Building the Cantonese Wordnet

Joanna Ut-Seong Sio
Palacky University
Olomouc, the Czech Republic
joannautseong.sio@upol.cz

Luis Morgado da Costa
Nanyang Technological University
Singapore
luis.passos.morgado@gmail.com

Abstract

This paper reports on the development of the Cantonese Wordnet, a new wordnet project based on Hong Kong Cantonese. It is built using the expansion approach, leveraging on the existing Chinese Open Wordnet, and the Princeton Wordnet's semantic hierarchy. The main goal of our project was to produce a high quality, human-curated resource – and this paper reports on the initial efforts and steady progress of our building method. It is our belief that the lexical data made available by this wordnet, including Jyutping romanization, will be useful for a variety of future uses, including many language processing tasks and linguistic research on Cantonese and its interactions with other Chinese dialects.

1 Introduction

1.1 Chinese and its Dialects

Chinese is generally treated as one language with many dialects for both cultural and political reasons. The dialects are spoken by people who mostly identify as a single nationality with a shared cultural history. Linguistically speaking, this unifying view is problematic as the dialects are not always mutually intelligible. Chinese is, more accurately, a family of genetically-related languages most probably descended from a form of late Old Chinese dating from the Han Dynasty or slightly earlier (with the possible exception of Min (Handel, 2015)). Various dialects, including Cantonese, had also been importing grammatical elements from neighboring languages (Yue-Hashimoto, 1991), creating dialectal variations that are more than the sum of language-internal changes. An arguably less confusing term 'Sinitic' is often used to refer to the Chinese languages (Handel, 2015). The term 'topolects' is coined by Mair (1991) to refer to Chinese dialects or, more generally, to speech varieties where the label of either 'language' or 'dialect' would be controversial. Nevertheless, for the purpose of this paper, we will continue to use the term 'Chinese' to refer to this family of languages and the term 'dialects' to refer to its variants while being fully aware of the complexity involved.

There are seven most recognised dialectal groups: Mandarin (or Northern Chinese), Xiang, Gan, Wu, Yue, Hakka and Min (Handel, 2015). Norman (1988) classifies the traditional seven dialectal groups into three larger groups: Northern (Mandarin), Central (Wu, Gan, and Xiang) and Southern (Hakka, Yue, and Min). Cantonese belongs to Yue, the Southern group, and it is often used as an alternative name for this whole group. The variety this Cantonese Wordnet is based on is Hong Kong Cantonese. Hong Kong Cantonese is often considered a prestige variety due to its association with the prosperous southern provinces as well as with the Cantonese culture of films and popular music.

1.2 Project Motivation

A few wordnets exist for Chinese languages. These efforts include some work on Pre-Qin Ancient Chinese (Zhang et al., 2017), Middle Ancient Chinese (Zhang et al., 2014), as well as multiple wordnets for Mandarin Chinese, namely: the Sinica Bilingual Ontological Wordnet (Huang, 2003; Huang et al., 2004, BOW), the Southeast University Chinese WordNet (Xu et al., 2008, SEW), the Chinese WordNet (Huang et al., 2010, CWN) and the Chinese Open Wordnet (Wang and Bond, 2013, COW). The Chinese Open Wordnet is the best and most recent effort to produce a high quality wordnet for Mandarin Chinese, learning from previous analogous experiences and developed alongside a sense-tagged corpus (Tan and Bond, 2014; Wang and Bond, 2014; Seah and Bond, 2014).

Unfortunately, scholarly efforts often seem to forgo Cantonese, as there is a chronic absence of digital resources to study and process this important Chinese dialect. To further stress this problem, it is important to note that the significant differences between Mandarin, for which there are plenty of resources, and Cantonese make the idea of using Mandarin resources to process Cantonese fairly useless.

