Corpus information:

Name of the corpus:	The Hong Kong Cantonese Child Language Corpus (CANCORP) [2012 version]
Investigator(s) involved:	Thomas Hun-tak Lee (Principal Investigator, CUHK), Colleen H. Wong (HKPU) and Samuel Cheung-Shing Leung (then HKU, now THEi); Patricia Yuk-hing Man, Alice Shuk-yee Cheung, Kitty Ka-sin Szeto, and Cathy Sin-Ping Wong.
Research students:	Patricia Yuk-hing Man, Alice Shuk-yee Cheung, Kitty Ka-sin Szeto, and Cathy Sin-Ping Wong.
Year of establishment of the corpus:	1996, revised 2012.

The nature and characteristics of the corpus and how it may be used: [1]

CANCORP is a set of audio-recordings of the longitudinal language development of eight Cantonese-speaking children who were each observed for around 12 months. The beginning age of observation was between 1;07 and 1;11 for four of the children, and between 2;02 and 2;08 for the remaining four children; the age at which observation ended was between 2;07 and 3;08. The mean number of observation sessions for each child, with each session lasting approximately one hour, was 21. Four of the child subjects were male, and the other four female.

The transcripts of CANCORP, consisting of 171 transcripts, were coded according to the CHAT format (Codes for the Human Analysis of Transcripts), and tagged with 33 parts-of-speech labels. CANCORP was made publicly accessible in 1996 and deposited in the CHILDES archive.

There have been several versions of the corpus. The original version of CANCORP was released in 1996 (‘The 1996 version of CANCORP’). This version went through further checking and corrections and has since been updated and revised. The updated standard

[1] This CANCORP description is based on a paper entitled “Longitudinal child Cantonese corpus: An update” presented by Thomas Lee at the Roundtable Conference on Linguistic Corpus and Corpus Linguistics in the Chinese Context held at the Hong Kong Institute of Education on May 6-8, 2011, with subsequent revisions by Margaret Lei.

version was released in 2012 (‘The 2012 updated standard version of CANCORP’) and comes with two encodings: a Big5 (asc) encoding and a Unicode (uni) encoding. The transcripts can be downloaded here: http://www.arts.cuhk.edu.hk/~lal/corpora.html.

Another version of the corpus, in the form of a zipped file ‘LeeWongLeung.zip’ under the East Asian Corpora of the CHILDES database (‘The CHILDES version of CANCORP’), received additional processing due to the work of Paul Fletcher’s research group at HKU. The transcripts of this version of the corpus, which contain parts of speech (POS) tags laid out in a different format than that of the 2012 updated standard version, are not identical to those of the latter on this website. Differences between them can be found in the detailed description of the corpus below.

Publications using CANCORP data should cite either of the following sources:

Lee, Thomas H.T., Colleen H. Wong, Samuel Leung, Patricia Man, Alice Cheung, Kitty Szeto, and Cathy S. P. Wong. 1996. *The Development of Grammatical Competence in Cantonese-speaking Children.* Report of RGC earmarked grant 1991-94.

Lee, Thomas H.T. and Colleen H. Wong. 1998. CANCORP: the Hong Kong Cantonese Child Language Corpus. *Cahiers de Linguistique Asie Orientale* 27(2):211-228.

1. Various versions of CANCORP

1.1 The 1996 version of CANCORP

The CANCORP corpus grew out of a project entitled "The development of grammatical competence in Cantonese-speaking children" funded by the Hong Kong Research Grants Council for the period 1991-93, and based at the Chinese University of Hong Kong. The project represented a joint effort of three local universities: The Chinese University of Hong Kong, The Hong Kong Polytechnic University, and The University of Hong Kong.

