日中韓辭典研究所
The CJK Dictionary Institute

CJKI 粵語讀音綜合數據庫

The CJKI Comprehensive Cantonese Readings Database

The most complete, accurate, and comprehensive Cantonese database

OVERVIEW

The large scale, linguistically accurate CJK Comprehensive Cantonese Readings Database (CCCRD) is, literally, second to none in both quantity and quality. CCCRD provides Cantonese readings for 300,000 compounds words and some 80,000 readings and romanized variants for some single 13,000 Traditional Chinese characters. It is ideal for such natural language processing applications as speech recognition, speech synthesis, and machine translation for use in input method editors, voice-driven car navigation systems, and voice-to-voice translation/interpretation software.

MAIN FEATURES

- 300,000 compound words and some 80,000 readings and romanized variants for 13,000 Traditional Chinese characters
- Phonemic transcriptions given in standard Jyutping romanization but available in up ten Cantonese romanization systems.
- Phonetic transcriptions provided in accurate IPA or SAMPA.
- Readings ordered by frequency and/or importance.
- Flags distinguish common readings from rare one.
- Compiled, proofread verified by team of Cantonese specialists special "sanity check" programs to detect errors.
- Distributed to software developers and academic organization for reasonable terms.

ACCURACY AND COMPREHENSIVENESS

Our institute has put much effort to ensure that this resource becomes the best Cantonese database available, based on solid principles of Cantonese phonology and semantics. We have examined many sources, including ones produced by the top universities in Hong Kong, and discovered that all the resources have errors and omissions. We have also thoroughly reviewed the database to ensure that it is complete, error-free and does not copy the errors of its predecessors.

COMPILATION METHODOLOGY

First, we have an excellent infrastructure and suite of tools fine-tuned to collect, process and validate Cantonese data. Second, our team includes four native Cantonese linguist and editors, some with a PhD in Chinese linguistics. Third, we have a full range of Cantonese electronic resources and reference works. Finally, to ensure accuracy and completeness, we use meticulous procedures based on our in-depth knowledge of Cantonese and our many decades of experience in compiling the world's largest Chinese dictionary resources.

LINGUISTIC ISSUES

When assigning readings to compound words, we not only take into account the phenomenon of polyphony (多音字), as in *cung4* and *zung6* for 重, but also consider *tone change* (変音). Unlike tone sandhi (変調), tone change is *lexically dependent* so that it is unpredictable and requires manual proofreading. Tone change can be of two types: tone assimilation and morphological tone change.

An example of tone assimilation is:

挨晚	*aai1 maan3* →	*aai1 maan1* 'evening'	(assimilation obligatory)
今晚	*gam1 maan3* →	*gam1 maan1* 'tonight'	(assimilation optional)

An example of morphological tone change is:

袋	*doi6* 'to bag sth' →	*doi2* 'bag' (part of speech change)
靚仔	*leng3 zai2* 'handsome boy' →	*leng1 zai2* 'young boy' (semantic change)

For single character readings, we not only need to ensure that the first reading is the most common one, as in:

行 *hang4 haang4 hong4 hang6 hong2 hong6*

but also that the remainder of the readings are arranged in order of frequency or importance, whenever applicable (some readings are of equal frequency). In addition we have a flag that shows which readings are rare (see samples).

- - -

In this manner, we have made significant efforts to ensure that the database is complete, accurate, and comprehensive, and fully ready for use in a host of NLP applications, especially Cantonese speech technology.

Table 1

HANZI	UNICODE	RARE	JYUTPING	GONGSIK	CPR
行	884C	N	hang4	hang4	hang4, hung4, heng4
		N	haang4	haang4	hang4
		N	hong4	hong4	hong4
		N	hang6	hang6	hang6, hung6, heng6
		N	hong2	hong2	hong2
		Y	hong6	hong6	hong6
使	4F7F	N	si2	si2	si2, see2, shi2, shee2, sze2
		N	sai2	sai2	sai2, say2, shai2, shay2
		N	si3	si3	si3, see3, shi3, shee3, sze3
重	91CD	N	cung4	chung4	tsung4, tsoong4, chung4, choong4
		N	cung5	chung5	tsung5, tsoong5, chung5, choong5
		N	zung6	jung6	tsung6, tsoong6, dzung6, dzoong6, chung6, choong6

The above table shows single characters polyphones (多音字), characters that have more than one reading) that whose reading depends on the context and meaning. As can seen the readings are available in standard Jyutping as well as in other romanization systems, such as Gonksik and CPR (Cantonese Popular Readings).

Table 2

WORD	JYUPTING	DEFAULT
重量	cung5 loeng6	cung4 loeng6
重複	cung4 fuk1	cung4 fuk1
重要	zung6 jiu3	cung4 jiu3
欺騙行爲	hei1 pin3 hang4 wai4	hei1 pin3 hang4 wai4
行家瞧門道	hong4 gaa1 ciu4 mun4 dou6	hang4 gaa1 ciu4 mun4 dou6
行大運	hang4 daai6 wan6	hang4 daai6 wan6
照功行賞	ziu3 gung1 hang4 soeng2	ziu3 gung1 hang4 soeng2
馴行	seon4 hang4	seon4 hang4
遊覽飛行	jau4 laam5 fei1 hang4	jau4 laam5 fei1 hang4
滔天罪行	tou1 tin1 zeoi6 hang6	tou1 tin1 zeoi6 hang4

The compounds shown here include characters having several readings (polyphones, 多音字) corresponding to different meanings which may be unrelated. The Default column shows the primary reading of each character, which is not always correct, whereas the Jyutping column shows the *actual* reading for that that compound. Note how the actual reading can depend on the meaning.

Table 3

WORD	JYUTPING	DEFAULT	COMMENT
說話	syut3 waa6	syut3 waa6	Includes a character having several readings corresponding to different meanings which are related.
廣東話	gwong2 dung1 waa2	gwong2 dung1 waa6	
刷牙	caat3 ngaa4	caat3 ngaa4	
牙刷	ngaa4 caat2	ngaa4 caat3	
坐低	co5 dai1	zo6 dai1	Includes a character having several readings that don't differ in meaning but one being more formal.
坐井觀天	zo6 zeng2 gun1 tin1	zo6 zeng2 gun1 tin1	
人力車	jan4 lik6 ce1	jan4 lik6 ce1	Includes a character having different readings that depend on the type of compounds it is.
女人	neoi5 jan2	neoi5 jan4	
老婆	lou5 po4	lou5 po4	
婆婆	po4 po2	po4 po4	

Some characters undergo tone change (變音), different from tone sandhi (變調), which is lexically dependent and unpredictable.