It was this chronic absence of Cantonese digital resources that ultimately fed our motivations to build the Cantonese Wordnet. Our motivation spawns from our belief that Cantonese should have plenty of open, computational tractable and linguistically rich resources, such as wordnets and corpora, that support scholarly work, as well as this language's maintenance and preservation – similar to what happens with Mandarin Chinese.

We would like our Cantonese Wordnet to support many Natural Language Processing tasks, such as speech recognition, word sense disambiguation, machine translation or information retrieval. And, at the same time, to also support the study of purely linguistic research topics, such as lexical semantics, tonal patterns, verb subcategorization, etc.

2 Cantonese: an Overview

Cantonese is the second most widely known Chinese dialect after Mandarin (Matthews and Yip, 1994). It is spoken in Guangdong Province, Guangxi Province, the Special Administrative Regions of Hong Kong and Macau, as well as diaspora communities in North America, Australia, Malaysia, Singapore, etc. According to Ethnologue,[1] there are 73 million Cantonese speakers worldwide. But despite the large number of speakers, credible online resources on Cantonese, free or otherwise, are limited, especially in comparison with Mandarin.

There is a considerable lexical overlap between Cantonese and Mandarin. Snow (2004, 49) mentions that the difference between Cantonese and Mandarin vocabulary ranges from 30-50%. Ouyang (1993, 23) estimates that about 1/3 of the lexical items used in regular Cantonese speech is not found in Mandarin. To give an example for a very common item,'umbrella' is yǔsǎn 雨伞 in Mandarin but ze1 遮 in Cantonese. In cases where they share the same lexical item, the item is always pronounced differently in the two dialects. For example, 'teacher' 老師 is pronounced as lǎoshī in Mandarin and lou5si1 in Cantonese. The vowels of the first syllables in each case are different, and the onsets of the second syllables are also different, not to mention tonal differences. Note that the romanization system is different here as well. Mandarin uses pīnyīn, which is based exclusively upon the pronunciation of the Beijing dialect. In Cantonese, Jyutping is used, a point we will come back to later.

In the existing Mandarin Chinese Wordnet, simplified characters are used. The simplified script, adopted in 1949, aims to alleviate some of the difficulty associated with use of the traditional script, as a measure to eradicate illiteracy. In Hong Kong, Macau and Taiwan, the traditional script is used, though in the former two, changes are happening rapidly since their return to China in 1997 and 1999, respectively.

Cantonese is primarily a spoken variant. A lot of lexical items, excluding those shared with Mandarin, do not have fixed agreed upon characters, these are often called 'characterless' words. It is not always easy to determine which character to use as there is no standardization. In some cases, multiple options are available, while in some other cases, no options are available. We will pick up on this issue in Section 4.3.

2.1 Jyutping Romanization System

Pīnyīn is the official romanization system of Mandarin Chinese or Pǔtōnghuà (lit. 'common speech'). And since Mandarin Chinese/Pǔtōnghuà and Cantonese have different phonological systems, a different romanization system is needed for Cantonese. Many romanization systems exists for Cantonese (e.g., Jyutping, S.L. Wong, Sidney Lau, Yale, the Government System, etc.) (Cheng and Tang, 2016). We adopt the Jyutping system, 粵拼, for the Cantonese Wordnet. Jyutping was

[1] https://www.ethnologue.com/language/yue

developed by the Linguistic Society of Hong Kong (LSHK) in 1993. Its formal name is The Linguistic Society of Hong Kong Cantonese Romanization Scheme.[2] Since its inception, it is used widely in academic papers as well as social media.

Cantonese syllables contain onset and rime. The rime can be further divided into the nucleus and coda. The lists of possible onset, nucleus and coda in Jyutping are shown in Table 1. /m/ and /ng/ are syllabic nasals, meaning they can appear on their own to form a syllable. Kataoka and Lee (2008) provide the correspondence between Jyutping, the International Phonetic symbol (IPA) and other Cantonese romanization systems.