The audio recordings consisted of conversational exchanges between the child subjects and adults, mostly involving the investigators talking to the children, often in the presence of other members of the family or other adults. Morphemic transcriptions of the data were produced based on these audio-recordings.[2]

The transcriptions were given in Chinese characters, and rendered in romanizations only for sentence final particles and other words whose morphemic status could not be uniquely represented by means of Chinese characters. The transcriptions followed the format of CHAT (Codes for the Human Analysis of Transcripts) (MacWhinney 1991) and were tagged with 33 parts-of-speech labels. The computer encoding for the Chinese characters was Big5, with Cantonese-specific characters made available by the Hong Kong Supplementary Character Set (HKSCS).[3] Except for the children's names, the romanizations are based on the Cantonese romanization scheme of the Linguistic Society of Hong Kong (LSHK) (1997, 2002), which can be accessed at: http://www.lshk.org/jyutping.

The original transcripts contained three tiers for each utterance, with Chinese characters in the main tier, and below it the tags corresponding to the word-like units in the main tier, followed optionally by other comment lines.

1.2 The CHILDES version of CANCORP

The current CANCORP on CHILDES is available under the Cantonese folder of the East Asian section of the downloadable corpora in a file named 'LeeWongLeung.zip'. These files, encoded in Unicode, contain a set of transcripts with utterances in Chinese each followed by a tier of tags aligned with romanized forms. The romanization of the Chinese characters in the transcripts was done by an automatic conversion software.

Given the fact that a Chinese character may correspond to more than one morpheme and have more than one pronunciation, there was no one-to-one correspondence between a Chinese character and its romanization. The conversion process resulted in more than one romanized form for a Chinese character, which posed problems for researchers working exclusively with romanized forms of the Chinese characters. A group of researchers led by

[2] The task was performed by three research students: Patricia Yuk-hing Man (for two of the children CGK and WBH), Alice Shuk-yee Cheung (for three of the children CKT, MHZ, and LTF), and Kitty Ka-sinn Szeto (for another three of the children CCC, HHC and LLY).

[3] The character set is downloadable from: http://www.ogcio.gov.hk/ccli/eng/hkscs/introduction.html.

Paul Fletcher, then at the University of Hong Kong (HKU), made an effort to disambiguate the romanizations.

The tagged lines in the files of 'LeeWongLeung.zip' were adapted from the original POS tagging of the 1996 version of CANCORP, which was done exclusively by the CANCORP team based at the Chinese University of Hong Kong. The work of the HKU team was limited to disambiguation of the romanized forms of Chinese characters (presumably based on linguistic context and not on the actual audio recordings), and on alignment of tags with the disambiguated romanized forms in a format commonly used in POS tagging. It should be emphasized that the CANCORP team was not involved in any of the additional processing work done in the CHILDES version of CANCORP involving romanization disambiguation, and hence we cannot vouch for the accuracy of the disambiguated romanizations in this version of CANCORP.

1.3 The 2012 updated standard version of CANCORP

A number of changes have been made to the 1996 version of CANCORP.[4] First, given that some of the Chinese characters have more than one romanized form as there does not exist one-to-one correspondence between a Chinese character and its romanization, the romanized forms of Chinese characters in the romanized tier of CANCORP were disambiguated to the extent this was possible, based on audio check and contextual information.

Some explanations are in order to illustrate the decisions that the CANCORP team had to make in revising the transcripts. For example, the original romanizations provided by machine conversion in the 1996 version of CANCORP for the Cantonese words '上' and '去' carry two possible pronunciations rendered as 'soeng5^soeng6 heoi2^heoi3'. However, as the two words can only be pronounced unambiguously as 'soeng5 heoi3' when they are used together to mean "go up", this form of romanization was used throughout in the revised transcripts. We

[4] The transcripts of four of the children: CKT, LTF, LLY and MHZ were checked against the audio recordings in the years 2001-2003 by a team of student assistants of the City University of Hong Kong (Barbara Ching-man Lee, Samantha Chan, Kin Sum Wong, Apple Leung, Billy Wong, Jasmine Yung, Coco Wong, Kwan-kit Li, Bowie Cheng, and Florence Fung). The transcripts of the other four children: CCC, CGK, HHC and WBH, were checked in 2011-2012 by the following research assistants of Chinese University of Hong Kong: Margaret Lei, Eva Lai, Kristen Cheng, and Mandy To. Since the checking of the transcripts were carried out by two different teams, there might be some discrepancies in the criteria for disambiguating romanization and the extent of corrections made between transcripts revised by the two teams.

also corrected a few of the romanizations carrying tonal information that did not conform to the Cantonese pronunciation of the character, e.g. ‘waan4^waan6’ was corrected as ‘waan2’ for ‘玩’.

