# MYCanCor: A Video Corpus of spoken Malaysian Cantonese

**Andreas Liesenfeld**
Nanyang Technological University
Singapore
LIES0002@ntu.edu.sg

**Abstract**

The Malaysia Cantonese Corpus (MYCanCor) is a collection of recordings of Malaysian Cantonese speech mainly collected in Perak, Malaysia. The corpus consists of around 20 hours of video recordings of spontaneous talk-in-interaction (56 settings) typically involving 2-4 speakers. A short scene description as well as basic speaker information is provided for each recording. The corpus is transcribed in CHAT (minCHAT) format and presented in traditional Chinese characters (UTF8) using the Hong Kong Supplementary Character Set (HKSCS). MYCanCor is expected to be a useful resource for researchers interested in any aspect of spoken language processing or Chinese multimodal corpora.

**Keywords:** Malaysian Cantonese, spoken corpora, naturally-occurring talk-in-interaction

## 1. Corpus description

The Malaysia Cantonese Corpus (MYCanCor) was conceptualized as a repository of naturally occurring conversations collected in the Malaysian Cantonese speech community. It consists of video recordings of a variety of everyday settings where spontaneous everyday conversation commonly takes place, such as:

- Family dinner conversations (private homes)
- Conversations over lunch or dinner with friends, relatives or colleagues (Restaurants and street food vendors)
- Car and public transport rides with family members, friends or colleagues (private car or public bus)
- Conversations at work between colleagues as well as between employees and manager (various office spaces)
- Conversations between customers and vendors (shop or factory setting)
- Conversations between local citizens and government representatives (local government office)
- Conversations between students as well as between students and staff (various educational institutions)

The conversations typically involve 2-4 speakers (up to a maximum of 8 speakers) of Malaysian Cantonese and are between 2 and 40 minutes long. Topics include a wide range of commonplace activities related to the setting in which the conversation takes place, frequently including elements of:

- storytelling and narrative sequences
- management of social obligations including greeting and leave segments
- social chat and expressions of emotional states
- task-oriented dialog such as directions while driving or task instructions in workspace environment
- management of authority in dialog, e.g. task assignment and task-oriented dialog in a workspace environment

The recordings were collected mainly in the city of Ipoh and other places in the state of Perak in 2017 and are now available to the research community as a collection of 56 video recordings (around 20 hours of video content).
Each conversation is transcribed in CHAT format at the minCHAT level (see 3. Transcription). Each video recording also comes with a scene description written in full sentence format that provides the following information:

- brief description of the setting in which the recording took place
- summary of the topics talked about
- specification of the task, in case the recording includes task-oriented sequences
- brief summary of emotional expressions used and brief description of the emotional states talked about, in case the recording features frequent expressions of emotional states
- information about age and gender of each participant

Each recording is available in both high definition video (.mov) and uncompressed audio format (.wav) with a transcription in CHAT format (.cha) (MacWhinney, 2000) and a scene description (.txt).

### 1.1. Technical specifications

All 56 conversations were filmed from one angle with one HD video camera (1080p) providing access to facial expression and gesture information of most (but not all) participants. The camera was set up in such a way that in almost all cases all participants appear on the screen.
Audio data was captured with one directional supercardioid microphone in uncompressed monophonic wave format (96kHz / 24-bit). The recordings took place in a real world environment and feature natural background noise (for example in a restaurant or workspace settings).

## 2. Data collection and sampling

The conversations featured in the corpus aim to present naturally-occurring everyday speech. All recordings took place after informed consent was obtained from each participant and all conversations were recorded after the researcher had left the scene. Several measures were taken to minimize the influence of the researcher on the naturalness of the recording and to mitigate the effects of the observer’s paradox (Labov, 1972). A basic level of rapport was built up with participants before the recording. Recordings that show an obvious influence of the researcher on the produced utterances have not been included in the corpus. Building on several months of fieldwork in Ipoh and other parts of Perak, a multiple-entry chain sampling method was used in order to recruit participants from a wider variety of socio-economic and educational backgrounds. However, no metrics have been recorded to attest for this. Most interlocutors of a single recorded conversation knew each other prior the recording as friends, family members or colleagues. The corpus is roughly balanced for gender and age group (see Table 1).

| Variable | Distribution |
|---|---|
| Gender (binary) | Male: 48%<br>Female: 52% |
| Age group (years) | under 30: 33%<br>30-60: 45%<br>60+: 22% |

Table 1: Gender and age group distribution in MYCanCor (rounded)

Reflecting common practice in Malaysia, the corpus also features code mixing of Cantonese speech with English, Malay, Mandarin and Southern Min. However, an estimated 95% of the recorded utterances are in Cantonese and the most frequent code mixing language is English (as well as local varieties of English such as Malaysian English and Singlish).

