colleagues who contributed to the project. Please refer to `https://digital.lib.hkbu.edu.hk/cepic/about.php#project` for a list of the team members.

## References

Claudio Bendazzoli. 2018. Corpus-based interpreting studies: Past, present and future developments of a (wired) cottage industry. In *Making Way in Corpus-based Interpreting Studies (New Frontiers in Translation Studies Series)*. Singapore: Springer, pages 1–19.

Claudio Bendazzoli and Annalisa Sandrelli. 2009. Corpus-based interpreting studies: Early work and future prospects. *Revista Tradumtica: L'aplicaci dels corpus lingstics a la traducci* (7). https://revistes.uab.cat/tradumatica.

Silvia Bernardini, Adriano Ferraresi, Mariachiara Russo, Camille Collard, and Bart Defrancq. 2018. Building interpreting and intermodal corpora: A how-to for a formidable task. In *Making Way in Corpus-based Interpreting Studies (New Frontiers in Translation Studies Series)*. Singapore: Springer, pages 21–42.

Maria Rosaria Buri. 2015. Interpreting in diplomatic settings. https://aiic.net/page/7349/interpreting-in-diplomatic-settings/lang/.

Daniel Gile. 1991. A communication-oriented analysis of quality in nonliterary translation and interpretation. In *Translation: Theory and Practice, Tension and Interdependence*. Amsterdam: John Benjamins Publishing Company, pages 188–200.

Kaibao Hu and Qing Tao. 2013. The chinese-english conference interpreting corpus: Uses and limitations. *Meta* 58(3):626–642.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations.*. Baltimore, Maryland: Association for Computational Linguistics, pages 55–60. https://aclweb.org/anthology/papers/P/P14/P14-5010/.

Jun Pan. 2007. Two styles of interpretation: Reflection on the influence of oriental and western thought patterns on the relationship between the speaker and the interpreter. *Foreign Language and Culture Studies,* 6:677–688.

Jun Pan. 2019. *The Chinese/English Political Interpreting Corpus (CEPIC)*. Hong Kong Baptist University Library [Retrieved on 19 June 2019]. https://digital.lib.hkbu.edu.hk/cepic.

Jun Pan, Fernando Gabarron Barrios, and Haoshen He. forthcoming. Part-of-speech (pos) tagging enhancement for the chinese/english political interpreting corpus (cepic). In *Translation Studies in East Asia: Tradition, Translation and Transcendence*. http://www.cbs.polyu.edu.hk/2019east/index.php.

Jun Pan and Billy T.M. Wong. forthcoming. Pragmatic competence in chineseenglish retour interpreting of political speeches: A corpus-driven exploratory study of pragmatic markers. *Intralinea* https://www.intralinea.org.

Mariachiara Russo, Claudio Bendazzoli, and Bart Defrancq (Eds.). 2018. *Making Way in Corpus-based Interpreting Studies*. Singapore: Springer.

Beatrice Santorini. 1990. *Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd revision, 2nd printing)*. Department of Linguistics, University of Pennsylvania. https://catalog.ldc.upenn.edu/docs/LDC99T42/tagguid1.pdf.

Robin Setton. 2011. Corpus-based interpreting studies (cis): Overview and prospects. In *Corpus-based Translation Studies. Research and Applications*. London: Continuum, pages 33–75.

Miriam Shlesinger. 1998. Corpus-based interpreting studies as an offshoot of corpus-based translation studies. *Meta: Journal des traducteurs* 43(4):486–493. https://doi.org/10.7202/004136ar.

Francesco Straniero Sergio and Caterina Falbo. 2012. *Breaking Ground in Corpus-based Interpreting Studies*. Bern: Peter Lang.

Binhua Wang. 2012. A descriptive study of norms in interpreting : based on the chinese-english consecutive interpreting corpus of chinese premier press conferences. *Meta* 57(1):198–212.

Fei Xia. 2000. The part-of-speech tagging guidelines for the penn chinese treebank (3.0). *RCS Technical Reports Series* (38). http://repository.upenn.edu/ircs$_reports/38$.

## A Supplemental Material

Examples and data listed in this paper are generated using the CEPIC online search engine, which should be cited as:

- Pan, Jun. (2019). The Chinese/English Political Interpreting Corpus (CEPIC). Hong Kong Baptist University Library, [Retrieved on 19 July 2019], Accessed from https://digital.lib.hkbu.edu.hk/cepic/

The following are links related to the CEPIC:

- Link to the CEPIC search engine: https://digital.lib.hkbu.edu.hk/cepic/search.php
- A Google Site page of the CEPIC: https://sites.google.com/a/hkbu.edu.hk/cepic-the-chinese-english-political-interpreting-corpus/