Corrections were also made to the main tier and/or the romanization tier when the corresponding character or the romanized form could not be properly shown, such as when a simplified character was used that could not be recognized by the machine during the romanization conversion procedures, e.g. ‘衭’ (“trousers”). In addition, the number which marks tonal information next to the Chinese characters in the main tier was deleted, as such information could be retrieved from the corresponding romanization tier. For example:

[Original]

*CHI:	整	爛爛	唔	得	嘅 3.
%mor:	zing2	laan6laan6	m4^ng4	dak1	ge33.
morpheme:	make	break.break	NEG	can	NOM
translation:	‘(You) are not allowed to break (it).’				

[Revised]

*CHI:	整	爛爛	唔	得	嘅.
%mor:	zing2	laan6laan6	m4	dak1	ge3.
morpheme:	make	break.break	NEG	can	NOM
translation:	‘(You) are not allowed to break (it).’				

Some romanizations in the 1996 CANCORP transcripts were found to be inconsistent with the LSHK *Jyutping* romanization, for example, the use of ‘dz’ (for ‘z’), ‘ts’ (for ‘c’), ‘eu’ (for ‘oe, eo’), ‘eou’ (for ‘eu’) and ‘i' (for ‘ji’). These inconsistencies were removed in the 2012 CANCORP.

In many instances, the exact transcription cannot be determined from the transcript itself. A unified romanization was adopted arbitrarily based on the common form of pronunciation, for consistency considerations. Some examples of unified romanizations of kinship terms and interjections are shown in Table-1 and Table-2 respectively.[5]

[5] The task of unifying romanizations was primarily carried out on the transcripts of four children: CCC, CGK, HHC and WBH.

Table-1. Examples of unified romanizations of kinship terms

Kinship terms	Unified romanization
哥哥 “elder brother”	go4go1
姐姐 “elder sister”	ze4ze1
弟弟 “younger brother”	dai4dai2
妹妹 “younger sister”	mui4mui2
公公 “grandfather”	gung1gung1
婆婆 “grandmother”	po4po1
姨姨 “aunt”	ji1ji1

Table-2. Examples of unified romanizations of interjections

Interjections	Unified romanization
哎吔	aai1jaa3
哦	o4
喂	wai3
嗱	laa4
噢	o4

The indeterminacy of transcription was also encountered with some of the sentence final particles. The appropriate tone in the romanized form was selected either based on audio check or the linguistic context that suits the corresponding form best. Some variants of the sentence final particles sharing the same Chinese character are illustrated in Table-3:

Table-3. Examples of variants of sentence final particles represented by the same Chinese character

Example	Romani-zation	Meaning
*INV: 我 幫 妳 搣 呀, 好 唔 好? Ngo5 bong1 nei5 mit1 aa1, hou2 m4 hou2? I help you tear SFP, yes NEG yes? ‘Should I help you tear (something)?’ / ‘Do you want me to help you tear (something)?’	aa1	Suggestion
*INV: 係 啊. Hai6 aa3. yes SFP ‘Yes.’	aa3	Statement
*INV: "唔該 晒" 咁 好 呀? “M4.goi1 saai3” gam3 hou2 aa4? thank.you completely so good aa4 ‘It was so nice of (someone) to say “thank you so much!” ’	aa4	Surprised

On the basis of these audio and context checks, a number of transcription and part-of-speech errors were corrected, and situational contextual information that was not included in the original transcripts was filled in. Some minor corrections were also made for consistency and accuracy.

