

# Cantonese LSP for LC-STAR II

Language:	HK Cantonese
Responsible LC-STAR II partner:	Nokia

# Table of Contents

1. Introduction.....	1
2. Features of lexical items in Cantonese.....	1
3. Orthography.....	2
3.1. Tokenization.....	2
4. Speech and syllabification.....	4
4.1. Consonants.....	4
4.2. Monophthongs and diphthongs.....	4
4.3. Tones.....	7
4.4. Tone change.....	8
5. Morphology.....	9
6. Grammar.....	9
7. References.....	11
Appendix I: Conversion table from IPA to SAMPA.....	13
Appendix II: Language Specific DTD.....	16

## 1. Introduction

As one of the southern dialects of the Chinese language, Cantonese is one of the most popular linguistic varieties spoken in Hong Kong and it is also the most common linguistic varieties that are used in daily communication. Currently, over 90% (over 6 million) of the Hong Kong population use Cantonese as their home language. Cantonese is also spoken all over the world outside Hong Kong.

Cantonese is mainly used in the spoken register and thus, no standard writing system has ever been devised for the dialect. Written Chinese used in Hong Kong publications, such as newspapers, magazines, books, official documents are based on Modern Standard Chinese (MSC). The Cantonese dialect may be used in some less formal or informal contexts such as personal Blog, chatroom, ICQ/QQ, MSN, SMS, as well as in some magazines. Therefore, the present so-called “Cantonese corpus” comprises both spoken Cantonese and written Chinese used in Hong Kong.

## 2. Features of lexical items in Cantonese

For historical reasons, Hong Kong has close contact with foreign culture and particularly the impact of English on Cantonese is not small. Thus, besides Chinese characters (including those colloquial items), there are quite a few loanwords used in contemporary Cantonese which are mainly borrowed from English. Most of them are transliterated with Chinese characters by mapping the pronunciations of the source language to the Cantonese pronunciations. Typical examples include 巴士 (bus), 的士 (taxi), 海洛英 (heroin), 士巴拿 (spanner), and 貼士 (tips).

Foreign personal names or place names are always transliterated with Chinese characters. However, it should be noted that the transliterations are usually different from those used by other Chinese communities such as Mainland China because the transliteration involves the mapping of the pronunciations from the source language to the target languages (i.e. Cantonese or Mandarin). For example, “Hollywood” is transliterated as 荷里活 in Cantonese but 好萊塢 in Mainland China because 活 (*wu:t* in IPA) in Cantonese sounds similar to the ending of “wood” in Hollywood. Moreover, the ending consonant of the character 活 (-t) corresponds to the stop ending of “wood” in “Hollywood”. “Beckham” is transliterated as 碧咸 in

Cantonese because 咸 has the –m ending in Cantonese which corresponds to its source “ham” in “Beckham”.

Besides, code-mixed items as well as words written in foreign languages are commonly found in Cantonese (or Hong Kong Chinese which includes both written and spoken forms). Typical examples include CD 碟 (CD disk), BB 機 (beeper). Some popular brand names such as LV, Chanel, Coca-cola, GAP, HP, are even written in their original languages directly without any transliteration.

### 3. Orthography

Cantonese is written with traditional Chinese characters which are encoded with Big-5 code. Cantonese shares most of the lexical items with MSC on the written level, such as 天 (heaven), 地 (earth), 人 (man), 電話 (telephone), and 學校 (school), though often with different pronunciations. On the other hand, due to the frequent use of colloquial Cantonese in informal writing in recent years, the Hong Kong government has devised the Hong Kong Supplementary Character Set (HKSCS) to include around 5000 characters that are found in Hong Kong Cantonese which have not yet been supported by the Big-5 coding system.<sup>1</sup>

#### 3.1. Tokenization

Unlike alphabetic languages, such as Indo-European languages, there is no word delimiter in the Chinese writing system. Therefore, the words cannot be obtained directly from the corpus. In other words, before extracting words from the corpus, we need to do **tokenization**. Tokenization in fact is a basic and the first step in Chinese natural language processing.

