Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2906–2913
Marseille, 11–16 May 2020

CantoMap: a Hong Kong Cantonese MapTask Corpus

Grégoire Winterstein, Carmen Tang, Regine Lai
Université du Québec à Montréal, The Chinese University of Hong Kong
Département de Linguistique, Department of Linguistics and Modern Languages
winterstein.gregoire@uqam.ca, tangkm.carmen@gmail.com, ryklai@cuhk.edu.hk

Abstract

This work reports on the construction of a corpus of connected spoken Hong Kong Cantonese. The corpus aims at providing an additional resource for the study of modern (Hong Kong) Cantonese and also involves several controlled elicitation tasks which will serve different projects related to the phonology and semantics of Cantonese. The word-segmented corpus offers recordings, phonemic transcription, and Chinese characters transcription. The corpus contains a total of 768 minutes of recordings and transcripts of forty speakers. All the audio material has been aligned at utterance level with the transcriptions, using the ELAN transcription and annotation tool. The controlled elicitation task was based on the design of HCRC MapTask corpus (Anderson et al., 1991), in which participants had to communicate using solely verbal means as eye contact was restricted. In this paper, we outline the design of the maps and their landmarks and the basic segmentation principles of the data and various transcription conventions we adopted. We also compare the contents of CantoMap to other available Cantonese corpora.

Keywords: Hong Kong Cantonese, corpus, interaction, MapTask, tone merging, word segmentation

1. Introduction

This paper introduces CantoMap, an audio and transcribed corpus of connected speech in Hong Kong Cantonese. Comparable previous works on Cantonese corpora include The Hong Kong Cantonese Adult Language Corpus (HKCAC, (Leung and Law, 2001)), The Hong Kong Cantonese Corpus (HKCanCor, (Luke and Wong, 2015)), the Linguistics Corpus of Mid-20th Century Hong Kong Cantonese (Chin, 2015) and the recordings of Adult-Child sessions from CHILDES (Lee et al., 1994; Fletcher et al., 2000). Some of these corpora were phonetically transcribed with IPA, while others were phonemically transcribed and glossed with parts of speech, and a subset of these were transcribed with Chinese characters. Some were manually segmented, others were semi-automatically segmented and some are not segmented at all. Thus, although each of these resources is precious in its own right, none of them is transcribed in the same way, and each lacks information that appears in the others. In addition, at the present moment, none of these corpora offers access to the recordings of the conversation transcribed in the form of quality audio files. The size of each corpus is also rather limited, and to date, there is still a lack of resources to apply data-intensive techniques of NLP to Hong Kong Cantonese. Addressing those issues is one of the goals of CantoMap, i.e. adding to the list of existing resources and offer a corpus that brings all levels of analysis in a single resource: access to the audio material as well as to aligned transcriptions at the phonemic, narrow phonetic, and orthographic level (using Chinese characters). Besides providing additional data, we designed the recording sessions in ways that allow the elicitation of linguistic patterns of interest. Specifically, we replicated part of the setting used in the HCRC MapTask corpus (Anderson et al., 1991). This was chosen to gather data on phonetic phenomena related to tone sequences, as well as data in the dialogical domain, especially about the use of Cantonese Sentence Final Particles in dialogue. In Sec. 2 , we describe these two aims in more detail, leading to Sec. 3 which describes the design of the corpus and the organization of recording sessions. In Sec. 4 , we describe the post-processing of the data of recording sessions, and Sec. 5 gives various quantitative measures about CantoMap. We conclude in Sec. 6 by listing future directions for CantoMap.

2. Purpose of CantoMap

Here, we describe the two main research objectives that determined the design of CantoMap. Note that we do not report results about these objectives in this paper, but only mention them for explanatory purposes when describing the structure of CantoMap.

2.1. Phonology

One of the goals of this corpus is to provide public access to the audio and phonemic transcriptions of spontaneous speech for researchers to investigate phonological phenomena in Hong Kong Cantonese. Among these, tone merging is of particular interest. Contemporary analyses of Hong Kong Cantonese distinguish six lexical tones, three of which are level tones (Tone 1: High level, Tone 3: Mid level, Tone 6: Low level), two are rising (Tone 2: High rising, Tone 5: Low rising) and one is a falling tone (Tone 4: Falling) (Matthews and Yip, 2011).[1] Many researchers have observed that the two rising tones (Tones 2 and 5) along with tones 3 and 6, and 4 and 6 are undergoing merging in modern Hong Kong Cantonese (e.g. Kej et al. (2002; Bauer et al. (2003; Yiu (2009; Mok and Wong (2010)). These studies investigated the production perception of the rising tones in experimental settings, yet it is unclear how they are produced in spontaneous and natural settings. As described in the next section, the task undertaken by the participants in the recordings was designed in order to elicit the production

