Language Specific Peculiarities Document for

Cantonese as Spoken in the Guangdong and Guangxi Provinces of China

1. Dialects

The name "Cantonese" is used either for all of the language varieties spoken in specific regions in the Guangdong and Guangxi Provinces of China and Hong Kong (i.e., the Yue dialects of Chinese), or as one particular variety referred to as the "Guangfu group" (Bauer & Benedict 1997). In instances where Cantonese is described as 'Cantonese "proper"' (i.e. used in the narrower sense), it refers to a variety of Cantonese that is spoken in the capital cities Guangzhou and Nanning, as well as in Hong Kong and Macau.

This database includes Cantonese as spoken in the Guangdong and Guangxi Provinces of China only (i.e. not in Hong Kong); five dialect groups have been defined for Cantonese (see the following table)[1]. Three general principles have been used in defining these dialect groupings: (i) phonological variation, (ii) geographical variation, and (iii) lexical variation. With relation to phonological variation, although Cantonese is spoken in all of the regions listed in the table, there are differences in pronunciation. Differences in geographic locations also correlate with variations in lexical choice. Cultural differences are also correlated with linguistic differences, particularly in lexical choices.

Area	Cities (examples)
Central Guangdong Group	Guangzhou, Conghua, Fogang (Shijiao), Longmen, Zengcheng, Huaxian
Northern Guangdong Group	Shaoguan, Qijiang, Lian Xian, Liannan, Yangshan, Yingde, Taiping
Northern Pearl River Delta Group	Nanhai, Gaoya, Sanshui, Shunde, Gaoming, Foshan, Wuchan, Huazhou
Southern Pearl River Delta Group	Shenzhen, Dongguan, Baoan, Zhongshan
Guangxi and Western Guangdong	Guangxi cities, Zhanjiang, Maoming, Lianjaing, Guazhou

[1] Although Chinese dialectologists (such as Yue-Hashimoto (1991) and Bauer and Benedict (1997)) classify the Siyi region, also known as Taishan, as one of the dialect groups for Cantonese, Bauer and Benedict note (1997: xxxvi – xxxvii) that the dialects spoken in this region are markedly different phonologically from Cantonese spoken in the other regions. The extent of these phonological differences has been considered significant enough to exclude the Siyi-Liangyang group of dialects from the current speech collection of Cantonese.

2. Deviation from native-speaker principle

No special deviation: only native speakers of Cantonese born in mainland China will be recruited for voice recording in this project. No speakers from Hong Kong will be included.

3. Special handling of spelling

Spelling is not commonly used for Cantonese words. Chinese characters can be decomposed into radicals, which are sequences of individual strokes.

When spelling foreign words, the Roman letters are pronounced in the English way; Cantonese syllables are used to mimic the English pronunciation (see Spelling alphabet table in Section 9). For Cantonese recordings, only English words and sequences of isolated letters from the English alphabet are used in the prompt files. The elicited spelling item is also replaced by an English word letter sequence.

4. Description of character set used for orthographic transcription

Unicode UTF8 is used for the orthographic transcription of Cantonese.

5. Description of Romanization scheme

For Cantonese, there is no recognized standard comparable to the Pinyin system used to Romanize Mandarin. A number of Romanization systems have been developed to represent Cantonese sounds, the most well known are Meyer-Wempe, Sidney Lau, LSHK (Jyut-ping) and Yale. The Hong Kong and Guangdong/Guangxi Province governments do not follow any particular Romanization system. Matthews and Yip (1994:7) admit that "[r]epresenting Cantonese in alphabetic form is an intrinsically challenging problem, and none of the current systems is ideal."

The Romanization system adopted for transcribing the current Cantonese database is based on the Yale system. This system was developed by Parker Huang and Gerald Kok and has the following advantages:

- its conventions resemble those of the IPA and Pinyin systems (Matthews and Yip 1994);
- most current textbooks and dictionaries adopt the Yale system or some modified version of it;
- it is relatively easy to learn to use and read. This is beneficial not only for those learning Cantonese as a foreign language, but also for native speakers of Cantonese[2], since they tend not to Romanize Chinese in practice; in typing email, for instance, Cantonese speakers either use Chinese characters or type in English.