Data collection and recruitment of participants was in line with the British Association for Applied Linguistics (BAAL) Recommendations on Good Practice in Applied Linguistics (https://baal.org.uk/resources/). Participants were recruited under two prerequisites:

- self-proclaimed ability to have a conversation in Cantonese (proficiency in written Chinese not required)
- having spent two-thirds of one’s life in Malaysia (self-proclaimed)

All participants are self-proclaimed speakers of Malaysian Cantonese Chinese. This variety of Cantonese Chinese somewhat differs from the varieties spoken in speech communities in China (such as the speech community in the Hong Kong SAR and those in Guangdong and Guangxi province). For instance, differences are evident in prosody, lexical token preference and code-mixing. However, speakers of this variety are generally able to fluently communicate with members of, for instance, the Hong Kong Cantonese speech community. Reflecting the Malaysian Cantonese Chinese speech community in general, the level of proficiency considerably varies among participants. Participants may be English-, Mandarin- or Malay-educated and may speak Cantonese as a second language, home language or work language alongside with other languages. Code-mixing with other languages such as English, Singlish, Mandarin Chinese, Hokkien Chinese, Hakka Chinese or Malay is a commonplace phenomenon in the speech community. Recruitment of participants was not tied to ethnic groups, reflecting the ethnically diverse nature of the local speech community.

The data collection mainly took place in the state of Perak (especially the city of Ipoh), the region has the highest percentage of Cantonese speakers in Malaysia (Tan, 2005). Cantonese Chinese is seldom used by educational institutions in Perak. Secondary education is commonly available only in Malay, English or Mandarin Chinese. Tertiary education only in Malay and/or English. Based on participant observation in 2017, however, Cantonese is the lingua franca of the ethnic Chinese community in Ipoh as well as in many (but not all) other cities and villages in Perak, such as Taiping, Kuala Kangsar, Teluk Intan and Kampar. Based on these observations, Perak features one of the biggest, if not the biggest, Cantonese speech communities outside of China.

## 3. Transcription

All conversations are transcribed using the minCHAT format, a basic version of the CHAT format (MacWhinney, 2000). The transcripts focus on the word level, aiming to identify words and sentences in the utterances, prioritizing lexis over, for example, phonological aspects. In this sense the transcripts aim to provide a modest starting point for possible further annotation of more aspects of the complexity of naturally-occurring speech. In order to achieve a high level of accuracy and consistency, all transcripts have been proofread by Malaysian Cantonese speakers with relevant experience and backgrounds. The transcripts are presented in Traditional Chinese characters (UTF8) including the Hong Kong Supplementary Character Set (HKSCS). Only identifiable units that can be presented by Chinese characters are transcribed. For example, the utterance (’gw’, ’o’, ’ng’, ’2’), (’d’, ’u’, ’ng’, ’1’), (’w’, ’aa’, ”, ’2’) (onset, nucleus, coda, tone) would be transcribed as 廣東話 . But the utterance (’gw’, ’o’, ’ng’, ’2’),(’d’, ”, ”, ’1’) would be transcribed as 廣xxx since (’d’, ”, ”, ’1’) has no defined corresponding character in this format. Completely unintelligible utterances are also transcribed as xxx (see Example 1).

Cantonese-specific characters that are not (yet) supported by HKSCS are presented as romanized strings following the Jyutping Romanization Scheme. Code switching utterances in Mandarin are presented in Traditional Chinese characters (UTF8), for Hokkien and Southern Min the Tai-wan Romanization system (Tai-lo) was used. Modal Particles (語氣詞) and Modal Particle Morphemes ( 語氣語素) are transcribed following UTF8+HKSCS standards. For all utterances that are supported by UTF8+HKSDS no separate romanization in Jyutping or Pinyin is provided.

Following minCHAT requirements, utterances have to end with an utterance terminator (such as period, exclamation mark or question mark). In order to fulfill this requirement, the transcript uses periods to terminate all utterances regardless whether they are questions or exclamations. No question or exclamation marks (標點符號) are annotated, but question and exclamatory particles are transcribed.

A separate CHAT transcription file was created for each of the 56 conversations, but it should be no problem to combine the transcripts into one file using one of the many tools available for CHAT. Each file (.cha) is verified as machine-readable using the CHECK function of the CLAN editor.