The 2012 version of CANCORP consists of the revised transcripts of the eight children in two versions:

a) The Chinese TAG version, with each utterance transcribed in Chinese characters in the main tier, and POS tags in a subsidiary tier followed by other optional comment tiers.

b) The Romanized TAG version, with each utterance transcribed in Chinese characters in the main tier, followed by a tier of romanized forms of the Chinese characters, a tier of corresponding POS tags, and other optional comment tiers.

Below is information concerning the background of the eight child subjects and the POS tag set used in the transcripts.

2. The background of the 8 Cantonese-speaking children

CCC was born in Hong Kong and was the only son in the family. His father was a businessman and his mother taught English in a secondary school. Both of the parents are monolingual Cantonese speakers. They lived with the child's maternal grandparents. He had not started going to a nursery during the period of data collection.

CGK is female and was brought up in a monolingual Cantonese-speaking working class family. The parents of CGK were both born in Hong Kong. CGK's father was a technician in an electronic company and her mother was a housewife. They lived with the child's grandmother. The child was not yet enrolled in a nursery during the whole period of data collection. She was entirely taken care of by her mother.

CKT was born in Hong Kong and was the only son of the family. His father was a Census & Survey Officer working with the government and his mother a secondary school teacher teaching Chinese and Religious Studies. Since his birth, he had been living in his maternal grandparents' house during weekdays and was taken care of by his grandmother. His parents visited him occasionally during the weekday evenings and took him back home on Friday nights to stay over the weekend. They communicated in Cantonese. When CKT was 1 year 10 months old, his mother went to study for a year in the United Kingdom. He started to attend a nursery at the age of 2 years 1 month.

HHC was born in Hong Kong and was the youngest child in the family. He had an elder sister who was seven years older. His father was an engineer and her mother was a typist. Both of the parents are monolingual Cantonese speakers. The family employed a Thai helper, who spoke Cantonese to the children. He had not started going to a nursery during the period of data collection.

LTF was born in Hong Kong and was the youngest child in the family. She had a sister who was four years older. Her father was an engineer working with the government and her mother was a piano teacher teaching at home. During the first one-and-a-half years from her birth, she was taken care of mostly by a Filipino helper while her mother worked as a school music teacher. After her mother had stopped working in school, LTF was mostly taken care of by her mother, except at times when her mother had to give piano lessons or had to go out, when the child would be looked after by her Filipino helper. LTF communicated in Cantonese except when speaking to her Filipino helper, for which she used 'something English-like' (as described by her mother). She started to attend kindergarten at the age of 2 years 9 months.

LLY was also born in Hong Kong and was the youngest child in the family. She had an elder brother who was ten years older and an elder sister who was four years older. LLY's father was a businessman and her mother was a housewife. Both of the parents are monolingual Cantonese speakers. The family employed a Filipino helper, who spoke some Cantonese and English to the children.

MHZ was born in Kent, United Kingdom and was brought back to Hong Kong at the age of eight and a half months old. He was the only son of the family. His father was a lecturer of the Hong Kong Polytechnic, and his mother an English language lecturer of the Chinese University of Hong Kong. He was then taken care of by his maternal grandmother at her house until the age of about 1 year 1 month. From that time to the age of 2 years 6 months, he was taken care of by a caretaker on weekdays. He communicated in Cantonese, though his parents occasionally introduced to him some English terms. He started to attend the nursery play-groups at the age of 2 years 6 months.

WBH is female and was also brought up in a monolingual Cantonese-speaking family. WBH's father worked in the warehouse of a mass transport company and her mother was a part-time piano teacher. The child had a younger brother who was about two years younger. They lived with the child's grandmother and uncle. The child had already started attending a nursery school when data collection started. After school, she was taken care of by her parents and grandmother.