[1] Though some descriptive studies mention 9 tones (Hashimoto, 1972), the three additional tones, dubbed "entering" tones, have the same prosodic quality as the level tones (e.g. contour and pitch levels) and their salient characteristic is segmental rather than tonal (i.e. the presence of a coda in a syllable). There is thus little incentive to treat them as separate tones. In addition, it is often remarked that the Cantonese spoken in mainland retains a seventh, high-falling tone, that has already disappeared in Hong Kong.

of the two rising tones in controlled, yet spontaneous contexts, thus providing a large amount of minimal pairs. In addition, other common phonological phenomena that are often found in connected speech such as syllable contractions instantiated by consonant deletion and syllable fusion can be investigated. From a more general perspective, information about Cantonese phonotactics and restrictions of tone cooccurrences can also be extracted from this corpus. Another aspect of CantoMap is that words are segmented, and this allows for a more fine-grained investigation of the effect of domains in which phonological processes occur.

2.2. Discourse Strategies

One characterizing feature of Cantonese is its large inventory of Sentence Final Particles (SFP). These elements number between 30 and 40 and appear in the right periphery of discourse units (i.e. sentences or utterances). SFP carry a wide range of semantic and pragmatic functions, and though not strictly obligatory, about two-thirds of utterances sport at least one of these elements (Kwok, 1984; Matthews and Yip, 2011; Winterstein et al., 2017). Among these, a number of SFP are adequately analyzed in a conversational perspective, i.e. by describing their effect in a dialogical setting in terms of notions such as grounding (Clark, 1996) or putting a content at issue or not (Horn, 2016). Typical SFP of that sort are the assertive SFP *ge3* and *gaa3* and the "softening" SFP *aa3* (Winterstein et al., to appear), or SFP related to the epistemic status of conversation participants such as *wo3* and *lo1* (Luke, 1990; Hara and McCready, 2017).

While some studies have looked at the use of these SFP in attested data (Kwok, 1984; Luke, 1990), the data used were not specifically created to elicit interactions with potentially complex grounding situations, and mostly consisted of free, open conversations, such as those found during call-in programs on radio shows. Because of its design, CantoMap offers a flurry of situations involving articulate grounding events. This is because the conversational situation involves potential mismatches and uncertainties in the information shared by participants, leading to regular explicit requests for grounding material. CantoMap also offers a measure of the degree of acquaintance between participants, thus allowing a study of how this dimension affects the use of SFP (e.g. to test claims regarding the effect of the perceived authority of a speaker on their SFP use Winterstein et al. (2017)).

Finally, beyond the use of SFP, and for the same reasons as above, CantoMap also offers many occurrences of clarification requests as well as self and other-repair, all of which are core elements when studying the grammar of dialogue (Ginzburg, 2012).

3. Design of CantoMap

CantoMap was designed by following the general setup used for the HCRC MapTask corpus (Anderson et al., 1991). The material in the corpus thus corresponds to conversations between two participants with asymmetrical roles: the Giver and the Follower. Both participants were given maps with various landmarks and names indicated in Chinese characters. The maps given to the participants differed in some controlled aspects (the choice of landmarks and the names of some landmarks). The Giver's map showed a path that the Follower had to replicate on their map following the instructions given verbally by the Giver. The participants were unable to make eye-contact and use gestures, but otherwise were free to communicate in any way they wanted. Each participant was recorded using two Sony PCM-D100 recorders, and the data was saved in wav format.

3.1. Maps

The landmarks in the maps were used to elicit pronunciation of nonce words of certain tone sequences of interest. They were nested into the maps and distributed evenly into four sets of maps. Each set consisted of 8 unique target landmarks and eight unique fillers, and there were two items representing each tone combination. There was only one minimal pair, and the two words were identical in segmental features and only differed in one tone. The landmarks on the maps were represented both graphically and orthographically in Chinese. The images were downloaded from various opensource web libraries (most notably http://clipart-library.com/), and the label of each landmark was located directly below the image. The route of the map was controlled for its complexity across the four maps as each of them had 15 90-degree turns. The maps were printed in grayscale colors.

The participants were explicitly told that the maps they received were not identical at the beginning of their session, but it was up to them to discover how the two maps differed. Each map consisted of landmarks labelled with their intended labels. The landmarks were considered as *common* if the identical form and label appeared in the identical location on both the Giver's and Follower's map. Landmarks which were not common differed in one of three ways:

- *Absent or Present* Features were found on one map but not the other;
- *Name Change* Landmarks were identical in image and location but had different labels on the two maps;
- *2:1* Landmarks appeared twice on the Giver's map, once in a position close to the route and once more distant, while the Follower had only the distant irrelevant one.

All map routes began with a starting point, marked on both maps and finishing point was marked only on the Instruction Giver's map. Both start and end points were adjacent to a common landmark. The landmark labels were designed in such a way that they can be used to test for phonological effect in connected speech, e.g. the pronunciations of tones 2 and 5 which are undergoing merging.

Figure 1 shows an example of a "Master" map, i.e. one that includes all the information of the map it (i.e. both that of Giver and Follower). Both the Giver and the Follower map were later produced from the Master for each of the four map sets (labeled A to D).