Tokenization of the present Cantonese corpus is carried out first by **auto-tokenization**. The auto-tokenization process is based on the principles of tokenization implemented in the LIVAC corpus maintained by the Language Information Sciences Research Center, City University of Hong Kong.<sup>2</sup> The auto-tokenized corpus then goes through **human verification** since there can be multiple readings of the same sentence whose

---

<sup>1</sup> <http://www.info.gov.hk/digital21/chi/hkscs/introduction.html>

<sup>2</sup> LIVAC has been contributing test corpus to the International Chinese Language Processing Bakeoff under SIGHAN since 2003. A reference guideline on the LIVAC tokenization can be found at the SIGHAN website.

correct reading cannot be simply determined by machine. For example, the sequence of characters 大同區長安西路 can be tokenized as:

- (a) 大同區 / 長安西路 (West Chang'an Road of the Datong district)
- (b) 大同區長 / 安西路 (An Xilu, the mayor of the Datong district)

Human verification of the auto-tokenized texts is necessary not just for disambiguation but also to identify new words that have not yet appeared in the dictionary.

The procedures for tokenization of the Cantonese corpus can be summarized as follow:

- (1) Automatic word tokenization
- (2) Text normalization: Remove digits, foreign language characters/words<sup>3</sup>
- (3) Human verification of the result of the auto-tokenization.
- (4) Extract the words from the processed corpora

---

<sup>3</sup> Punctuation marks are not segmented in our tokenization process.

## 4. Speech and syllabification

Cantonese is a monosyllabic language in the sense that each character is represented by one syllable.

### 4.1. Consonants

There are 20 initial consonants and only p, t, k, m, n and ŋ can be used as final consonants in Cantonese. These 20 consonants are tabled below<sup>4</sup>:

p	p <sup>h</sup>	m	f
t	t <sup>h</sup>	n	l
ts	ts <sup>h</sup>	s	j
k	k <sup>h</sup>	ŋ	
k <sup>w</sup>	k <sup>wh</sup>	w	
0			h

### 4.2. Monophthongs and diphthongs

Cantonese is well-known for having long-short distinction in its vowel system. The vowels can be divided into two major groups: Monophthongs and diphthongs. These two groups are tabled below:<sup>5</sup>

---

<sup>4</sup> The transcription used here is based on the IPA system which can be found in Zee (2001). A conversion table of the IPA into SAMPA is attached at the end of this document.

<sup>5</sup> See Zee (2001), p.59.

Monophthongs (in IPA)	Examples
a:	家 [ka:]
ɛ:	車 [ts <sup>h</sup> ɛ:]
œ:	靴 [hœ:]
ɔ:	梳 [so:]
u:	夫 [fu:]
i:	屍 [si:]
y:	豬 [tsy:]

The above monophthongs are long vowels (represented by the colon) and can stand alone to form syllables. There are four more monophthongs that are short vowels and have to form syllables with other vowels or consonants.

e	西 [sei]
ɐ	恤 [sɐt]
ɪ	色 [sɪk]
ʊ	風 [fʊŋ]

The above monophthongs can combine with other monophthongs or one of the six final consonants to form syllables. The table below lists diphthongs of Cantonese transcribed with IPA.<sup>6</sup>

---

<sup>6</sup> See Zee (2001), p.59.

Diphthongs (in IPA)	Examples
ei	西 [sɛi]
a:i	街 [ka:i]
ɐu	周 [tseu]
a:u	交 [ka:u]
ei	機 [kei]
ɛ:u	[tɛ:u ] a colloquial word for “to throw”
əy	水 [səy]
ɔ:i	鯉 [sɔ:i]
ou	高 [kou]
i:u	要 [ji:u]
u:i	每 [mu:i]

There are certain finals representing some colloquial pronunciations in Cantonese which have no character realization. For example, [ɛ:u] is a diphthong of the syllable [tɛ:u] for the colloquial pronunciation of 掉 (to throw). [ɛ:n] is a final used to transcribe the Cantonese pronunciation [fɛ:n si:] (written as “fan 屎”) of the English word *fans*.

