The Chinese/English Political Interpreting Corpus (CEPIC)
An Online Corpus for the Study of Interpreted Political Speeches

Jun PAN
Associate Professor
Department of Translation, Interpreting and Intercultural Studies
Hong Kong Baptist University
janicepan@hkbu.edu.hk
CEPIC: An Online Corpus for the Study of Interpreted Political Speeches

- Rationale
- About the CEPIC
- How to use the CEPIC
- Potential uses
Rationale

- Personal obsession
- Challenges involved in political interpreting
  - Both interpreters and translators are under continuous scrutiny in diplomatic settings. …
  - Moreover, interpreters and translators may be easily transformed into scapegoats especially when there are misunderstandings or friction between parties – straightforwardly attributed to misinterpretation.
  - Buri (2015, para 14)
Rationale

- President Bush: Thank you, very good question. (applause) First of all, I want to compliment you on your English. Very good. The first thing that is important on the Taiwan issue is that my government hopes there is a peaceful, as I said, dialogue, that there is a settlement to this issue. But it must be done in a peaceful way. That’s why I keep emphasizing peaceful. And, by the way, “peaceful” is a word intended for both parties, that neither party should provoke that … go ahead, I’m sorry.

- Interpreter: First of all – sorry.

- President Bush: She’s correcting my English. (laugh)

- Interpreter: I’m sorry, Mr. President.

2002.2.22 Bush visits Qinghua University (Pan 2007)
Rationale

- President Bush: It’s so important for people to realize in both our countries that we’re dealing with human beings that have got desires and loves and frustrations. Even old citizens like me and the Vice President (laugh)
- Interpreter: I’m sorry, sir?
- President Bush: Even old citizens like me and the Vice President (laugh) can benefit by spending time getting to know each other. Obviously, there are some issues in our relationship that we don’t see 100 percent, don’t have a 100 percent agreement on. But it is so much better to discuss these issues after you get to know a person, as a person.

- 2002.2.22 Bush visits Qinghua University  (Pan 2007)
Rationale

- Small number of open access interpreting corpora
Rationale

An open access corpus covering Chinese and English language political interpreting

A possible solution to future collection of interpreting corpora

Solutions to spoken data transcription and annotation

Templates of metadata collection
The Chinese/English Political Interpreting Corpus

https://digital.lib.hkbu.edu.hk/cepic/
About the Project

The Chinese/English Political Interpreting Corpus (CEPIC), with about 6.5 million word tokens in size, is designed for the study of Chinese/English political interpreting and translation.

CEPIC Data

The CEPIC consists of transcripts of speeches delivered by top political figures from Hong Kong, Beijing, Washington DC and London, as well as their translated/interpreted texts.

The main speech types of CEPIC include the reading of government reports such as policy addresses and budget speeches, Q&A at press conferences, parliamentary debates, as well as remarks delivered at bilateral meetings (For details, please refer to the section Basic Statistics).

The corpus features a parallel display of up to six versions of the same speech segment, aligned at paragraph level. Apart from POS tagging, the corpus is also annotated with different prosodic and paralinguistic features that are of concern to the study of spoken language as well as interpreting.

The CEPIC can be used to investigate matters relating to Chinese/English political translation/interpreting and political discourse at large. It can also serve students, teachers, as well as people working in political settings, in aspects of political speech delivery and translation/interpreting production. Users can also download search results from the corpus for their own teaching/research purposes.
About the CEPIC

Introduction

Basic statistics

The CEPIC consists of parallel representation of speech segments in Cantonese, Putonghua and English. The following table shows the number of words (word token) and unique words (type) in each language.

Table 1. The composition of the CEPIC by language

<table>
<thead>
<tr>
<th>Language</th>
<th>Word (Word Token)</th>
<th>Unique Word (Type)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese</td>
<td>2,578,911</td>
<td>83,312</td>
</tr>
<tr>
<td>Cantonese</td>
<td>1,072,368</td>
<td>61,837</td>
</tr>
<tr>
<td>Putonghua</td>
<td>1,506,541</td>
<td>30,320</td>
</tr>
<tr>
<td>English</td>
<td>3,815,083</td>
<td>32,748</td>
</tr>
<tr>
<td>Total</td>
<td>6,393,994</td>
<td>116,060</td>
</tr>
</tbody>
</table>

The main speech types of CEPIC include the reading of government reports such as policy addresses and budget speeches, Q&A at press conferences, parliamentary debates, as well as remarks delivered at bilateral meetings. The following table shows the current composition of the corpus and some basic statistics of each subset.
About the CEPIC