3.2. Repetitions and Trick Landmarks

In order to test for any differences in careful and casual connected spontaneous speech, two of the stimuli (one target

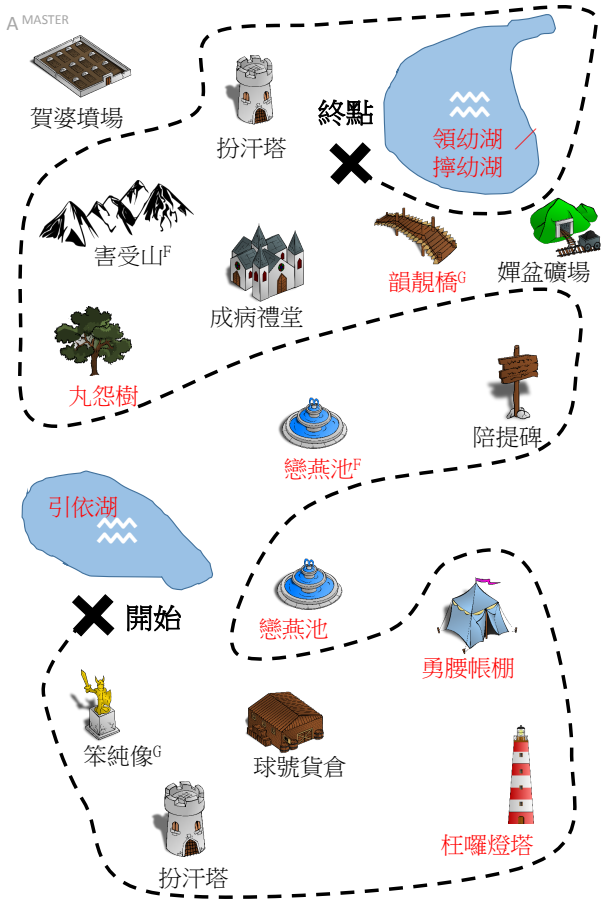

Figure 1: A map example. Landmark names in red correspond to material that was controlled. Landmark names that appear only on the F(ollower) or G(iver) map are indicated with the relevant superscript.

and one filler) were repeated. For one pair, the repeated landmarks were placed far from each other on the map (condition: `distant`), while for the other pair they were placed close to each other (condition: `close`). Out of the four maps that each pair of participants were given, two had close repetitions, and two had distant repetitions. The prediction is that while a Giver is providing information on the repeated landmark (located differently from its first mention), the Follower would ask for clarification on its pronunciation as the repeated occurrence of the same landmark should not be expected. This was designed as such in order to elicit any acoustic differences in casual and careful speech production and to investigate whether the distance between close and distant repetitions have an effect on the degree of acoustic difference between casual and careful speech in a spontaneous setting. The repetitions of the landmarks were set up as follows (Table 1).

A *trick landmark* was included in each set of maps for a similar purpose. A trick landmark refers to a landmark that was represented by the same image on both Giver's and Follower's map, but whose labels differed across maps. The different labels were minimally different in that they only contrasted in T2/T5. The purpose of the trick landmarks is twofold. First, we aim to test whether pairs of landmarks that were minimally different in one merging tone was perceptually detected by either party; second, whether minimal pairs were produced with acoustic differences given their merging status. The distribution of the trick landmarks and the tone sequences of which they were made up are summarized in Table 2.

Map	Close	Distant
A	T2T3	T6T6
B	T4T4	T2T1
C	T5T1	T6T4
D	T4T6	T5T5

Table 1: Repetitions of tone sequences on four Maps

Map	Trick Landmarks
A	T2T1/T5T1
B	T2T3/T5T3
C	T2T1/ T5T1
D	T2T3/T5T3

Table 2: The types of tone combinations used as the trick landmark for each map

Both the Giver's and the Follower's map had some *missing* items. Each of them was missing a target landmark and a filler landmark. The missing items must not be the same on the G's and the F's map as all items need to appear in one of these maps. Since there are 4 tone combinations types of targets and 4 types of fillers, each type was missing on one map, see Table 3.

Fill/Target	Map	Missing in G	Missing in F
Target	A	T5T3	T2T3
	B	T2T3	T2T1
	C	T5T1	T5T3
	D	T2T1	T5T1
Filler	A	T6T4	T6T6
	B	T4T6	T6T4
	C	T6T6	T4T4
	D	T4T4	T4T6

Table 3: The missing tone combinations in each map

3.3. Recording Procedure

Each pair of participants completed all four maps. Each participant took turns to be the Giver. They were given instructions that the goal of the task was to draw the route of Giver's map on the follower's map through verbal collaboration. The two participants were seated across from each other with approximately 1.5 m apart. A cardboard screen was placed between the two participants to prevent any communication by eye contact and gestures. Each participant was recorded with a Sony PCM-D100 recorder at the sampling rate of 44 100 Hz. Two recorders were used since the pair of participants were sitting at a distance. Both sound tracks are available in `CantoMap`. The duration of the task varies, possibly due to the friendship status of the

participants with whom they were paired up. The mean duration of recording for each pair is 37.25 min, the mean for friends is 30.78 min and the mean for strangers is 39.63 min. It should be noted that the numbers of participants who are friends and of those that who are not are not balanced. Table 4 provides a summary of the recording duration and their degree of acquiantance.