Cantonese is a tonal language in which most morphemes are one syllable long. As in all languages, each syllable is composed of one initial and one final (or 'rime') and an inherent tone.

The following table provides the Romanization symbol for each of the initials and rimes in Cantonese. The first column provides the phones and the second column provides the (distinctive) phonemes.

Phone (IPA)	Phoneme		Romanization	Example word
			Initials	
p	/b/[1]	1	b	**b**a0 (*father*)
p^h	/p/	2	p	**p**a3 (*fear*)
t	/d/[1]	3	d	**d**a2 (*hit*)

[2] All transcribers are native speakers of Cantonese.

t^h	/t/	4	t	**t**ong4 (*sugar*)
k	/g/[1]	5	g	**g**am0 (*gold*)
k^h	/k/	6	k	**k**am2 (*cover*)
k^w	/gw/[1]	7	gw	**gw**a0 (*melon*)
k^{hw}	/kw/	8	kw	**kw**a0 (*boast*)
m	/m/	9	m	**m**a0 (*mother*)
n	/n/	10	n	**n**ei5 (*you*)
ŋ	/ŋ/	11	ng	**ng**a4 (*tooth*)
f	/f/	12	f	**f**a0 (*flower*)
s	/s/	13	s	**s**a0 (*sand*)
ɕ			s [2]	**s**yu0 (*book*)
h	/h/	14	h	**h**a0 (*shrimp*)
ts	/dz/[1,3]	15	j	**j**aak3 (*narrow*)
tʃ			j	**j**yu0 (*pig*)
ts^h	/ts/[3]	16	ch	**ch**aang2 (*orange*)
$tʃ^h$			ch	**ch**eung4 (*long*)
w	/w/	17	w	**w**aak6 (*draw*)
l	/l/	18	l	**l**aang5 (*cold*)
j	/j/	19	y	**y**ut6 (*month*)
			Rimes	
i:	/i:/ [4]	1	i	j**i**6 (*Chinese character*)
		2	ip	d**i**p6 (*plate*)
		3	im	l**i**m4 (*blinds, curtains*)
		4	it	ch**i**t3 (*cut*)
		5	in	n**i**n4 (*year*)
e^j		6	ik	s**i**k6 (*eat*)
		7	ing	s**i**ng1 (*star*)
ɛ:	/ɛ:/	8	e	s**e**2 (*write*)
		9	ek	s**e**k6 (*rock*)
		10	eng	p**e**ng4 (*cheap*)
a:	/a:/	11	a	f**a**0 (*flower*)
		12	aap	l**aa**p6-s**aa**p6 (*garbage*)
		13	aam	s**aa**m1 (*clothes*)
		14	aat	b**aa**t3 (*eight*)
		15	aan	ng**aa**n5 (*eye*)
		16	aak	j**aa**k3 (*narrow*)
		17	aang	ch**aa**ng0 (*support*)
ɐ	/ɐ/	18	ap	y**a**p6 (*enter*)
		19	am	g**a**m0 (*gold*)
		20	at	gw**a**t1 (*bone*)
		21	an	y**a**n4 (*person*)
		22	ak	h**a**k1 (*black*)

		23	ang	d**a**ng1 (*lamp*)
u:	/u:/ [5]	24	u	g**u**0 (*aunt*)
		25	ut	f**u**t3 (*wide*)
		26	un	w**u**n2 (*bowl*)
o[w]		27	ung	f**u**ng0 (*wind*)
		28	uk	l**u**k6 (*six*)
y:	/y:/	29	yu	s**yu**0 (*book*)
		30	yu	h**yu**t3 (*blood*)
		31	yu	d**yu**n2 (*short*)
ɔ:	/ɔ:/	32	o	ng**o**5 (*I*)
		33	on	g**o**n0 (*dry*)
		34	ong	b**o**ng0 (*help*)
		35	ot	h**o**t3 (*thirsty*)
		36	ok	w**o**k6 (*frying pan*)
œ:	/œ:/ [6]	37	eu	h**eu**0 (*boots*)
		38	eung	g**eu**ng0 (*ginger*)
		39	euk	g**eu**k3 (*foot*)
ɵ		40	eung	ch**eu**n0 (*spring*)
		41	eut	ch**eu**t1 (*go out*)
			Diphthongs	
i:w	/i:w/	42	iu	t**iu**3 (*jump*)
ej	/ej/	43	ei	n**ei**5 (*you*)
ɐw	/ɐw/	44	au	s**au**2 (*hand*)
ɐj	/ɐj/	45	ai	s**ai**3 (*small*)
a:j	/a:j/	46	aai	d**aai**6 (*big*)
a:w	/a:w/	47	aau	m**aau**1 (*cat*)
ɔ:j	/ɔ:j/	48	oi	ch**oi**3 (*vegetable*)
ow	/ow/	49	ou	d**ou**1 (*knife*)
u:j	/u:j/	50	ui	b**ui**1 (*cup*)
ɵy	/ɵy/	51	eui	s**eui**2 (*water*)
			Colloquial and loanwords only	
ɛ:m	/ɛ:m/	52	em[7]	k**em**1 (*game*)
ɛ:p	/ɛ:p/	53	ep[7]	t**ep**1 (*taste, sip*)
ɛw	/ɛw/	54	ew[7]	t**ew**6 (*throw away*)