Table 2. The composition of the CEPIC by speech types

<table>
<thead>
<tr>
<th>Speech Type</th>
<th>Word (Word Token)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 HK SAR Policy Addresses (HKPA)</td>
<td>1,290,774</td>
</tr>
<tr>
<td>2 Press Conferences of HK SAR Policy Addresses (HKPAPC)</td>
<td>326,194</td>
</tr>
<tr>
<td>3 HK SAR Budget Speeches (HKBS)</td>
<td>1,167,530</td>
</tr>
<tr>
<td>4 Press Conferences of HK SAR Budget Speeches (HKBSPC)</td>
<td>419,236</td>
</tr>
<tr>
<td>5 PRC Reports on the Work of the Government (PRCWoG)</td>
<td>782,794</td>
</tr>
<tr>
<td>6 Press Conferences of PRC Reports on the Work of the Government (PRCWoGPC)</td>
<td>448,111</td>
</tr>
<tr>
<td>7 US State of the Union Addresses (USSoUA)</td>
<td>275,018</td>
</tr>
<tr>
<td>8 Press Conferences of US State of the Union Addresses (USSoUAPC)</td>
<td>266,639</td>
</tr>
<tr>
<td>9 US Budget Speeches (USBS)</td>
<td>73,115</td>
</tr>
<tr>
<td>10 Press Conferences of US Budget Speeches (USBSPC)</td>
<td>328,850</td>
</tr>
<tr>
<td>11 UK State Opening Addresses of Parliament (UKSOoP)</td>
<td>31,006</td>
</tr>
<tr>
<td>12 Debates on the UK State Opening Addresses of Parliament (UKSOoPD)</td>
<td>53,941</td>
</tr>
<tr>
<td>13 UK Budget Speeches (UKBS)</td>
<td>469,452</td>
</tr>
<tr>
<td>14 Debates on the UK Budget Speeches (UKBSD)</td>
<td>376,721</td>
</tr>
<tr>
<td>15 Bilateral Meetings between PRC Key Politicians and their Counterparts in US (BMPRCUS)</td>
<td>70,138</td>
</tr>
<tr>
<td>16 Bilateral Meetings between PRC Key Politicians and their Counterparts in UK (BMPRCUK)</td>
<td>14,473</td>
</tr>
<tr>
<td>Total</td>
<td>6,393,994</td>
</tr>
</tbody>
</table>
The following is a list of words that have specific meaning in the CEPIC.

1. **Annotated**: It refers to a revised or newly transcribed version (when there are no readily available transcripts) of political speeches and their interpreted texts based on audios/videos collected from government websites and TV programme archives (also see **Raw**). The texts were annotated with prosodic and paralinguistic features that are of concern to the study of spoken language as well as interpreting.

2. **CI**: It refers to **Consecutive Interpreting**, a mode of interpreting employed in political settings when the interpreter speaks after the speaker finishes (part of) his/her speech (also see **Interpreting Mode** and **SI**).

3. **Delivery Mode**: It refers to the ways how a speech is delivered and includes mainly **Monologue** and **Dialogue**.

4. **Dialogue**: It refers to the **Delivery Mode** in which one or more political figures debate with or answer questions from the audience or each other.

5. **Interpreter Language**: It refers to the language that the interpreter interprets into.

6. **Interpreting Mode**: There are usually two modes of interpreting employed in political settings, i.e., **CI** and **SI**.

7. **Location**: It refers to the (capital) city where the political speech/interpreting is delivered.

8. **Monologue**: It refers to the **Delivery Mode** in which a political figure delivers the speech to the audience, with almost no interactions with the audience or other political figures.

9. **Part of Speech (POS)**: It refers to the grammatical category a word belongs to, which is also called a word class. The CEPIC is POS tagged, i.e., each word is tagged with its part of speech. (Also refer to the section **POS tagging** for a full list of POS tags used in the CEPIC).

10. **Raw**: It refers to the speech transcripts and their translations collected from government websites, with no annotations of prosodic or paralinguistic features (also see **Annotated**).

11. **SI**: It refers to **Simultaneous Interpreting**, a mode of interpreting employed in political settings when the interpreter speaks (usually in an interpreting booth) while the speaker delivers his/her speech (also see **CI** and **Interpreting Mode**).

12. **Speaker Language**: It refers to the language that the speaker uses.