In addition to the actual map task, participants were asked to read aloud a wordlist containing all the names of the landmarks from the maps after they completed the main task. Each name of the landmark appeared twice in randomized order. The participants were instructed to read aloud the list slowly and carefully.

Participant #	Gender	Degree of acquaintance	Duration (mm:ss)
1 and 2	F and F	Strangers	64.27
3 and 4	F and F	Strangers	64.12
5 and 6	M and F	Strangers	110.39
7 and 8	F and F	Friends	36.37
11 and 12	M and M	Strangers	29.35
13 and 14	F and F	Strangers	41.40
15 and 16	F and M	Friends	42.55
17 and 18	M and F	Strangers	26.23
19 and 20	F and F	Strangers	29.44
21 and 22	F and F	Friends	28.44
23 and 24	F and F	Strangers	52.31
27 and 28	F and M	Friends	21.15
29 and 30	F and F	Strangers	33.25
31 and 32	F and M	Strangers	41.10
33 and 34	F and M	Strangers	18.21
35 and 36	M and F	Strangers	13.02
37 and 38	F and F	Friends	24.23
39 and 40	F and F	Strangers	18.38
41 and 42	F and F	Strangers	21.39
43 and 44	F and F	Strangers	29.45

Table 4: Length of recording and degree of acquaintance for each pair of participants in `CantoMap`.

3.4. Participants

Forty participants were recruited. Thirty-eight of them were native speakers of Hong Kong Cantonese and two were native speakers from Guangzhou, China. They were students of the Education University of Hong Kong at the time of recording, aged between 18-29. Nine of them were male participants. The two Guangzhou speakers were not identified until after the recording had ended. We included them in the corpus as they each interacted with our target participants, but researchers should take their status into consideration when using our data. With that said, their Cantonese had no noticeable differences compared to the rest of the participants'. Ten of them were acquainted with their partners in the task before they joined the experiment.

Participants gave their consent for the release of their data after being instructed about the project and its goals, and they all signed a form to that effect. The project received ethical approval from the Human Research Ethics Committee of the Education University of Hong Kong (project number RG- 61/2014–2015R).

4. Data Processing

The recordings were first manually transcribed into Chinese characters, and manually word segmented. The audio and the transcripts were aligned at utterance level using ELAN (Max Planck Institute for Psycholinguistics, 2019). Phonemic transcriptions in the form of Jyutping romanization were then produced automatically by matching words in our corpus with the entries in the Cantonese dictionary `yedict https://writecantonese8.wordpress.com/2012/02/04/cantonese-cedict-project/`. For all the unknown terms in the corpus (including our landmarks), we either manually enriched the dictionary, or reconstructed their Jyutping romanization by using a MaxMatch algorithm searching for the biggest known sub-parts of the term in `yedict`. The processed data is presented on one tier of segmented Chinese characters (the "orthographic" tier) and a tier of romanized transcription. Each of these tiers exists for the three speakers involved in the recording: the *experimenter* (E), the *giver* (G) and the *follower* (F). Each ELAN file only contains one unique giver and follower, which means participants did not switch roles within the same file.

In the following subsections we provide additional details about the information present in each tier.

4.1. Orthographic Transcription

Orthographic transcriptions are provided in traditional Chinese characters, and each Chinese character is mapped to a syllable. Although written Cantonese is not standardized, transcribing them into characters was not much of a problem as most of the words are cognates with Standard Chinese (Mandarin), which we represented with the standard characters. In this section, we discuss some of the conventions we used in orthographic transcriptions.

4.1.1. Written Cantonese

For words which are unique in Cantonese, we largely follow the conventions used in (Matthews and Yip, 2011), (Chin, 2015) and in the media. However, we did not distinguish the active passive use of *bei2* (俾 and 畀) by characters as some researchers do (e.g. (Leung and Law, 2001)). We used 俾 for both forms. In cases where a syllable does not have a clear corresponding character, e.g. *soe4* ('to slide' as in *soe4lok6heoi3* 'slide down'), *doeng1* (as in *zim1doeng1* 'tip (of an object)', *bat1* 'to scoop'), and *tan3* 'to budge' (*tan3hau6* 'to budge backwards'), we opted transcribe with Jyutping signalled by a `&` sign.