[1] Non aspiration has been represented as voiced
[2] [s] and [ɕ] : allophones, [ɕ] for GZ speakers
[3] [ts] and [tʃ]: allophones; [ts[h]] and [tʃ[h]]: allophones
[4] [i:] and[e[j]] are treated as allophonic
[5] [u:] and[o[w]] are treated as allophonic (NB no –p or –m)
[6] [œ:] and[ɵ] are treated as allophonic (NB no –p or –m)
[7] No official Yale Romanization for these colloquial rimes.

To mark tones of words, the Yale system uses diacritics. Traditionally, Cantonese has been considered to have nine distinctive tones, with three tones belonging to a category called “entering tone”, “stopped

tones", "dead syllables" or "checked syllables"; these syllables end abruptly with an unreleased *–p, -t* or *–k* stop. However, these three tones are now usually considered to be allotones of the three level tones for the live syllables. Accordingly, the Yale system provides for seven tones: three high, three low and one mid. However, Matthews and Yip (1994) recognise only six tones, arguing that the contrast between the high falling tone and the high level tone is not contrastive (p8). In acknowledgement of the difference between the well-known Yale notation and Matthews and Yip's description, our numbering system provides for the high falling tone (tone 0), but we have normalized virtually all instances of the high falling tone in transcripts to the high level tone (tone 1). (The one exception is the syllable 拜 *baai0*, which we have corrected to the mid level tone *baai3*). Consequently, no syllables in the high falling tone actually appear in the database.

There are also two (sometimes three are recognized) other tones called "changed tones" (Ramsay 1987, Matthews and Yip 1994) or pi:n-jɐm (Bauer and Benedict 1997), which are associated with colloquial Cantonese speech and are generally not used in formal speech contexts. The changed tone of certain words modifies the original meaning either grammatically or lexically, and can mean "that familiar thing one often speaks of" or can indicate the speaker's attitude towards it. While some Romanization systems do have diacritics for these changed tones, the Yale system does not.

The numbering system is as follows:

TONE	Yale diacritic	Numbering
high falling	à	a0
high level	ā	a1
high rising	á	a2
mid level	a	a3
low falling	àh	a4
low rising	áh	a5
low level	ah	a6

Note then that the Romanization system adopted for this database is a modified version of Yale.

6. Description of method for word boundary detection

Word boundaries in Chinese character orthography are not marked in the text. Word boundaries in the Romanized orthography are determined by localization of white spaces (blank, tab, etc.). They have been added manually in accordance with principles developed by Appen. Compounds and reduplicated word groups are treated as one unit; furthermore, affixes and verbal complements are joined to their noun/verb 'stems'. Tense-aspect markers, however, are not considered affixes by Appen and are not joined to their 'stems'. The word boundaries have been verified manually for the phonetically rich material.

7. Table containing all phonemes in the stipulated notation

The phonemic transcription of the words in this database uses the official SAMPA symbol set for Cantonese, which can be found at http://www.phon.ucl.ac.uk/home/sampa/cantonese.htm. The total number of phonemes is 27 (19 consonants and 8 vowels). There are also 11 diphthongs.