In Cantonese, syllabification is well established. The possible syllable structures in Cantonese are:<sup>7</sup>

V      烏  
 VC     安  
 GVC    人  
 GV     右  
 CV     書  
 CVC    十

Only one consonant is allowed as an initial (i.e. no consonantal cluster) or ending consonant in Cantonese.

#### 4.3. Tones

Cantonese is a tonal language bearing 6 tonal contours. The 6 tonal contours are plotted on a 5-point scale with 5 being the highest and 1 the lowest. The 6 tonal contours can be described as follows:

Tonal contours	Descriptions	Examples
55 <sup>8</sup>	high level	詩 識
35	high rising	史
33	mid level	試 洩
11	low level	時
13	low rising	市
22	low-mid level	事 十

識, 洩 and 十 are entering tone characters (入聲字) with final stop endings -k, -t, -p respectively.

<sup>7</sup> V stands for both vowels and diphthongs while G for glides such as /w/, /j/ in the initial.

<sup>8</sup> The 55 tone is similar to the first tone in Mandarin. It is claimed that Cantonese has 53 instead of 55 for the *yin-ping* tone (陰平). However, in Hong Kong Cantonese, the *yin-ping* tone is now generally pronounced with a level tone rather than a falling tone.

#### 4.4. Tone change

Unlike Mandarin, there is no *regular* tone change in Cantonese. In Mandarin, tone change is phonologically driven. The typical example is that when two third tones appear together, the first one has to undergo tone change and becomes a second tone. The other typical examples are the character *yi* (一) and *bu* (不). The tone change of these two characters thus depends on the tone of the character following it.

Cantonese also has tone change but the phenomenon is not as regular as that in Mandarin. Generally speaking, tone change mainly takes place for those low tone characters. Thus, only tonal contours of 33, 11, 13, 22 can undergo tone change and the resulting tone is always 35 (陰上). However, the tone change in Cantonese is not phonologically driven. There is no regular phonological rule that can be spelt out when the character can undergo tone change. And even for the same character, not every Hong Kong Cantonese speaker will pronounce it with the changed tone. For example, 鴨 (duck, *a:p3*) can become *a:p2* or have no tone change at all and this varies from person to person.<sup>9</sup>

When the character that can undergo tone change forms compound word, the character in question can be pronounced:

- (a) with the **original tone only**;
- (b) with **either** the original tone **or** the changed tone;
- (c) with the **changed tone only**.

Let us take 魚 (fish, *jy:11*) as an example. When 魚 (fish) appears in the word 鯨魚 (whale), only the original tone is adopted by most speakers. For the word 鯊魚 (shark), some speakers will pronounce it with the original tone (i.e. 11) while some with the changed tone (i.e. 35). Finally, for the word 金魚 (gold fish), it is almost exclusively pronounced with the changed tone (i.e. 35) but rarely with the original tone.

The same phenomenon can be found for 房 (room). Generally speaking, the character 房 (room) in 睡房 (bedroom), 廚房 (kitchen), 書房 (study room) can undergo tone change to become *fɔ:ŋ35* (the citation tone is *fɔ:ŋ11*). However, for words like 殮房 (mortuary) and 車房 (garage), the character in question seldom

---

<sup>9</sup> Notice that in general, tone change only takes place on the character that occurs at the end of the compound or phrase. Thus, 鴨腳 (duck leg) does not have tone change on 鴨 (duck) even for those speakers who have tone change for 鴨 (duck).

undergoes tone change.

The tone change phenomenon is a complicated issue and cannot be formulated easily. To reflect the actual scenario of the tone change phenomenon, the followings will be adopted for the transcription of the lexicon:

- (a) For **single character**, both the original tone and changed tone will be given even though the character is always pronounced with the changed tone. Thus, the character 魚 (fish) will be transcribed with both *jy:11* and *jy:35*.
- (b) For **compound words**, only the tone that is observed will be transcribed. As shown by the above examples, 金魚 (gold fish) will be transcribed with the changed tone, i.e. *kam55 jy:35* while for 鯨魚 (whale), *k<sup>h</sup>ɿŋ11 jy:11* will be provided.
- (c) If the character in question can be pronounced with either the citation or changed tones within a given word, then both pronunciations will be included in the lexicon. Thus, for 鯊魚 (shark), both *sa:55 jy:35* and *sa:55 jy:11* will be shown in the lexicon.