13. **Speaker Role**: It refers to the position of the speaker in a government.
<table>
<thead>
<tr>
<th>Language</th>
<th>Text</th>
<th>Annotated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cantonese</td>
<td>我們要重新認識八十年代關於香港發展的一些重要概念，放棄二元對立分析方法。(<em>HKSAR Policy Address, 2008-10-15</em>)</td>
<td>[...]我哋要重新認識八十年代關於香港發展嘅一啲嘅重要嘅概念，係放棄二元對立嘅分析嘅方法。(<em>HKSAR Policy Address, 2008-10-15</em>)</td>
</tr>
<tr>
<td>English</td>
<td>So that is the big difference in our approach and the approach that I think might have been debated about. (<em>Press Conference of US Budget Speech, 1997-02-06</em>)</td>
<td>[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. (<em>Press Conference of US Budget Speech, 1997-02-06</em>)</td>
</tr>
</tbody>
</table>
About the CEPIC

The CEPIC (Commodity Exchange and Physical Infrastructure Corporation) is an organization focused on the trading, distribution, and physical infrastructure of commodities. Its primary goal is to facilitate the smooth flow of goods and services, ensuring fair and efficient transactions. The CEPIC operates in various locations around the world, including Hong Kong, Washington DC, and London. Its activities span multiple sectors, providing services that are crucial for the global economy. The organization emphasizes the importance of freedom, human rights, and equality, and it works towards creating a more prosperous and sustainable future for all.
About the CEPIC

The CEPIC is POS tagged with the assistance of Stanford CoreNLP 3.9.2 (Manning et al. 2014).

A semi-automatic process was employed to enhance the accuracy rate of machine tagging, in which all tags were checked and revised based on subsets of manually checked testing data. Please click here (available soon) for a detailed account of the semi-automatic process employed in the POS tagging of CEPIC.

The following table provides a list of the POS tags that appeared in the English subset of CEPIC, which is based on the Part-of-Speech Tagging Guidelines for the Penn Treebank Project (Santorini 1990:6-7).

Table 1. POS tags that appeared in the English subset of CEPIC (based on Santorini 1990:6-7)

<table>
<thead>
<tr>
<th>POS tagger</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>/AD</td>
<td>Adverb</td>
</tr>
<tr>
<td>/AS</td>
<td>Aspect Particle</td>
</tr>
<tr>
<td>/BE</td>
<td>被 in be-construction</td>
</tr>
<tr>
<td>/CC</td>
<td>Coordinating conjunction</td>
</tr>
<tr>
<td>/CD</td>
<td>Cardinal number</td>
</tr>
<tr>
<td>/CS</td>
<td>Subordinating conjunction</td>
</tr>
<tr>
<td>/DEC</td>
<td>的 as a complementizer or a nominalizer</td>
</tr>
<tr>
<td>/DEG</td>
<td>的 as a genitive marker and an associative marker</td>
</tr>
<tr>
<td>/DEV</td>
<td>Manner地 (before VP)</td>
</tr>
<tr>
<td>/DT</td>
<td>Determiner</td>
</tr>
<tr>
<td>/ETC</td>
<td>For words 等,等等</td>
</tr>
<tr>
<td>/FW</td>
<td>Foreign word</td>
</tr>
<tr>
<td>/LR</td>
<td>Adjective</td>
</tr>
<tr>
<td>/LIR</td>
<td>Adjective, comparative</td>
</tr>
<tr>
<td>/EX</td>
<td>Existential there</td>
</tr>
<tr>
<td>/FW</td>
<td>Foreign word</td>
</tr>
</tbody>
</table>

Details of the POS tagging:
About the CEPIC

<table>
<thead>
<tr>
<th>Part of Speech</th>
<th>Speaker Language</th>
<th>Delivery Mode</th>
<th>Interpreter Language</th>
<th>Interpreting Mode</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freedom</td>
<td>All</td>
<td>All</td>
<td>All</td>
<td>All</td>
<td>1997</td>
</tr>
<tr>
<td>Raw</td>
<td>All</td>
<td>All</td>
<td>All</td>
<td>All</td>
<td>2017</td>
</tr>
</tbody>
</table>
How to use the CEPIC

Keyword Search

Word Collocation

Expanded Keyword in Context
### Keyword Search

<table>
<thead>
<tr>
<th>Keyword in Context</th>
<th>Sort by: Year</th>
<th>Location</th>
<th>Speaker Name</th>
<th>Context</th>
</tr>
</thead>
</table>
| 1. 2009 London     | Symons Elizabeth |          |                  | ...[whitespace removed]...[
| 2. 2011 London     | Tyrie Andrew   |          |                  | That's...[
| 3. 2011 London     | Bell Stuart    |          |                  | ...the private sector...[
| 4. 2011 London     | Bell Stuart    |          |                  | ...when the young...[
| 5. 2014 London     | Tyrie Andrew   |          |                  | ...announced...[
| 6. 2014 London     | Tyrie Andrew   |          |                  | ...And then...[
| 7. 2014 London     | Balls Ed       |          |                  | ...[this]...[
| 8. 2014 London     | Balls Ed       |          |                  | ...net...[