4.1.2. Sentence Final Particles

We also decided to not transcribe SFPs with characters for two reasons. First, there is relatively little consensus on which characters to be used to represent each of these SFPs. Second, many of the SFPs which differ in tones and meanings are written with the same character, e.g. the three SFP *wo3/wo4/wo5* are all represented by the character 喎 (even though their meaning is strikingly different). As mentioned,

the use of SFPs is one of our research objectives, so we kept all SFP orthographically distinct by using the same convention we outlined above for syllables which cannot be written by Chinese characters, e.g. &wo3 for *wo3*. This will make the extraction of specific SFPs easier in further investigations. Furthermore, some of the SFPs were transcribed more narrowly with *r* as in *&gr3*. We used *r* to refer to schwa [ə] as they occur in contexts which are felicitous for both *gaa3* and *ge3*. We leave these instances open to future researchers' interpretation.

4.1.3. Code-mixing

Code-mixing of Cantonese and English is a common language phenomenon in Hong Kong (e.g. (Gibbons, 1987), (Li, 2000)), in which our participants were also observed to engage occasionally. We transcribed the instances of English words used in normal English orthography in both the character and phonemic tiers, see example (1).

(1) 引意水塘 嘅 exactly 下面 # 南方 就係
landmark GEN exactly below south is
射頹餐室&aa4
landmark SFP
jan5ji3seoi2tong4 ge3 exactly haa6min6 # naam4fong1 zau6hai6 tui4se6can1sat1 aa4
'*The reservoir is exactly below... south of the restaurant.*'

4.1.4. Phonological Contractions

Contractions are common in casual connected speech. Some of the frequently occurring ones are readily accepted and have specific characters assigned to represent them in written Cantonese. Examples include *ji6sap6* 二十 'twenty', which is contracted to *jaa6* 廿, *mat1je5* 乜嘢 'what' contracted to *me1* 咩. Cases like these are transcribed with their conventionally accepted contracted forms, 廿 (2) and 咩 (3) in the corpus.

(2) 廿九 號
twenty-nine number
jaa6gau2 hou6
'*Number twenty nine.*'

(3) 咩 城堡 &waa2 咩 &lai4 &gaa3
what castle SFP what SFP SFP
me1 sing4bou2 waa2 me1 lai4 gaa3
'*What castle? What is it?*'

Other cases of contractions are less frequent or do not have a consistent orthographic representations. For example, 即係 *zik1hai6* 'then it is' are sometimes contracted to *ze1hai6* 姐係 or *ze1* 姐, but even when the contracted forms were produced, they were transcribed the full form, not as 姐係 or 姐. This was done to facilitate the identification of the underlying form. The phonemic transcription of these cases was manually handled to ensure it matched the actual production of the participants.

4.1.5. Special Symbols

We only used two special symbols in the transcription, one to mark pauses `#`, see example (1) and the other to mark unintelligible speech `xxx` (4).

(4) 上面 仲 有 一 個 xxx
above still have one CL
soeng6min6 zung6 jau5 jat1go3 xxx
'*There's another xxx above here.*'

4.2. Phonemic Transcription

Cantonese is a tonal language, and each syllable contains an optional onset, a nucleus formed by either a monophthong, a diphthong or a syllabic consonant, an optional coda and an obligatory lexical tone. `CantoMap` follows the Jyutping romanization scheme developed by the Linguistics Society of Hong Kong (`https://www.lshk.org/jyutping`) to transcribe phonemic information. Jyutping can straightforwardly be mapped back to an IPA phonemic transcription. See Table 5 for the conversion of Jyutping to IPA.

C Jyutping	C IPA	V Jyutping	V IPA	T Jyutping	T in Chao's letter
p	p^h	*i*	i	*1*	55
t	t^h	*yu*	y	*2*	25
k	k^h	*u*	u	*3*	33
b	p	*e*	ɛ	*4*	21
d	t	*oe*	œ	*5*	23
g	k	*eo*	ɵ	*6*	22
c	ts^h	*o*	ɔ		
z	ts	*a*	ɐ		
m	m	*aa*	a		
n	n				
ng	ŋ				
l	l				
w	w				
j	j				
h	h				
f	f				
s	s				
kw	k^{wh}				
gw	k^w				

Table 5: Conversion of Jyutping to IPA

4.3. Segmentation

Cantonese, like Standard Chinese, does not separate words with spaces in orthography. What constitutes a word is still a controversial topic in Chinese linguistics, however, it is clear that a character which often is associated to a morpheme, does not always form a word. We adopted the segmentation rules proposed by Academia Sinica (Huang et al., 2017), and made necessary changes to meet the morpho-syntactic parameters of Cantonese. We have also made reference to (Matthews and Yip, 2011) when making decisions about word boundaries.

4.3.1. Basic Principles

Strings of characters were segmented based on the following basic principles.

- *Compositionality* Character strings with meanings which cannot be derived compositionally were treated

as one word. These refer to cases where individual characters can stand alone as words but when combined, their meanings differ from their literal meanings. For example, *fong1hoeng3* 方向 (*lit.* location-towards) 'direction', *gaak3lei4* 隔離 (*lit.* separate-part) 'next to', *lou6sin3* 路線 (*lit.* road-line) 'route'.