Symbol	Word	Transcription	Remark
Consonants			
Plosives			
b p	**b**a0 (*father*) **p**a3 (*fear*)	ba:_0 pa:_3	Non aspiration has been represented as voiced.

d t g k gw kw	**d**a2 (*hit*) **t**ong4 (*sugar*) **g**am0 (*gold*) **k**am2 (*cover*) **gw**a0 (*melon*) **kw**a0 (*boast*)	da:_2 tO:N_4 g6m_0 k6m_2 gwa:_0 kwa:_0	
Fricatives			
f s h	**f**a0 (*flower*) **s**a0 (*sand*) **h**a0 (*shrimp*)	fa:_0 sa:_0 ha:_0	[s] and [ɕ] are allophones. GZ speakers use [ɕ]
Affricates			
dz ts	**j**aak3 (*narrow*) **ch**aang2 (*orange*)	dza:k_3 tsa:N_2	[ts] and [tʃ]: allophones; [ts^{h}] and [tʃ^{h}]: allophones
Nasals			
m n N	**m**a0 (*mother*) **n**ei5 (*you*) **ng**a4 (*tooth*)	ma:_0 nej _5 Na:_4	
Lateral			
l	**l**aang5 (*cold*)	la:N_5	
Semi-vowels			
w j	**w**aak6 (*draw*) **y**ut6 (*month*)	wa:k_6 ju:t_6	
Vowels			
i:	d**i**p6 (*plate*)	di:p_6	[i:] and[e^{ɹ}] allophones of this phoneme
E:	s**e**k6 (*rock*)	sE:k_6	
a:	b**aa**t3 (*eight*)	ba:t_3	
6	g**a**m0 (*gold*)	g6m_0	
u:	w**u**n2 (*bowl*)	wu:n_2	[u:] and[o^{w}] allophones of this phoneme
y:	h**yu**t3 (*blood*)	hy:t_3	
O:	ng**o**5 (*I*)	NO:_5	
9:	g**eu**ng0 (*ginger*)	g9:N_0	[œ:] and [ɵ] allophones of this phoneme
Diphthongs			
iw	t**iu**3 (*jump*)	tiw_3	
ej	n**ei**5 (*you*)	nej_5	
Ew	[no Yale Romanization] (*throw away*)	tEw_6	colloquial only
6w	s**au**2 (*hand*)	s6w_2	
6j	s**ai**0 (*west*)	s6j_0	
a:j	d**aai**6 (*big*)	da:j_6	
a:w	m**aau**1 (*cat*)	ma:w_1	
O:j	ch**oi**3 (*vegetable*)	tsO:j_3	
ow	d**ou**1 (*knife*)	dow_1	
u:j	b**ui**1 (*cup*)	bu:j_1	
9y	s**eui**2 (*water*)	s9y_2	

8. Complete list of all rare phonemes

/kw/ is a relatively rare phoneme in Cantonese.
According to Bauer and Benedict (1997:486-487) some rimes (finals) occur in loan/colloquial words only, but they are combinations of phonemes, rather than individual phonemes.

9. Other language specific items

9.1. Spelling alphabet

A	ei1
B	bi1
C	si1
D	di1
E	yi1
F	e1fu4
G	ji1
H	ek1chu3
I	aai1
J	chei1
K	kei1
L	e1lou4
M	em1
N	en1
O	ou1
P	pi1
Q	kiu1
R	a1-lou4
S	e1-si4
T	ti1
U	yu1
V	wi1
W	dak1-bi1-yu1
X	ek1si1
Y	waai1
Z	yi1-set1

9.2. Chinese written script

There are two versions of Chinese written script: the Simplified-characters version, which is used in Mainland China (including the Guangdong Province) and the Traditional-characters version, which is used in Hong Kong.

Because recordings will take place in mainland China only, the simplified character set is to be used for both display of the prompts and orthographic transcription.

The lexicon for the Cantonese database thus contains three fields: simplified characters, Romanization, and SAMPA representation.