## 5. Morphology

It is well known that Cantonese is an isolating language without any significant morphology. Therefore, unlike Indo-European languages, Cantonese does not have distinctions in number, tense, gender, case. Thus, when parsing a Chinese sentence, it is not necessary to do any morphological analysis as in parsing English sentences. On the other hand, there is no need to carry out lemmatization. Each segmented word can be regarded as a lemma.

## 6. Grammar

In the Chinese language, all words do not have any morphological change no matter how they are used. In this sense, all words and their part-of-speech, unlike western languages, do not have any relation with the Number, Gender, Person, Case. Words carrying the same syntactic functions will be assigned with the same POS tag no

matter it is extracted from Cantonese or MSC sub-corpora.

The following POS tag sets are proposed for Cantonese.

### **NOM (Common and proper nouns)**

- This set is further broken down into the following sub-classes:  
*common*: common nouns; *PER*: Person name; *GEO*: Geographic Name; *COU*: Country; *CIT*: City; *STR*: Street Name; *COM*: Organisation; *BRA*: Brand; *TOU*: Cultural/historic place, *HLD*: Holidays.

### **ADJ (Adjectives)**

### **NUM (Numerals)**

- Since numerals are not included in the lexicon, NUM here refers to those lexical items that indicate order such as 首, 次.

### **MEW (Measure words / classifiers)**

### **VER (Verbs)**

### **AUX (Auxiliary verbs)**

- This includes those verbs showing modality such as 可能, 需要, 應該 ...

### **AUW (Auxiliary words)**

- AUW includes the following 3 types of Chinese auxiliary words: (1) Aspectual words indicating aspects in Cantonese such as 咗, 緊, 開, 住, 親, 翻, 了, 過, 著; (2) Structural auxiliary word 地, 的, 嘅, 得; (3) Comparison auxiliary word 過 (i.e. 我大過他, I am older than him);

### **PRO (Pronouns)**

- PRO covers both personal pronouns 我, 你, 他 and demonstrative pronouns such as 這, 那.

### **ADV (Adverbs)**

- ADV also covers time words and place words like 今天, 昨天 ...

### CON (Conjunctions)

- CON also covers “以” and “而” as in the following sentences:  
以增強總體競爭實力 / 為生存下去而不得不採取的

### ADP (Adpositions)

- ADP is further broken down into 2 sub-classes:  
*PRE*: Preposition (在, 向, 沿著, 從 ...); *POST*: Postposition (上, 下, 前, 後, 東, 南, 西, 北 ...).

### PAR (Particles)

- PAR is further broken down into 2 sub-classes:  
*general*: General particles (第, 者, 們, 所 as in 他所用的是最新的); *SFP*: Sentence Final Particles, which include those particles appearing at the end of sentences to express speakers' emotions, attitudes, e.g. 咩 (mɛ:55), 喎 (wɔ:22, wɔ:13), 啱 (kʷa:22), 噃 (pɔ:22, pɔ:35), 啫 (tsɛ:55, tsɛ:k55).

### INT (Interjections)

### ONO (Onomatopoeias)

### IDI (Idioms)

## 7. References

- [1] Bauer, Robert & Paul Benedict. *Modern Cantonese Phonology*. Mouton: de Gruyter (1997).
- [2] Cheung Hung-nin 張洪年. 《香港粵語語法的研究》. 香港 :香港中文大學出版社 (1972).
- [3] 《廣州方言詞典》. Jiangsu: Jiangsu jiaoyu chubanshe (1998).
- [4] Leung Chung-sum 梁仲森. *A Study of the Utterance Particles in Cantonese as Spoken in Hong Kong* (當代香港粵語語助詞的研究). Hong Kong: Language Information Sciences Research Center, City University of Hong Kong (2005).