Jyutping	phonemes
Onset	b, p, m, f, d, t, n, l, g, k, ng, h, gw, kw, w, z, c, s, j
Nucleus	aa, i, u, e, o, yu,oe, a, eo
Coda	p, t, k, m, n, ng, i, u

Table 1: Jyutping Syllable Struture

In Jyutping, tones are expressed numerically, using numbers 1 to 6. Table 2 shows how these numbers relate to their respective aptonal contour using Chao's number (1 is the lowest and 5 is the highest) together with their description.

Jyutping	Chao's	description
1	53/55	high falling/high level
2	35	mid rising
3	33	mid level
4	21	low falling
5	13	low rising
6	22	low level

Table 2: Cantonese Tones

Traditional Chinese philology treats syllables with final stops (p, t, k) as distinct tone classes (checked tones), yielding a nine-tone system. Until recently, there was also a contrast between high level (55) and high falling (53). However, this distinction has collapsed for most speakers today.

Cantonese has a lot of homophones, characters that have the same pronunciation but have different meanings. To uniquely identify a lemma, both its Jyutping representation and its graph (character) are needed. For example, sing1 can mean 'to rise' 升 or 'star' 星. Without the character, it is ambiguous.

3 Methodology

There are two main methods to build wordnets (Vossen, 1998). The first method is known as the 'expansion' approach, where the structure of another wordnet is used as 'pivot', and the main work is essentially a translation effort – conserving the structure of the pivot wordnet and translating nodes of the hierarchy. The Princeton Wordnet (Fellbaum, 1998, PWN) is, by far, the most frequently used 'pivot' for projects that employ this approach. The second method is known as the 'merge' approach. This is usually a slower method, since no pivot structure is assumed, but it ensures a higher degree of freedom to more carefully model the structure of the wordnet based on the language in question, without depending on pre-assumed semantic relations. One of the immediate benefits of this approach is the ability to add new concepts that are not part of the 'pivot' language, a problem many wordnet projects that followed the 'expansion' approach have struggled with.

And while the 'merge' approach is perhaps more principled in theory, the major drawback from this approach is that it does not benefit from the parallel translations available from all other projects that used the same pivot. The best example of this benefit is the Open Multilingual Wordnet (Bond and Foster, 2013, OMW), a project that links dozens of open wordnets using PWN as the common structure. This language alignment is very useful for many NLP tasks, such as Machine Translation and Word Sense Disambiguation.

A recent addition to this discussion is the conception of the Collaborative Interlingual Index (Bond et al., 2016, CILI) – an open, language agnostic, flat-structured index that links wordnets across languages without imposing the hierarchy of any single wordnet. And even though PWN was the main contributor to the initial set of concepts present in CILI, this set is no longer constrained by it – multiple projects are now able to contribute to CILI's set of concepts, and gain the ben-

[2] https://www.lshk.org/jyutping

efits of multilingual alignments without the penalty of being frozen within some imposed structure. To the best of our knowledge, the quickest and easiest way to link to CILI and to access these language alignments without an imposed structure is, interestingly enough, to use the expansion approach with PWN hierarchy has pivot because all PWN concepts have direct links to CILI. And this was what we decided to do.

As we had no urgent need for a high coverage wordnet, for which multiple bootstrapping techniques are available to quickly create high coverage lower quality resources, we decided to build a high quality resource fully checked by native speakers. And although we knew from the start that building a wordnet from scratch would be very time-consuming, without going against our commitment to high quality, we decided to ease our task by leveraging on existing resources as much has possible.

We used the Chinese Open Wordnet (Wang and Bond, 2013, COW) as pivot. COW is a high quality hand-checked resource for Mandarin Chinese that was also created through the expansion approach (using PWN as pivot). This means that by linking our wordnet to COW, we would have easy access to PWN's concept IDs and, as a result, also to CILI.