3. Parts-of-speech (POS) tags

The words in the utterances were tagged with 33 parts-of-speech labels. It should be noted that the issue of word boundaries is a complex one in the analysis of Chinese languages and dialects. Each utterance in a transcript was segmented into word-like units which may be free forms (like nouns and verbs) or bound forms such as many of the adverbs, affixes, connectives and particles. The design and assignment of tags primarily took into account the needs of grammatical analysis related to the research project that gave rise to the corpus.

It should also be noted that adult criteria were used in assigning POS tags to word-like units. In other words, a word assigned the label of a certain syntactic category means that the word would have been assigned that category in the adult language. The tags in the transcripts should not be taken to mean that the children have acquired the syntactic categories represented by the tags.

Below is a summary list of the syntactic categories used in coding the corpus.

Table-4. List of syntactic categories used in the CANCORP transcripts

Category	Examples
1. adj = adjective	紅 *hung4* ‘red’
2. advf = focus adverb	仲 *zung6* ‘still’ 都 *dou1* ‘also’ 又 *jau6* ‘again’ 再 *zoi3* ‘again’
3. advi = adverb of intensity	好 *hou2* ‘very’ 幾 *gei2* ‘a bit’ 咁 *gam3* ‘so’ 真 *zan1* ‘really’
4. advm = adverb of manner	慢慢 *maan6maan2* ‘slowly’ 麻麻哋 *ma4ma4dei2* ‘not great’
5. advs = sentential adverb	不如 *bat1jyu4* ‘how about’ 咁(樣)*gam2(joeng2)* ‘in this manner, so’ 一齊 *jat1cai4* ‘together’
6. asp = aspectual marker	咗 *zo2* (perfective) 住 *zyu6* (durative) 緊 *gan2* (progressive) 過 *gwo3* (experiential) 開 *hoi1* (habitual)

7. aux = auxiliary / modal verb	應該 *jing1goi1* (should) 肯 *hang2* (willing) 可以 *ho2ji5* (can) 會 *wui5* (will) 駛 *sai2* (need)
8. cl = classifier	個 *go3* (classifier$_{\text{GENERAL}}$) 隻 *zek3* (classifier$_{\text{ANIMAL}}$) 本 *bun2* (classifier$_{\text{BOOK}}$) 杯 *bui1* (classifier$_{\text{CUP}}$) 啲 *di1* (classifier$_{\text{PLURAL}}$)
9. com = comparative morpheme	過 *gwo3* ‘than’ (as in 大過 *dai6 gwo3* ‘bigger than’) 啲 *di1* ‘a bit more’ (as in 紅啲 *hung4 di1* ‘a bit redder’)
10. conj = connective	但係 *daan6hai6* ‘but’ 同埋 *tung4maai4* ‘and’ 或者 *waak6ze2* ‘or’
11. corr = correlative	越…越 *jyut6...jyut6* ‘the more…the more’ 又…又 *jau6...jau6* ‘and…and…’
12. ctc = clitic	倒 *dou2* ‘attain’ 到 *dou3* ‘extent’
13. det = determiner	呢 *nei1* ‘this’ 嗰 *go2* ‘that’ 第 *dai6* (ordinal)
14. dir = directional verb	落 *lok6* ‘down’ 上 *soeng5* ‘up’ 出 *ceot1* ‘out’ 入 *jap6* ‘in’ 嚟 *lai4* ‘come’
15. ex = expressive utterance	拜拜 *baai1baai3* ‘goodbye’ 早晨 *zou2san4* ‘good morning’
16. gen = genitive marker	嘅 *ge3* ‘nominalizer’