- [5] Linguistic Society of Hong Kong Jyutping Editorial Group. *Guide to LSHK Cantonese Romanization of Chinese Characters*. 《香港語言學學會粵語拼音字表》 Second edition. Hong Kong: Linguistic Society of Hong Kong (2002).
- [6] Matthews, Stephen & Virginal Yip. *Cantonese: A Comprehensive Grammar*. London: Routledge (1994).
- [7] Rao Bingcai 饒秉才, Ouyang Jueya 歐陽覺亞 & Zhou Wuji 周無忌. 《廣州話詞典》. Guangzhou: Guangdong renmin chubanshe (1997).
- [8] Snow, Don. *Cantonese as Written Language: The Growth of a Written Chinese Vernacular*. Hong Kong: Hong Kong University Press (2004).
- [9] Yue-Hashimoto, Anne. *Phonology of Cantonese*. London: CUP (1972).
- [10] Zee, Eric. “Chinese (Hong Kong Cantonese)”. *Handbook of the International Phonetic Association*, pp.58-60. Cambridge: CUP (2001).
- [11] Zheng Dingou 鄭定歐. 《香港粵語詞典》. Jiangsu: Jiangsu jiaoyu chubanshe (1997).

## Appendix I: Conversion table from IPA to SAMPA10

IPA symbol	SAMPA symbol	Keyword in SAMPA	Orthography	English gloss
<b>CONSONANTS</b>				
p <sup>h</sup>	p	pa:_3	怕	to fear
p	b	ba:_1	爸	father
t <sup>h</sup>	t	tO:N_4	糖	sugar
t	d	da:_2	打	to hit
k <sup>h</sup>	k	k6p_1	給	to give
k	g	g6m_1	金	gold
k <sup>wh</sup>	kw	kwa:_1	誇	to boast
k <sup>w</sup>	gw	gwa:_1	瓜	melon
m	m	ma:_1	媽	mother
n	n	nej_5	你	you
ŋ	N	Na:_4	牙	tooth
f	f	fa:_1	花	flower
s	s	sa:_1	沙	sand
h	h	ha:_1	蝦	shrimp
ts	dz	dza:k_3	窄	narrow

<sup>10</sup> The list of consonants, monophthongs, diphthongs in IPA is based on Zee (2001).

ts <sup>h</sup>	ts	tʂa:N_2	橙	orange
j	j	ju:t_6	月	month
w	w	wa:k_6	畫	to draw
l	l	la:N_5	冷	cold

### MONOPHTHONGS

i:	i:	di:p_6	碟	plate
y:	y:	hy:t_3	血	blood
ɛ:	E:	sE:_2	寫	to write
œ	9:	g9:k_3	腳	foot
a:	a:	ba:t_3	八	eight
e	6	j6p_6	入	to enter
◁:	O:	gO:n_1	乾	dry
u:	u:	gu:_1	姑	aunt
ɪ	i	sik_1	色	colour
ə	9	s9t_1	恤	sympathy
e	6	s6p_1	濕	wet
ʊ	u	suk_1	叔	uncle

## DIPHTHONGS

a:i	a:j	da:j_6	大	big
ɛi	ɛj	sɛj_1	西	west
ɛu	ɛw	sɛw_2	手	hand
a:u	a:w	ma:w_1	貓	cat
ei	ej	nej_5	你	you
ɛu	Ew	tEw_6	(This colloquial word has no written form)	to throw away
ɛy	ɛy	sɛy_2	水	water
ɔi	ɔj	tsɔ:j_3	菜	vegetable
ui	uj	bu:j_1	杯	cup
i:u	i:w	ti:w_3	跳	to jump
ou	ow	dow_1	刀	knife
<b>Pitch</b>	<b>SAMPA</b>	<b>Keyword in SAMPA</b>	<b>Orthography</b>	<b>Gloss</b>
high level	_1	si:_1 / sik_1	詩 / 識	poem / to know
high rising	_2	si:_2	史	history
mid level	_3	si:_3 / sit_3	試 / 洩	to try / to leak
low level	_4	si:_4	匙	a key
low rising	_5	si:_5	市	a city
low-mid level	_6	si:_6 / sik_6	是 / 食	to be / to eat

