Words.hk: a Comprehensive Cantonese Dictionary Dataset with Definitions, Translations and Transliterated Examples

Chaak Ming Lau*, Grace Wing-yan Chan*,
Raymond Ka-wai Tse†, Lilian Suet-ying Chan‡

The First Workshop on Dataset Creation for Lower-Resourced Languages @LREC2022 (24 Jun 2022)

*The Education University of Hong Kong
†The Hong Kong University of Science and Technology
‡Words.hk

Outline

- We built a dictionary for a language that is *widely spoken but usually unwritten* . We compiled more than *50,000 entries, with definitions and examples* , with the help of *devoted volunteers and the community* .

1. Introduction
2. Compilation
3. Linguistic considerations
4. Dataset
5. Conclusion

Introduction

§1 Cantonese

Cantonese (ISO-639-3: yue)

- 68M speakers
- Widely spoken in Hong Kong, Macau, Guangdong, Guangxi (Southeastern China and part of S.E. Asia)
- Both Traditional Han and Simplified Han scripts are used
- *Diglossia*
  - Written Cantonese is discouraged in the education system
  - Most writing is done in Standard Written Chinese
- Tonnes of spoken data (radio, movies, online videos), but close to no usable written data

§1.1 Existing Resources

Name	Type	Size	License
Cifu (Lai and Winterstein, 2020)	Lexicon	51,798 words	GPLv3
Rime-Cantonese	Lexicon (Input Method)	185,809 items	CC-BY-4.0
CyberCan (Shen et al., 2021)	Lexicon	133,212 words	CC-BY-4.0
Cantonese WordNet (Sio and da Costa, 2019)	Wordnet	3,500 concepts, 12,000 senses	CC-BY-4.0
HKCanCor (Luke and Wong, 2015)	Corpus	230,000 words	CC-BY-4.0
CantoMap (Winterstein et al., 2020)	Corpus	105,000 words	GPLv3
HKCAC (Leung and Law, 2001)	Corpus	170,000 words	Proprietary
ABC Cantonese-English Comprehensive Dictionary (Bauer, 2020)	Dictionary	15,000 entries	Proprietary
CC-Canto	Dictionary	34,335 words	CC-BY-SA-3.0
CantoDict	Dictionary	60,714 words	Proprietary

Table 1: Selected Cantonese language resources

§1.2 The Project

- A resource developed by the Cantonese dictionary project 粵典 (*words.hk*, *cantowords.com*), founded in 2014
- Bilingual (Canto-Canto, Canto-Eng)
- 53,000 dictionary entries (of which more than 11,550 thoroughly reviewed and published)
- Other entries are available for research purposes

Compilation

§2.2 Crowd-editing

- *Where?* A wiki-like online system (since 2014)
  - People who have made at least one edit: 300+
  - Cumulative number of revisions: 150,000+
- *Who?* Online volunteers and Interns
- *Training?* Regular meetups (pre-COVID) and Apprenticeship
- *Division of Labour*
  - Editors can make edits to parts that they are confident with
  - Specialisation is needed: English translation, romanisation, etc.

§2.3 Web-scraping

- *Problems:*
  - resources are locked up
  - segmentation
  - "Diglossia" (Newspapers are 99% Standard Written Chinese, online forums contain too many typos)
- *What we did:* Bigrams and Trigrams extracted from online forums for manual filtering
  - Editors must sift through false positives of word combinations (e.g. 屋企喺 *uk1kei2 hai2*, "home is at"), typographic errors and proper names (celebrities or user handles)

§2.4 Quality Control

- All entries are “unpublished” by default
- Editors can review or edit the entries, and leave comments for discussion (random assignment)
- Reviewed entries are published by senior editors or Chief Editor

Linguistic considerations

§3 Segmentation

Cantonese orthography does not use the space, and it is well known that inter-annotator consistency for Chinese languages is low for this task (Sproat et al. 1996)

- *Our strategy:*
  Lexicographic functions (Bergenholtz and Tarp, 2003).
  e.g. Duplicates are acceptable, if keeping them is beneficial to the users.

§3 Segmentation

- 女朋友 (*neoi5pang4jau5*, “girlfriend”) can be listed as one entry, and the constituents 女 (*neoi5*, “female”) and 朋友 (*pang4jau5*, “friend”) will be listed as separate entries. The second part is further divided into 朋 and 友, and recorded as separate morphemes in the dataset.
- This creates **redundancy**, but it is fine as it does not harm the project.
- *Drawback:* Number of entries will not reflect actual number of word ”types” in Cantonese.