- *Derivational Affixes* Strings of characters were treated as words if affixation results in semantic or word class change. For example, the suffixation of *dei2* 地 to a reduplicated adjective *ce3* 斜 'slanted' gives the distinct meaning 'slanted-ish'.

- *Frequency* High frequency combinations were treated as compounds, e.g. *mat1je5* 乜嘢 what-thing 'what', *zeoi3hau6* 最後 superlative-behind 'last'.

- *Productivity* We also considered the productivity of the combinations as one of the segmentation criteria. For sets of morphemes that have limited size, we treated their members as words. For example, the sets of morphemes indicating directions are limited to the combinations of north/east/south/west, and we treated these as words, e.g. *dung1naam4* 東南 'southeast', *sai1bak1* 西北 'northwest' were all individual words.

Below are some examples of strings of characters that were segmented into separate words.

- When a word is inserted with aspect markers, classifiers etc., they were segmented into separate words, e.g. *zyun3waan1* 轉彎 'to turn' is a word, but when an aspect marker is inserted as in *zyun3 zo2 waan1* 轉 咗 彎 'turned' or when a classifier is inserted as in *zyun3 go3 waan1* 轉 個 彎 'to turn once', each of the morpheme is segmented as a word.

- When words combine and their meanings are compositional, e.g. *jyun4gung2 jing4* 圓拱形 'arch shape'.

In the subsections below, the segmentation criteria will be discussed with regards to some specific constructions.

4.3.2. A-not-A

A-not-A constructions are used in Cantonese to form yes-no questions. There are two treatments for A-not-A constructions following Academia Sinica's segmentation guidelines. When A is a modal verb, the construction is not segmented due to the fact that set of modals is limited in size. For example, *wui5m4wui5* 會唔會 'will-not-will', *ho2m4ho2ji3* 可唔可以 'can-not-can', *jing1m4jing1goi1* 應唔應該 'should-not-should' are all treated as a single units.

When A is not a modal verb, the construction was treated as separate words. Thus, when A is a verb or an adjective, e.g. *heoi3 m4 heoi3* 去 唔 去 'go-not-go', they are segmented as three separate words. Multisyllabic verbs or adjectives that are formed by a sequence of bound morphemes are exceptions to this rule. They are kept as a single unit in these constructions. An example with bound morphemes is: *zung1m4zung1ji3* 鍾唔鍾意 'like-not-like'.

4.3.3. Suffixation

As mentioned above, we mainly made decisions for segmentation based on whether the affixation induces semantic or word class change. When the meaning is altered after affixation, we kept the characters in the new construction as a word, but when the meaning of combination can be derived compositionally, we segmented the characters as different words. For example, *ze2* is a suffix which roughly translates to 'person', similar to the '-er' agentive suffix in English. When attached to the word *gan1ceoi4* 跟隨 'follow', as in *gan1ceoi4 ze2* 跟隨 者, the meaning 'follower' is derived. The above example contrasts with *gei3ze2* 記者'journalist'which resembles the structure of V-ze2, however, the meaning of *gei3ze2* is different from the literal meanings of the two morphemes (record-person), and is therefore treated as one single word. The combinations containing *dim2* 點 'point'and *sin3* 線 'line'as suffixes are treated as follows. When appearing in high frequency combinations such as *zung1dim2* 終點 'destination', *hei2dim2* 起點 'starting point', they are treated as single words. In cases where the suffixes appear in technical terms such as *saam1dim2gung6sin3* 三點共線 'colinearity', they are not segmented into separate words. When these suffixes occur in structures like *waan1 sin3* 彎 線 'curved line'and *zung1gaan1 dim2* 中間 點 'middle point'), they are segmented due to their compositional meaning.

4.3.4. Verbal Particles

Aspect markers are separated from the verb, and treated as a single word. This is because this kind of construction is highly productive, and the particles could be added to any verbs. For example: both *waak6 jyun4/zo2/gan2/saai3* 畫 完/ 咗/ 緊/ 晒 'draw-ASP' and *zou6 jyun4/zo2/gan2/saai3* 做 完/ 咗/ 緊/ 晒 'do-ASP' can bear four distinct aspect particles.

Other verbal particles such as directional particles *faan1* 返 'back', *maai4* 埋 'close-by', *gwo3* 過 'pass-by', *soeng5* 上 'up', *lok6* 落 'down' are considered as individual words and they can be combined with *heoi3* 去 'away' in post-verbal position. Such combinations are considered as one word, e.g. *haang4 faan1heoi3* 行 返去 'walk back', *haang4 lok6heoi3* 行 落去 'walk down'. (5) is an example with two directional particles *faan1* and *seong5* following the main verb *dau1*.