### 3.1. Word segmentation and annotation

The transcripts do not include word segmentation or chunking of Chinese characters. Utterances are transcribed as continuous strings of Chinese characters in UTF8 encoding. The minCHAT transcript does, however, separate segments of utterances based on pauses of considerable length (more than 0.1 seconds) (see Example 1). Also, no Part-of-Speech tags or further syntactic annotation is provided as part of the transcript. No additional (systematic) annotation of gestures, facial expression or events is provided.

```
@Begin
@Languages: zho-yue
@Participants: P1 Wong Older Sister,
P2 Chan Younger Sister
@ID: zho-yue|mycancor|P1|27;1.10||||
Target_P2|||
@ID: zho-yue|mycancor|P2|39;2.||||
Target_P1|||
*P2: 你食咩啊.
%com: every utterance ends with
  an utterance terminator (period).
*P1: 白果薏米.
*P2: 同怡保好似好唔同 哈哈哈.
*P1: 唔同啊(0.1)冇得比啦.
%com: Utterances are segmented by
      pauses exceeding 0.1 seconds.
*P2: 依佢係.
*P1: 白 但因為 因為佢冇煮溶個.
%com: All lexical items
      (onset, nucleus, coda, tone) are
      transcribed as Chinese
      characters.
%act: P2 points at the bowl.
%com: Gestures may be annotated
      as informal descriptions.
*P2: xxx睇下 個腐竹.
%com: Unintelligible or incomplete
      lexical units are
      transcribed as xxx.
*P1: 個腐竹 呃 係咯 同埋唔知點解佢
     唔係白色咯.
%com: Modal Participles and Modal
      Particle Morphemes are
      transcribed following
      UTF8+HKSDS conventions.
@End
```

Example 1: MYCanCor transcription example in minCHAT format. Each file begins with a header. Comments are marked with %com. Gestures, actions and scene descriptions with %act.

## 4. Possible applications

The Malaysia Cantonese Corpus might be of interest to any researcher interested in video corpora or Malaysian Cantonese. To the author's knowledge, MYCanCor is in fact the first language resource available that is concerned with the Cantonese speech community in Malaysia. Although the corpus is of rather small size, the data that went into its making should be of sufficient quality to make it a useful resource for a wide variety of applications.

The dataset was compiled as part of several research projects in the area of interactional linguistics and spoken language processing. The described transcription files were produced as part of these projects.

Considerable effort also went into providing accurate minCHAT transcription for each conversation. One of the advantages of the CHAT format is that the provided minCHAT can be adjusted or expanded with relative ease to encode additional phenomena, should the need arise. The design choices made should allow a combination with other resources with relative ease. The minCHAT transcription, for example, can be expanded to midCHAT or even a full conversation analytical (CA) transcription. Eventual additional layers or syntactic, semantic, phonological or prosodic annotation can also be integrated in CHAT. Facial expressions, gaze or gestures can be added for all conversations given that the video data provides sufficient visual information.

A problem in building Asian language resources is the existence of multiple and often competing writing systems and encoding systems. This is also true for Cantonese. Unicode has only recently included the special Hong Kong characters and replaced BIG-5 as the most popular encoding standard for Cantonese. There are also several completing romanization schemes for Cantonese, with the Jyutping Romanization scheme developed by the Linguistic Society of Hong Kong probably being the most widely used standard now. MYCanCor currently uses Unicode encoding without romanization, but several tools exist to facilitate the creation of a romanization if needed.

A good way to get started with processing Cantonese speech data are the resources compiled in the pycantonese library in Python (Lee, 2015) that can be found here: http://pycantonese.org. This resource also incorporates another Cantonese speech corpus focusing on Hong Kong Cantonese that features a somewhat similar design, the HKCanCor (K.K. Luke and May L.Y. Wong, 2015). Because of the similar design choices, HKCanCor would probably be a fitting counterpart for a comparative study of the Hong Kong and Perak Cantonese speech communities.

Several research projects involving MYCanCor are planned in the area of action formation and ascription. As part of these projects, additional transcription data that focuses on the annotation of social action and communicative intention might become available in the future.

## 5. Data access and future updates

MYCanCor is currently hosted at Nanyang Technological University in Singapore and maintained by the author. Pending license restrictions, the dataset is available to the research community in three different forms. Transcription data is available in the form of anonymized CHAT format transcription files. In order to protect the identity of all participants, these .chat transcripts do not contain any references related to the identity of the participants. In addition, all location-specific information such as place names and other personal identifiers have been removed.

Audio data is available in form of .wav files that have been edited to protect the identity of the participants. This includes, for example, the adjustment of the pitch range. Video files are only available as post-edited .mov files with a range of non-reversible chromatic, saturation and blur filters applied. Please refer to Figure 1 for a reference example of the visual information that such post-edited video files contain.