§4 Orthographic Representation

These issues have been handled:

- Character Choice (Hant is used, Eng or Jyutping if necessary, Hans will not be included)
- Misspelling
- Character Variants, mainly discrepancies between Taiwan and Hong Kong (both use Hant), and difference between Big5 glyphs and actual usage. (Note: *OpenCC* is used for the conversion)

§5 Pronunciation

The pronunciations are written in LSHK Jyutping.

- We are descriptive, in the sense that we describe what the majority feels to be correct.
- Accept newer rhymes in loanwords (e.g. oem)
- Actual pronunciations are not necessarily listed
  - Context: Mergers of onsets (e.g. n- and l-) and codas (e.g. -n and -ng, -t and -k) are common, e.g. 你 (“you”) is often pronounced as lei5, but we list nei5.
  - *traditional* pronunciations will be listed due to lexicographic functions.

§6 Part-of-speech Tagging

- Follows the POS system in Tang (2015)
- Can be roughly mapped to the Universal Dependencies (UD) Cantonese-HK tag-set (Wong et al., 2017)

POS	English Translation	UD
名詞	nouns	NOUN
區別詞	distinguishing words	ADJ
數詞	numerals	NUM
量詞	quantifiers	NOUN
代詞	pronouns	PRON, DET
動詞	verbs	VERB
形容詞	adjectives	ADJ
副詞	adverbs	ADV
介詞	prepositions	ADP
連詞	conjunctions	CCONJ, SCONJ
助詞	particles	PART
擬聲詞	onomatopoeia	INTJ
感嘆詞	interjection	INTJ
詞綴	affixes	PART, AUX
語素	morpheme	N/A
語句	expressions	N/A

Table 3: Part-of-speech tags and their corresponding POS tags in UD-Cantonese

§7 Definition Crafting

Guiding questions for less-experienced volunteers

- Is this a common, mid-range or rare word, in terms of perceived frequency in speech?
  - For a *common* word, list out different senses of the word with ample collocations and examples.
  - For a *mid-range* word, explain the word in plain language, and give one or two example sentences.
  - For a *rare* word, explain the word in a way that can describe its precise sense without using any other rare words.
- If it is an abstract concept, how would you explain it to a five year-old child?
- Is your definition too broad or restrictive for the word?

Dataset

§8 Data Format

Col1	Index	76359
Col2	Orthographic representation & Jyutping	一般來説:jat1 bun1 loi4 syut3
Col3	Entry-data (POS, Label, Synonyms, Antonyms, Explanation, and Examples)	(pos: 語句)(label: 書面語)(sim: 一般而言) <explanation> yue: 用嚟引起下文，表示只係睇普遍情況，唔考慮個別例子 eng:in general, in most situations <eg> zho: 一般來説，男生都喜歡漂亮的女孩子。(jat1 bun1 loi4 syut3, naam4 sang1 dou1 hei2 fun1 piu3 loeng6 dik1 neoi5 haai4 zi2.) yue: 一般嚟講，男仔都鍾意靚嘅女仔。(jat1 bun1 lai4 gong2, naam4 zai2 dou1 zung1 ji3 leng3 ge3 neoi5 zai2.) eng:In general, boys like beautiful girls.
Col4	Character variations	一般來說
Col5	Review status	OK

§9 Licensing

- Dataset and supporting files: *https://github.com/wordshk/data2021*
- The snapshot provided is CC-BY-NC 4.0
- The core dataset has a proprietary licence.

Conclusion

§10 Conclusion and Discussion

- The process of the compilation of this dictionary.
- *Use cases*
  - simplistic (longest string matching) word segmentation
  - training of text-to-speech models with verified pronunciation mapping data
  - Written Cantonese resources
  - WordNet or knowledge base projects

  We also plan to migrate this to a simpler format in the future (e.g. TEI Lex-0).

§A.2 Broader Impact

Through this project, we hope that we can achieve these:

- encourage more people to write in their own native language
  - The *words.hk* is one of the projects that proved the viability of using Cantonese in educational settings and paved the way for widespread use of written Cantonese.
  - As of today, Cantonese is commonly used in a variety of ways, ranging from literature to government publications.
- invite non-expert community members to contribute to a monolingual dictionary
  - editors need not be proficient in every aspect
  - open up projects to community members (who may not be directly involved in the process, including those who have the technical skills to work with the online system) for their views and comments