Enriching Linguistic Representation in the Cantonese Wordnet
And
Building the New Cantonese Wordnet Corpus

香港粵網

Joanna Ut-Seong Sio and Luis Morgado da Costa
Palacký University Olomouc
LREC 2022, 20-25 June, Marseille, France

Marie Skłodowska-Curie Actions

Univerzita Palackého v Olomouci

The Cantonese Wordnet

- Cantonese is a widely known Chinese regional variant, close to 80 million speakers worldwide
- We focus on Hong Kong Cantonese (about 7 m. population)
- Large yet under-resourced language
- Project started 2019, and is ongoing
- We include both characters and romanization (Jyutping, including tonal information, 花, faa1, ‘flower’)
- We try to capture as much variation as possible

New: Increasing sense coverage

- The current version of the Cantonese Wordnet contains over 16,000 senses (one sense = character + jyutping, over 32800 forms) distributed over more than 5,000 concepts (12,092 senses, distributed across 3,533 concepts in the previous version)

POS	No. synsets	%	No. senses	%
nouns	2,776	(52.9%)	7,067	(43.3%)
verbs	1,360	(25.9%)	4,200	(25.7%)
adjective	801	(15.3%)	4,071	(24.9%)
adverb	218	(4.2%)	896	(5.5%)
non-referential	97	(1.8%)	102	(0.6%)
Total	5,252	-	16,336	-

What is new?

- This new version of the Cantonese Wordnet includes functional categories that do not exist in the Princeton WordNet:
  - sortal classifiers
  - measure words
  - post-verbal particles
- And also a new open corpus of example sentences
  - enrich verbal representation

Classifiers

- **Sortal classifiers** (doesn't exist in English) and **measure words** (some exist in English)
  - Sortal classifier: one unit book/cat,

    一本書 (jat1 bun2 syu1) **vs.** 一隻貓 (jat1 zek3 maau1)
  - Measure word: one cup tea 一 杯 茶 (jat1 bui1 caa4)
- Sortal classifiers and measure words differ in syntactic behaviors
- We largely follow the Chinese Open Wordnet (COW) in the treatment of classifiers

Sortal Classifiers

- A direct mapping of Cantonese sortal classifiers to their Mandarin counterparts is not possible because:
  - many classifiers are unique to either variant;
  - identical classifiers in both variants are used with different nouns.
    - The Mandarin classifier *zhī* (隻) is used with small animals
    - Its Cantonese counterpart *zek3* (隻) covers both large and small animals (Erbaugh, 2013)

Measure Words

- Container measure words are both nouns and classifiers
- Many container words in Cantonese have equivalents in PWN (*cup* and *cupful/cup*), but some concepts are missing in the PWN: e.g., *Wok* but no *wok-ful* while *wok6* 鑊‘wok’ is a measure word in Cantonese;
- The opposite problem also exists: *roomful* exists in PWN but *fong2* 房 ‘room’ is not a very natural measure word in Cantonese
- A full overlap PWN and the Cantonese Wordnet is not possible
- But we use the relation 'eq_synonym’ to link the new Cantonese concepts to the existing English ones whenever possible

New Classifiers in the Cantonese Wordnet

- 41 sortal classifiers; 25 container measure words
- Receive the part-of-speech '**x**' (used for non-referential concepts)
- Standardized definitions encapsulating their type and meaning
- Added a small nominal hierarchy for the different types of classifiers
- Use concept relation '**exemplifies**' (used to indicate the usage of this word) to link the newly added classifiers to their respective types

Romanization: *gaa3*
Character: 架
Definition: a sortal classifier used for wheeled vehicles such as a car, a motorcycle or a wheelchair

Post-verbal particles

- Cantonese has a very rich inventory of post-verbal particles. They appear between the verb and the object (when present): V-P-O
- Post-verbal particles are needed in order to accomplish accurate cross-lingual linking of concepts, e.g., the equivalent of the English achievement verb ‘to find’ is *wan2 dou2* 搵到,
  - 搵 *wan2* means to look for / to search;
  - 到 *dou2* is a post-verbal particle, denoting that the action has been brought to a successful end

Post-verbal particles

- We have added one nominal concept that introduces ‘post-verbal particles’, and four other nominal concepts to introduce the four sub-types (Matthews and Yip, 2013).
- 32 post-verbal particles were added as non-referential concepts (taking ‘x’ as part-of-speech).
- They have a definition that alludes both to their category and their meaning. E.g.,:
    - 咗 *zo2,* defined as ‘a post-verbal aspectual particle used to express the perfective aspect’

Enriching verbal representations

- Cantonese has many compound verbs (e.g., *paau2 bou6* 跑步 'to run', lit. 'run a step')
- These compounds are puzzling in that even though they all correspond to single English lemmas, some of them are separable, allowing various elements to be inserted in-between the characters/morphemes
- Separability is not totally predictable
- Information on separability is important for the Cantonese Wordnet to be used as a linguistic resource

Enriching verbal representations

- We started with the perfective aspectual particle *zo2* (one of the most common aspectual particles)
- For each verb sense, we checked if it was compatible with *zo2*, and if so, we indicated where *zo2* was placed (in-between the compound or after)
- We also added information on transitivity, indicating whether each verbal sense accepts a complement (regardless of its type: e.g., nominal or clausal)

Enriching verbal representations

Summary of results (verbs):

- 82.4% are transitive
- 11.5% do not take *zo2* at all.
- Among compound verbs that take *zo2*, 79.5%, take *zo2* only at the end.
- About 8.9% of verbs are separable (i.e., take *zo2* in the middle)
- Only 0.3% (only 12 verbs) seem to be able to take *zo2* both in between and after the verb

Reasons for incompatibility: aspectual properties of the compound verbs, registers, etc.

Cantonese WordNet Corpus

- **Cantonese Wordnet Corpus**: a corpus of handcrafted examples where individual senses are shown in context

Synset: 00005815-v
Character: 咳

English lemma: cough
Romanization: *kat1*

Example sentence: 一起身咳咗好耐, 所以打電話返公司請假。

- **3,570** hand-crafted example sentences by two native speakers; begin with sentences with *zo2*

Cantonese Wordnet Corpus, why?

- We need corpora to do some types of linguistic research
- Existing Cantonese corpora are mostly sourced from speech/spoken data ↠ natural (include filler and pauses) but not ideal to extract clean examples
- Some are made up of texts collected from the Internet, using Cantonese seed words for crawling texts ↠ the text is often a mix of Cantonese and Mandarin
- Not all are under open license

Future

- Continue to extend coverage
- Adding concepts present in Cantonese but not in English by using the Collaborative Interlingual Index (Bond et al., 2016)
- Build a sense-annotated corpus from the example sentences
- Working towards making the Cantonese Wordnet a resource useful in linguistic research/education by adding, more grammatical features, pronunciation recordings and possible integration with online learning applications

Release

- The Cantonese Wordnet is released under a Creative Commons Attribution 4.0 International License (CC BY 4.0). This new version of the Cantonese Wordnet will be available on its Github repository.

  https://github.com/lmorgadodacosta/CantoneseWN

Contacts:

Joanna Ut-Seong Sio joannautseong.sio@upol.cz

Luis Morgado da Costa lmorgado.dacosta@gmail.com

Acknowledgement

We would like to thank **Dennis Lam Ngai** Wa for helping us to check data and provide example sentences. The research described here is supported by the European Regional Development Fund - Project '**Sinophone Borderlands – Interaction at the Edges**' CZ.02.1.01/0.0/0.0/16_019/0000791 and by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement H2020-MSCA-IF-2020 CHILL –No.101028782.