Note: The context text has been truncated for clarity.
Top 20 collocates of "interesting" based on this search.
(You can click on a collocate to narrow down the search)
<table>
<thead>
<tr>
<th>Keyword in Context</th>
<th>Sort by: Year</th>
<th>Location</th>
<th>Speaker Name</th>
<th>(2 records found)</th>
</tr>
</thead>
</table>
| 1. 2000 HK         | Tung Chee-hwa |          |              | "This will make learning more interesting."
| 2. 2005 HK         | Tang Ying-yen |          |              | "The community has actually been very interesting."

The Chinese/English Political Interpreting Corpus (EPIC)
## Expanded Keyword in Context

### Cantonese Raw

<table>
<thead>
<tr>
<th>Series</th>
<th>Date</th>
<th>Location</th>
<th>Speaker</th>
<th>Gender</th>
<th>Interpreter</th>
<th>Language</th>
<th>Mode</th>
<th>Speaker</th>
<th>Name</th>
<th>Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>HKSAR Policy Address</td>
<td>2000-10-11</td>
<td>Hong Kong</td>
<td>Male</td>
<td>Cantonese</td>
<td>Putonghua</td>
<td>SI</td>
<td>SI</td>
<td>Male</td>
<td>Tung Chee-Iwa</td>
<td>Chief Executive, Hong Kong SAR</td>
</tr>
</tbody>
</table>

### Putonghua Raw

<table>
<thead>
<tr>
<th>Series</th>
<th>Date</th>
<th>Location</th>
<th>Speaker</th>
<th>Gender</th>
<th>Interpreter</th>
<th>Language</th>
<th>Mode</th>
<th>Speaker</th>
<th>Name</th>
<th>Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>HKSAR Policy Address</td>
<td>2000-10-11</td>
<td>Hong Kong</td>
<td>Male</td>
<td>Cantonese</td>
<td>Putonghua</td>
<td>SI</td>
<td>SI</td>
<td>Male</td>
<td>Tung Chee-Iwa</td>
<td>Chief Executive, Hong Kong SAR</td>
</tr>
</tbody>
</table>

### Cantonese Annotated

We provide additional resources and professional services to help schools reform their education curriculum and improve teaching methods. This will make learning more interesting, allowing students to enjoy a full development. At the senior secondary level, the full development will be more pronounced.

### Putonghua Annotated

We also provide additional resources and professional services to help schools reform their education curriculum and improve teaching methods. This will make learning more interesting, allowing students to enjoy a full development. At the senior secondary level, the full development will be more pronounced.

### English Raw

We will also provide additional resources and professional services to help schools reform their education curriculum and improve teaching methods. This will make learning more interesting, allowing students to enjoy a full development. At the senior secondary level, the full development will be more pronounced.

### English Annotated

We will also provide additional resources and professional services to help schools reform their education curriculum and improve teaching methods. This will make learning more interesting, allowing students to enjoy a full development. At the senior secondary level, the full development will be more pronounced.
### Expanded Context

<table>
<thead>
<tr>
<th>Series:</th>
<th>Press Conference of HK SAR Budget Speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaker</td>
<td>Male</td>
</tr>
<tr>
<td>Gender:</td>
<td></td>
</tr>
<tr>
<td>Interpreter</td>
<td>Unspecified</td>
</tr>
<tr>
<td>Date:</td>
<td>2005-03-16</td>
</tr>
<tr>
<td>Location:</td>
<td>Hong Kong</td>
</tr>
<tr>
<td>Delivery:</td>
<td>Dialogue</td>
</tr>
<tr>
<td>Language:</td>
<td>English</td>
</tr>
<tr>
<td>Mode:</td>
<td>SL</td>
</tr>
<tr>
<td>Speaker Role:</td>
<td>Financial Secretary, Hong Kong SAR</td>
</tr>
</tbody>
</table>

#### Cantonese Raw
- NIL

#### Cantonese Annotated
- 有關於貧窮
- 當方方面

#### Putonghua Raw
- NIL

#### Putonghua Annotated
- [interpreter] 扶貧
- 方面
- 自從
- 董卿
- 先生
- 说
- 成立
- 貧富
- 委员会
- 以后
- 社会
- 税收
- 反应
- 而
- 很
- 有意思
- 的
- 大部分
- 人
- [interesting]
- 觉得
- 发现
- 中产
- 阶级
- 反应
- 很
- 强烈
- 他们
- 知道
- 他们
- 要
- 交税
- 
- 税款
- 是
- 他们
- 辛苦
- 赚来
- 的
- 钱
- 人

#### English Raw
- As far as poor is concerned. We have seen many interesting things happen. I think it took us by surprise that there was such a strong reaction from the middle-class.