## Appendix II: Language Specific DTD

```
<?xml version="1.0" encoding="UTF-16"?>
<!-- Language-independent specification of contents of lexica -->

<!-- Entity Declarations (BEGIN) -->
<!ENTITY % ns "not_specified">
<!ENTITY % pos "NOM | AUW | ADP | ADV | VER |
                INT | PRO | CON | MEW | ADJ | PAR |
                AUX | NUM | IDI | ONO">
<!ENTITY % subdomain "1.1.1. | 1.1.2. | 1.1.3. | 1.1.4. | 1.2.1. |
                    1.2.2. | 1.3. | 1.4. | 1.5.1. | 1.5.2. | 1.6. |
                    3.1.1. | 3.1.2. | 3.1.3. | 3.1.4. | 3.1.5. | 4.1.1. |
                    4.1.2. | 4.1.3. | 4.1.4. | 4.1.5. | 4.1.6. | 5.1.1. |
                    5.1.2. | 5.1.3. | 5.1.4. | 5.1.5. | 5.2. | 6.1.1. | 6.1.2.
                    |
                    6.1.3. | 6.1.4. | 6.1.5. | 6.2.1. | 6.2.2. | 6.2.3. | 6.2.4.
                    |
                    6.2.5. | 6.2.6. ">
<!ENTITY % class_noun "common | PER | GEO | COU |
                    CIT | STR | COM | BRA | TOU | HLD">
<!ENTITY % class_adp "PRE | POST ">
<!ENTITY % class_sfp "general | SFP ">

<!-- Entity Declarations (END) -->

<!ELEMENT LEXICA (ENTRYGROUP)+>
<!ATTLIST LEXICA xml:lang NMTOKEN #IMPLIED>
<!ELEMENT ENTRYGROUP (ALT_SPEL*, (ENTRY | ENTRY_COMP | ABB)+)>
<!ATTLIST ENTRYGROUP orthography CDATA #REQUIRED
                    xml:lang NMTOKEN #IMPLIED >
<!ELEMENT ALT_SPEL (#PCDATA)>
<!ELEMENT ENTRY_COMP (PHONETIC, LEMMA*, ENTRY_EL, ENTRY_EL, ENTRY_EL*,
APP?)>
<!ELEMENT PHONETIC (#PCDATA)>
<!ELEMENT ENTRY_EL (%pos;)>
<!ATTLIST ENTRY_EL orthography CDATA #REQUIRED>
```

```

<!ELEMENT ABB (EXP)+>
<!ELEMENT EXP (ENTRY_COMP | ENTRY)>
<!ATTLIST EXP expansion CDATA #IMPLIED>
<!ELEMENT ENTRY ((%pos;), LEMMA, PHONETIC, APP?)>
<!ELEMENT LEMMA (#PCDATA)>
<!ELEMENT APP (SBD+)>
<!ELEMENT SBD EMPTY>
<!ATTLIST SBD
    type (%subdomain;) #REQUIRED
    entries CDATA #REQUIRED>

<!-- POS DEFINITION BEGIN -->
<!ELEMENT NOM EMPTY>
<!ATTLIST NOM
    class (%class_noun;) #REQUIRED>
<!ELEMENT ADJ EMPTY>
<!ELEMENT NUM EMPTY>
<!ELEMENT VER EMPTY>
<!ELEMENT AUX EMPTY>
<!ELEMENT PRO EMPTY>
<!ELEMENT ADV EMPTY>
<!ELEMENT CON EMPTY>
<!ELEMENT ADP EMPTY>
<!ATTLIST ADP
    class (%class_adp;) #REQUIRED>
<!ELEMENT INT EMPTY>
<!ELEMENT PAR EMPTY>
<!ATTLIST PAR
    class (%class_sfp;) #REQUIRED>
<!ELEMENT ONO EMPTY>
<!ELEMENT MEW EMPTY>
<!ELEMENT AUW EMPTY>
<!ELEMENT IDI EMPTY>

<!-- POS DEFINITION END -->

```