9.3. Spoken and written Cantonese

It is generally agreed (e.g. Bauer and Benedict 1997, Matthews and Yip 1994, Ramsey 1987, Yue-Hashimoto 1991) that Cantonese is different from Mandarin not only in the pronunciation of written words, but also in the extent to which the spoken register diverges from the standard written register. The differences between the spoken and written varieties of Cantonese apply not only to the syntax but also to the lexicon, phonology and graphology. In other words, Cantonese can be regarded as having two "realization systems": one typically used in spoken or informal contexts[3] and the other in written or formal contexts. Note, however, that the former also has a written realization and the latter also has a Cantonese spoken realization! For simplicity's sake, we use the term "spoken form" for the informal variety, and "written form" for the formal variety.

The written form is represented graphically by standard Chinese characters, those which are readable by members of other Chinese dialects, with one caveat: Cantonese speakers in Hong Kong use the traditional characters, whereas those on the mainland use the simplified characters. When read aloud, these characters have Cantonese pronunciation, unintelligible to speakers of other Chinese dialects.

Words in the spoken form may or may not be represented by written characters. Bauer and Benedict (1997:xiv) call such characters "colloquial" or "dialectal" characters. These are often borrowed or adapted homophonous standard Chinese characters. These colloquial characters have not been officially standardized and are not taught to Cantonese speakers at school. Cantonese speakers are exposed to them in popular print media texts, e.g. magazines, certain sections of newspapers, song lyrics, personal letters. The use of colloquial characters in newspapers and magazines has become much more common over the last decade or so, particularly in Hong Kong publications.

There are also a number of commonly used colloquial words that do not have a written form. Since it is not unlikely that these will be used when speaking to a computer, they need to be included in the reading scripts. For such words, the Romanization will be used for identification, and a homophonous character will be used to represent the vernacular morpheme.

9.4. Spelling Chinese characters

When explaining (orally) how to write a new word, Cantonese speakers have a number of options:

(i) Give a homograph of the word:
For instance, this method is used for describing proper names.

(ii) Use a distinguishing radical:
This method is the most common, and is based on the assumption that Cantonese speakers are familiar with the written characters associated with each set of homophones. To use this method, the speaker will refer to the radical that distinguishes the character to be written from its homophones.
Some radicals are themselves characters. For instance, the character for heart "sam1" 心 ('heart') can function as a radical in many other characters, such as the word "seung2" 想 ('to think').
Other radicals may not be words themselves but have names. For instance, one of the radicals in the character河 "ho4" ('river'), is made up of three strokes, and is commonly referred to as

[3] Note that it is not uncommon to find the word "Cantonese" confusingly being used to refer to this spoken, informal variety only.

"saam1 dim2 seui2" 三点水 ('the three water strokes'); it is found in many words associated with water and the sea.
If the speaker does not know the name of a particular radical of a character, s/he may use the specific characteristics of a character to indicate which character s/he is referring to. For instance, there are two different characters for the surname "Wong", a speaker may say "saam1 waak6 Wong" 三画王 (three-strokes Wong) to indicate his/her surname is this character王, and may say "daai6 tou5 Wong"大肚黄 (big-stomach Wong) to refer to the other Wong character, i.e., 黄.

(iii) Describe the component strokes:
NB: This method is used as a last resort. The speaker says the names of individual strokes in the order in which they are written to form the character. For instance, this character 谭 "taam4" can be broken down into three parts: 言 "yin4" ('speech'), 西 "sai0" ('west'), and 早 "jou2" ('morning').

There are six principles which determine the stroke order of Chinese characters: (i) left to right, (ii) top to bottom, (iii) horizontal before vertical, (iv) from outside to inside, (v) middle before two sides, and (vi) inside before closing. A radical can be situated on the top-centre, or left, or right, or at the bottom part of a character. Most (literate) Cantonese speakers will know whether the stroke or radical "spelled" first is situated on the top left-hand-side or the top centre of the word.

11. References

Bauer, Robert S. & Benedict, Paul, K. 1997. *Modern Cantonese Phonology.* Mouton de Gruyter, Berlin/NY

Matthews, Stephen & Yip, Virginia. 1994. *Cantonese: a comprehensive grammar.* Routledge, London/New York

Ramsey, S. Robert. 1987. *The Languages of China.* Princeton University Press, Princeton, New Jersey

Yue-Hashimoto, Anne. 1991. "The Yue dialect." In *Languages and Dialects of China*, [Journal of Chinese Linguistics Monograph Series Number 3], ed. William S-Y Wang. Berkeley, CA, University of California, pp. 294-324.