Corpus of Hong Kong Cantonese 香港粵語語料庫

- 180,000-word corpus of Cantonese Speech
- 52 spontaneous conversations
- 42 radio programmes
- Transcribed (UTF-8); Transliterated Segmented; POS tagged
- English translation described in paper, not in downloadable corpus
- Available directly for download (CC BY)
  https://github.com/fcbond/hkcancor
- Produced by Luke Kang Kwong and ML Wong
  put online and licence clarified by Francis Bond

Creation

- 30 hours of recordings (March 1997 — August 1998)
- Native speakers of Cantonese
- ordinary settings with family members, friends and colleagues talking with each other freely on everyday topics such as current affairs, work and study, and personal hobbies
- Some parts selected

Meta-Data/Annotation

- Meta-Data
  - Tape number (of recording); Date of recording
  - Number of Speakers; List of Speakers (Code-Sex-Age-Origin)
    (e.g. A-M-22-HK says A is a 22-year-old male speaker from Hong Kong)
- Annotation
  - Each Utterance has the speaker code
  - Utterances are segmented, POS tagged and transliterated
    基本上/d/ge3i1bun2soeng6/
    哩個/r/ni1go3/ze
    ...
- The whole corpus is wrapped in xml (but not very well)

Usage

- Used to examine the uses of the frequently used sentence final particles wǒ and bǒ in the 1990s in Hong Kong Cantonese by examining speech data.
- Question: are wǒ (喎) and bǒ (嘸) variant forms?
- Answer: No

  > "[...] the two SFPs carry and serve different meanings and functions in modern Hong Kong Cantonese, and thus they are not exactly the same particles and not interchangeable as previously assumed." (Leung, 2010, p21)

- Also used as a corpus in the *PyCantonese* Project: Working with Cantonese corpus data using Python, by Jackson L. Lee (https://github.com/pycantonese/pycantonese)

References

- Luke, K. K. and Wong, M. L. (2015). The Hong Kong Cantonese corpus: Design and uses. *Journal of Chinese Linguistics*. (to appear)

- Wong, P.-W. (2006). The specification of POS tagging of the Hong Kong University Cantonese corpus. *International Journal of Technology and Human Interaction*, 2(1):21–38. DOI10.4018/jthi.2006010102

- Leung, W.-M. (2010). On the identity and uses of Cantonese sentence-final particles in the late 20th century: The case of wo and bo. *Asian Social Science*, 6(1):13–23. (http://www.ccsenet.org/journal/index.php/ass/article/view/4765)

ISLRN

Title	HKcancor
Full Title	Hong Kong Cantonese Corpus
Resource Type	Speech
Source/URL	compling.hss.ntu.edu.sg/hkcancor
Format/MIME Type	text/xml
Size/Duration	230,000 words
Access Medium	online
Description	The Hong Kong Cantonese Corpus was collected from transcribed conversations that were recorded between March 1997 and August 1998. About 230,000 Chinese words were collected in the annotated corpus. It contains recordings of spontaneous speech (51 texts) and radio programmes (42 texts), which involve 2 to 4 speakers, with 1 text of monologue. The text were word-segmented, annotated with part-of-speech tagging and Cantonese pronunciation using the romanisation scheme of the Linguistic Society of Hong Kong (LSHK).
Version	?
Media Type	Transcribed Speech
Language	yue (Cantonese)
Resource Creator	KK Luke
Distributor	KK Luke
Rights Holder	KK Luke