#### English Annotated
- [um] As far as poor is concerned. We have seen many interesting things happen. I think it took us by surprise that there was such a strong reaction from the middle-class.
Extra information

- **To search:**

- **Manual:**

- **High frequency words:**
  [https://digital.lib.hkbu.edu.hk/cepic/about.php#WordFrequencyData](https://digital.lib.hkbu.edu.hk/cepic/about.php#WordFrequencyData)
Potential uses
For (student) interpreters/translators

- Large database of high-quality Chinese and English texts delivered by native politicians and interpreters/translators
- Using the CEPIC to prepare translation and interpreting tasks & obtain online support
- Etc.
Potential uses
For interpreting/translation (student) researchers

- Using the CEPIC to study interpreted language
- Using the CEPIC to compare translation with interpreting
- The CEPIC and political discourse
- Etc.
Potential uses

Others

- Language learning
- The study of politician language
- The study of spoken language (e.g., fillers, etc.)
- Semi-automatic language processing (transcription, POS tagging, annotation, etc.)
Links to relevant corpora

- The European Parliament Interpreting Corpus
- The European Parliament Translation and Interpreting Corpus
- The WAW corpus
- The European Comparable and Parallel Corpora
- The Corpus of Political Speeches
- The Digital Corpus of the European Parliament
- The UN Parallel Text
- 中国外交话语语料库
Acknowledgements

- The CEPIC is developed with the funding and support from:
  - The Early Career Scheme (ECS) of the Research Grants Council (Project Title: Interpreting into the B language: A corpus-oriented study of pragmatic markers in interpreted political speeches from Chinese to English, Project No.: 22608716);
  - The Digital Scholarship Grant of the Hong Kong Baptist University (Project Title: The Chinese-English Political Interpreting Corpus (CEPIC): An Online Corpus for the Study of Interpreted Political Speeches); and
  - The Faculty Research Grant of the Hong Kong Baptist University (Project Title: The Use of Pragmatic Markers in Chinese-English Political Interpreting: A Corpus-oriented Study, Project No.: FRG2/17-18/046)
Project Team

- Principal Investigator: Dr. Jun PAN (Associate Professor, Translation Programme, Hong Kong Baptist University)
- Co-Investigator & Special Consultant to the CEPIC: Dr. Billy Tak Ming WONG (Research Coordinator, University Research Centre, Open University of Hong Kong)
- Senior Adviser to the CEPIC: Ms. Rebekah WONG (Head of Digital and Multimedia Services Section, Library, Hong Kong Baptist University)
- Research Assistants:
  - Mr. Fernando GABARRON BARRIOS
  - Mr. Steven Haoshen HE
  - Miss Chris Chencheng KUANG
  - Miss Hannah Qiuhan LIN
  - Mr. William Dongpeng PAN
  - Miss Jennifer Lok Man WONG
  - Miss Grace Jing ZHANG
- Student Helpers:
  - Mr. Antonio Yijiao GUO
  - Mr. Hank Lin HAN
  - Miss Rigel Chung Ting PAK
  - Miss Gladys Hiu Man SHIU
  - Miss Jess Lin Wing SZE
- Miss Tammy Cho Ying TANG
- Miss Janny Chi Wai WONG
- Miss Alice Yuxin YANG
- Mr. Niko Donghuan ZHANG
- Library Colleagues:
  - Mr. Wing Chung YIP
  - Mr. Timothy Sit YEUNG
  - Miss Sharon Suk Man YU
  - Miss Katie Kee Yee CHENG
<table>
<thead>
<tr>
<th>謝謝你們</th>
<th>謝謝你們</th>
<th>Thank you</th>
</tr>
</thead>
<tbody>
<tr>
<td>多謝你哋</td>
<td>謝謝大家</td>
<td>Thank you</td>
</tr>
</tbody>
</table>

**Join the CEPIC community**
- https://twitter.com/CEPICcorpus
- https://www.facebook.com/CEPIC
- Cthechineseenglishpoliticalinterpretingcorpus/

**Feedback & enquires**
- janicepan@hkbu.edu.hk