(5) 兜 返 上去機場 嘅 上方
detour back up airport GEN above
dau1 faan1 soeng5heoi3 gei1coeng4 ge3 soeng6fong1
'*Go back up to the location above the airport.*'

4.3.5. Sentence Final Particles

For clusters of SFPs, we only treated aa1maa3 and laa3wo3 as bi-syllabic particles, the others were segmented as combinations of two or more particles, such as the sequence of ga3 (laa3) bo3.

5. Corpus information

5.1. Basic statistics and comparison with other resources

Table 6 summarizes basic counts of tokens and types in CantoMap.

	Tokens	*Types*
Chinese characters	121,189	1,046
Word	106,197	2,361

Table 6: Basic statistics about CantoMap

As can be seen in table 7, the majority of Chinese words in CantoMap are mono or bisyllabic.

Word Length (in syll.)	no. of tokens
1	69,237
2	18,734
3	4,330
4	3,205
5+	379

Table 7: Frequency of Chinese words of different lengths in CantoMap

Table 8 compares the information available in CantoMap (including the information currently under development for its second version) to the information found in other major Cantonese corpora.

Corpus[2]	**Pm**	**Pn**	**Ch**	**Seg.**	**PoS**	**Aud.**	**Size**
CantoMap	✓	v2	✓	✓	v2	✓	106k wds
HKCAC	×	✓	✓	×	×	×	170k chars
HKCanCor	✓	×	✓	✓	✓	×	150k wds
Mid-20	×	×	✓	✓	×	×	140k wds
Canto. TB	✓	×	✓	✓	✓	×	14k wds
LWL	✓	×	✓	✓	✓	✓	1M wds
HKU-70	✓	×	✓	✓	✓	✓	186k wds

Table 8: Comparison of the information in CantoMap and other Cantonese corpora. **Pm**: Phonemic level, **Pn**: Phonetic level, **Ch**: Chinese characters transcription, **Seg.**: segmented, **PoS**: Part of Speech tagged, **Aud.**: audio available.

[2]References for the corpora are as follows. HKCAC: (Leung and Law, 2001), HKCanCor: (Luke and Wong, 2015), Mid-20: (Chin, 2015), Canto. TB (Wong and Leung, 2016), LWL: (Lee et al., 1994), HKU-70: (Fletcher et al., 2000).

5.2. Availability

CantoMap is available on a github repository at the following address: https://github.com/gwinterstein/CantoMap. CantoMap is released under the GNU General Public License v3.0. Future versions of the resource will be hosted in the same place, and contributions by interested parties are welcome.

6. Conclusion and outlooks

The primary purpose in releasing CantoMap is to provide additional Cantonese spoken data to the research community. There are around 62 million Cantonese speakers in China (Asher and Moseley, 2018), and around 5 million in Hong Kong (HKSAR, 2019). Even though among Chinese languages, Cantonese has the second largest speaker population after Mandarin, its accessible language resources are relatively scarce when compared to other languages that have a comparable size of speaker population, such as Italian. To this end, CantoMap provides an additional 13 hours of quality audio recording, word-segmented orthographic and phonemic transcriptions of the speech produced by young Cantonese speakers between the ages 18 to 29. The corpus adds to the existing resources by providing more recent interactive conversational data with specific relevance to the ongoing tone-merging phenomenon and the use of SFP. We have described how the design of the map task has enabled the elicitation of structured, controlled, but also natural and spontaneous spoken data. In terms of data processing, we have introduced the conventions we adopted for the orthographic transcription and word segmentation. These tasks have proven to be non-trivial compared to languages which have a standard writing systems like Mandarin Chinese and orthographic systems such as English, which explicitly demarcate word boundaries. We also have discussed how we obtained phonemic transcriptions automatically by exploiting existing resources, such as yedict. Currently, work is underway to add information to CantoMap. As mentioned in the paper, the goal is for CantoMap to offer information at all levels. The next release of the corpus will add a narrow phonetic transcription tier, along with a Part of Speech tag tier. For that latter part, there exists no consensual set of PoS tags for Cantonese. Existing works either used a custom set (HKCanCor, (Luke and Wong, 2015)), or relied on the set of PoS from the Universal Dependencies project (Wong et al., 2017). Yet another option would be to directly use those used for tagging Mandarin Chinese. Deciding for the optimal set to use remains an open question so far.

7. Bibliographical References

Anderson, A. H., Bader, M., Gurman Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. S., and Weinert, R. (1991). The Hcrc Map Task Corpus. *Language and Speech*, 34(4):351–366.

Asher, R. E. and Moseley, C. (2018). *Atlas of the world's languages*. Routledge.

Bauer, R. S., Cheung, K.-H., and Cheung, P.-M. (2003). Variation and merger of the rising tones in Hong Kong

Cantonese. *Language Variation and Change*, 15(2):211–225.

Clark, H. H. (1996). *Using language*. Cambridge University Press, Cambridge.

Fletcher, P., Leung, S. C.-S., Stokes, S. F., and Weizman, Z. O., (2000). *Cantonese pre-school language development: A guide.* Hong Kong University, Department of Speech and Hearing Sciences, Hong Kong.