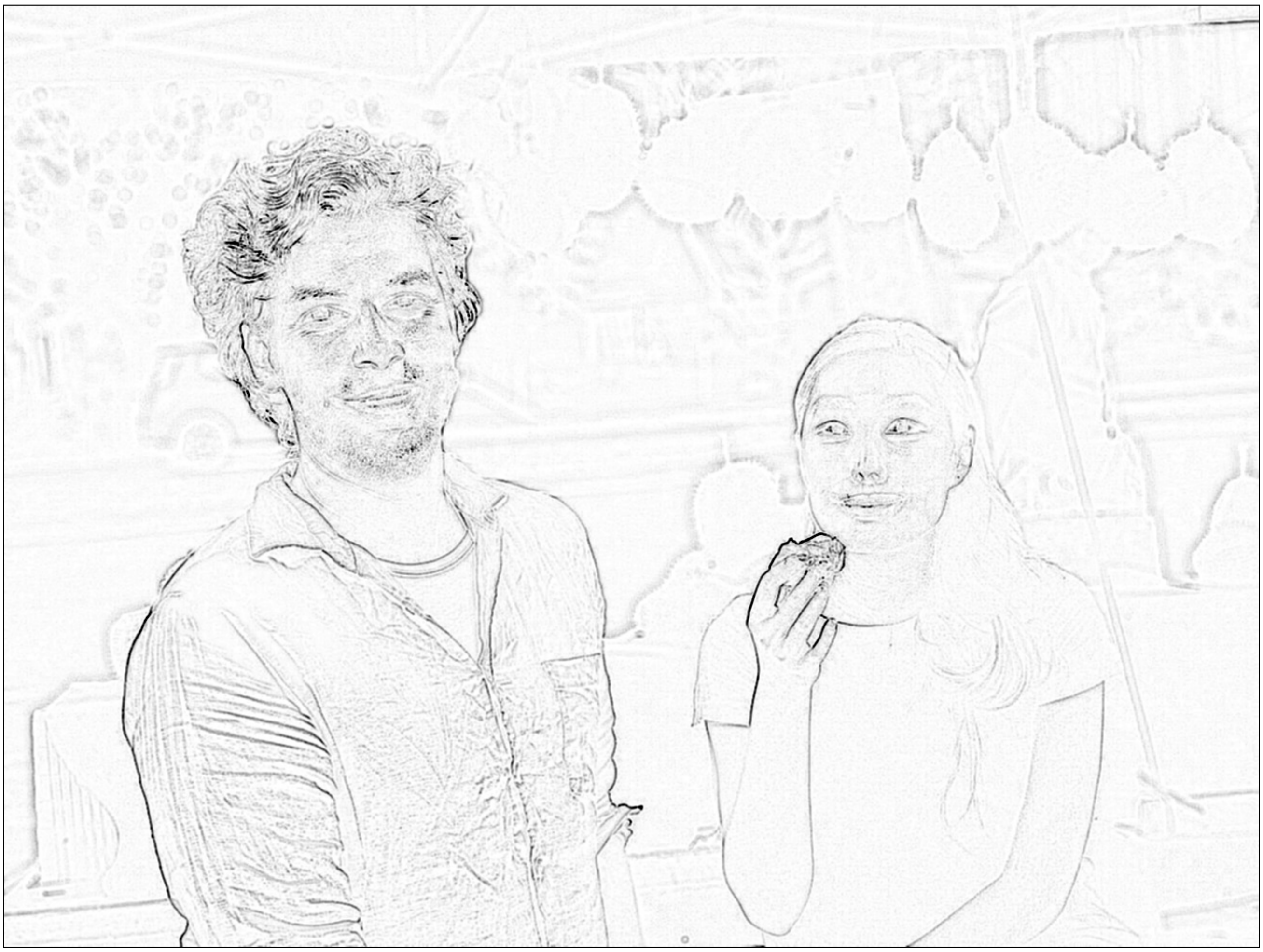

Figure 1: Post-edited visual information in MYCanCor.

For more information regarding licensing please visit http://mycancor.andreasliesenfeld.com. This website also provides further information on how to request access to the dataset, lists all current and planned research projects that involves the use of MYCanCor data and provides a detailed description of all available transcription and annotation files.

Future updates regarding MYCanCor will be made available either on this website or on https://github.com/liesenf/mycancor.

## 6. Bibliographical References

Labov, W. (1972). *Sociolinguistic patterns*. Number 4. University of Pennsylvania Press.

Lee, J. L. (2015). Pycantonese: Cantonese linguistic research in the age of big data. *Talk at the Childhood Bilingualism Research Centre, the Chinese University of Hong Kong*.

MacWhinney, B. (2000). *The CHILDES project: The database*, volume 2. Psychology Press.

Tan, C.-B., (2005). *Chinese in Malaysia*, pages 697–706. Springer, US.

## 7. Language Resource References

K.K. Luke and May L.Y. Wong. (2015). *The Hong Kong Cantonese Corpus: Design and Uses*. Journal of Chinese Linguistics, Monograph no. 25.