17. ins= emphatic inserted marker	鬼 *gwai2* (as in 好鬼靚 *hou3 gwai2 leng3* ‘so beautiful’)
18. nn = noun	蘋果 *ping4gwo2* ‘apple’ 爸爸 *baa4baa1* ‘father’
19. nnloc = locative noun phrase	上面 *soeng6min6* ‘above’ 裏面 *leoi5min6* ‘inside’
20. nnpr = pronoun	我 *ngo5* ‘I’ 你 *nei5* ‘you’ 佢 *keoi5* ‘s/he’
21. nnpp = proper name	天凡 *tin1faan4* ‘Tin Fan’ 駿駿 *zeon3zeon3* ‘Jun Jun’
22. neg = negative morpheme	唔 *m4* ‘not be’ 咪 *mai2* ‘do not’ 冇 *mou5* ‘not have’
23. prt = post-verbal particle	返 *faan1* ‘again’ 晒 *saai3* ‘all’ 親 *can1* ‘every time’ 埋 *maai4* ‘also’ 過 *gwo3* ‘before’ 吓 *haa5* (tentative marker)
24. prep = preposition	同埋 *tung4maai4* ‘also’ 喺 *hai2* ‘in’ 俾 *bei2* ‘let’
25. q = quantifier	一 *jat1* ‘one’ 三 *saam1* ‘three’ 十 *sap6* ‘ten’ 幾 *gei2* ‘few’ 每 *mui5* ‘every’
26. rfl = reflexive pronoun	自己 *zi6gei2* ‘oneself’

27. sfp = sentence final particle	喇 *laa3* ‘change of situation or past event’ 咖嘛 *gaa1maa3* ‘for emphasis’ 呢 *ne1* ‘how about?’ / ‘where?’
28. vd = ditransitive verb	擺 *baai2* ‘put’ 俾 *bei2* ‘give’
29. verg = ergative verb	跌 *dit3* ‘drop’
30. vf = function verb	係 *hai6* ‘be’ 有 *jau5* ‘have’ 喺 *hai2* ‘at’
31. vi = intransitive verb	笑 *siu3* ‘smile/laugh’
32. vt = transitive verb	推 *teoi1* ‘push’
33. wh = wh words	乜 *mat1* ‘what’ 乜嘢 *mat1je5* ‘what’ 點 *dim2* ‘how’ 點解 *dim2gaai2* ‘why’ 點樣 *dim2jeong2* ‘how’

4. Research theses related to CANCORP project:

Man, Yuk-hing Patricia. 1993. *Subject-object distinctions and empty categories in child Cantonese.* MPhil thesis, Hong Kong Polytechnic University.

Cheung, Shuk-yee Alice. 1995. *Acquisition of Wh-words by Cantonese-speaking childre*n. MPhil thesis, Hong Kong Polytechnic University.

Szeto, Ka-sinn Kitty. 1998. *The acquisition of Cantonese classifiers*. MPhil thesis, University of Hong Kong.

5. Acknowledgments

The creation of this corpus was made possible by a three-year grant to Thomas Hun-tak Lee (Chinese University of Hong Kong, PI), Colleen H. Wong (Hong Kong Polytechnic University), and Samuel Cheung-Shing Leung (University of Hong Kong) [RGC earmarked grant CUHK 335/95H]. The project was supported by two studentships from the Hong Kong Polytechnic awarded to Patricia Man and Alice Cheung, and a studentship from the University of Hong

Kong awarded to Kitty Szeto. In addition, funding for the later stages of the project was provided by a direct grant from Faculty of Arts, Chinese University of Hong Kong, a grant from the Freemason's Fund for East Asian Studies, as well as research assistantships from the Hong Kong Polytechnic University. The subsequent checking of the CANCORP transcripts was supported in part by a faculty start-up grant of the City University of Hong Kong (2000-2003) and one from the Chinese University of Hong Kong (2005-2007), as well as an RGC grant entitled "Issues of continuity in early lexical and syntactic development" [CUHK4583/06H]. The support of these funding agencies is hereby gratefully acknowledged.

We are grateful to the computer help and technical advice we received from Professor Brian MacWhinney of Carnegie Mellon University at different stages of the preparation of CANCORP for archiving in CHILDES.

Feedback, comments and queries on CANCORP would be most welcome.

Thomas Hun-tak Lee, Principal Investigator
The CANCORP project
huntaklee@cuhk.edu.hk
December, 2014