Gibbons, J. (1987). *Code-mixing and code choice: A Hong Kong case study*, volume 27. Multilingual Matters Clevedon.

Ginzburg, J. (2012). *The interactive stance: meaning for conversation*. Oxford University Press, Oxford.

Hara, Y. and McCready, E. (2017). Particles of (Un)expectedness: Cantonese Wo and Lo. In M. Otake, et al., editors, *New Frontiers in Artificial Intelligence. JSAI-isAI 2015. Lecture Notes in Computer Science*, volume 10091, pages 27–40. Springer, Berlin.

Hashimoto, O.-k. Y. (1972). *Phonology of Cantonese*, volume 1. Cambridge University Press.

HKSAR. (2019). Thematic Household Survey Report No. 66. Technical report, Census and Statistics Department, The Government of the Hong Kong Special Adminstrative Region.

Horn, L. R. (2016). Information Structure and the Landscape of (Non-)-at-issue Meaning. In Caroline Féry et al., editors, *The Oxford Handbook of Information Structure*, pages 108–127. Oxford University Press.

Huang, C.-R., Hsieh, S.-K., and Chen, K.-J. (2017). *Mandarin Chinese words and parts of speech: A corpus-based study*. Routledge.

Kej, J., Smyth, V., So, L. K., Lau, C., and Capell, K. (2002). Assessing the accuracy of production of Cantonese lexical tones: A comparison between perceptual judgement and an instrumental measure. *Asia Pacific Journal of Speech, Language and Hearing*, 7(1):25–38.

Kwok, H. (1984). *Sentence Particles in Cantonese*. Center of Asian Studies, University of Hong Kong.

Lee, T., Wong, C., Leung, S., P., M., A., C., Szeto, K., and Wong, C. (1994). The Development of Grammatical Competence in Cantonese-speaking Children. Technical report, RGC, Hong Kong.

Leung, M.-T. and Law, S.-P. (2001). HKCAC: The Hong Kong Cantonese Adult Language Corpus. *International Journal of Corpus Linguistics*, 6(2):305–325.

Li, D. C. (2000). Cantonese-English code-switching research in Hong Kong: A Y2K review. *World Englishes*, 19(3):305–322.

Luke, K. K. and Wong, M. L. (2015). The Hong Kong Cantonese Corpus: Design and Uses. *Journal of Chinese Linguistics*, 25:312–333.

Luke, K. K. (1990). *Utterance Particles in Cantonese Conversation*. John Benjamins Publishing Company, Amsterdam, Philadelphia.

Matthews, S. and Yip, V. (2011). *Cantonese: A Comprehensive Grammar*. Routledge, 2nd edition, dec.

Max Planck Institute for Psycholinguistics, (2019). *ELAN*. Nijmegen. version 5.7.

Mok, P. P.-K. and Wong, P. W.-Y. (2010). Perception of the merging tones in Hong Kong Cantonese: Preliminary data on monosyllables. In *Speech Prosody 2010-Fifth International Conference*.

Winterstein, G., Lai, R., Luk, Z., and McCready, E. (2017). Authority and Gendered Speech: Cantonese Particles and Sajiao. CSSP 2017.

Winterstein, G., Lai, R., and Luk, Z. P.-s. (to appear). Softness, assertiveness and their expression via Cantonese sentence final particles. In Elin McCready, et al., editors, *Discourse Particles in Asian Languages*. Routledge.

Wong, T.-s., Gerdes, K., Leung, H., and Lee, J. (2017). Quantitative Comparative Syntax on the Cantonese-Mandarin Parallel Dependency Treebank. In *Proceedings of the Fourth International Conference on Dependency Linguistics*, pages 266–275, Pisa, Italy.

Yiu, C. Y.-m. (2009). A preliminary study on the change of rising tones in Hong Kong Cantonese: An experimental study. *Language and Linguistics*, 10(2):269–291.

8. Language Resource References

Chin, Andy. (2015). *A Linguistics Corpus of Mid-20th Century Hong Kong Cantonese*. LML Department, The Education University of Hong Kong, 2.0.

Fletcher, P. and Leung, S. C-S. and Stokes, S. F. and Weizman, Z. O. (2000). *HKU-70*. 1.0.

Lee, T.H.T. and Wong, C.H. and Leung, S. and Man. P. and Cheung A. and Szeto, K. and Wong, C.S.P. (1994). *Lee/Wong/Leung Corpus*. 1.0.

Leung, Man-Tak and Law, Sam-Po. (2001). *HKCAC: The Hong Kong Cantonese Adult Language Corpus*. Hong Kong University, 1.0.

Luke, Kang Kwong and Wong, May L.Y. (2015). *HKCanCor: The Hong Kong Cantonese Corpus*. 1.0.

Wong, Tak-sum and Leung, Herman H.M. (2016). *Cantonese-HK UD treebank*. The City University of Hong Kong, 1.0.