The basic assumption of our method was that while Mandarin Chinese and Cantonese are fairly different languages, and it was clear from the start that resources from one language would not perform well in tasks for the other language, there is still a fair amount of overlap in the lexical usage. This is not without caveats, since Cantonese uses traditional characters and Mandarin Chinese uses simplified characters – and conversion from simplified to traditional characters is inherently lossy. That being said, we decided to automatically convert the lemmas in simplified Mandarin Chinese to traditional Chinese, and use this to jumpstart the manual construction of our Cantonese Wordnet.

Since the other Mandarin Chinese wordnets such as the Sinica Bilingual Ontological Wordnet (Huang, 2003; Huang et al., 2004, BOW), the Southeast University Chinese Wordnet (Xu et al., 2008, SEW) and the Chinese Wordnet (Huang et al., 2010, CWN) included lemmas that were not present in COW, we started by building a small 'Chinese Wordnet' from the union of all four Chinese wordnets: COW, BOW, SEW and CWN. This was fairly easy since all these wordnets were linked to PWN's hierarchy. Next, we used Hanziconv[3] to convert all lemmas from simplified Mandarin Chinese to traditional characters. And finally, we generated a list of all candidate senses (with lemmas converted into traditional characters) that satisfied any of these three criteria:

- Senses that belong to the 4,960 'core' concepts in Princeton WordNet (Boyd-Graber et al., 2006) – a usual measure for coverage of wordnet resources;
- Senses from all concepts in two sense tagged Sherlock Holmes stories, as reported by NTUMC (Tan and Bond, 2014); and
- Senses from any concept with sense sum-frequency score of one or higher, as reported by the PWN (i.e. most concepts yield sum-score of 0);

3.1 Human Validation and Jyutping

The data generated by the process explained above generated a list of 47,499 candidate senses, spanning over 9,340 synsets. Based on this information, we created a spreadsheet for our human validation task. As of this moment, a single Cantonese native speaker, who is also a trained linguist with extensive work on Cantonese language is manually checking, correcting and adding to this data.

An example of this spreadsheet is shown in Table 3. This spreadsheet contains the candidate Cantonese lemmas (converted to traditional characters from one of the existing Mandarin lemmas), English lemmas (provided by the PWN), Mandarin lemmas (provided by the collection of Chinese wordnets), English definitions and examples (provided by the PWN), and the synset ID of the PWN3.0.

The Jyutping romanization is not produced automatically. It is, in fact, being added by hand by the lexicographer. To our knowledge, there are no open Jyutping dictionaries available under an open license. For this reason, we decided to include this valuable resource in our wordnet. Having Jyutping romanization

[3] https://pypi.org/project/hanziconv/

Cantonese Lemma	English Lemmas	Mandarin Lemmas	English Definitions	English Examples	Synset
今夜 [deleted]	[deleted]	[deleted]	[deleted]	[deleted]	[deleted]
今晚, gam1 maan5	this night; tonight; this evening	今夜; 今晚	during the night of the present day	drop by tonight	00079499-r
今晚, gam1 maan1	this night; tonight; this evening	今夜; 今晚	during the night of the present day	drop by tonight	00079499-r
今晚黑, gam1 maan5 hak1	this night; tonight; this evening	今夜; 今晚	during the night of the present day	drop by tonight	00079499-r
今晚黑, gam1 maan1 hak1	this night; tonight; this evening	今夜; 今晚	during the night of the present day	drop by tonight	00079499-r

Table 3: Human Validation and Jyutping (example)

for each sense will not only facilitate searching, but can also be very useful for a variety of other tasks, such as speech recognition or even for educational purposes. We will make use of the new structure provided by the WN-LMF format to cluster Jyutping romanizations as variants inside the canonical lemma (i.e. the traditional Chinese characters).

The process explained in the section above generated one candidate Cantonese sense for each available Mandarin sense inside each concept. In the example shown in Table 3, the concept 00079499-r contained two Mandarin senses: jīn yè 今夜, and jīn wǎn 今晚. Both these lemmas have the same form in simplified and traditional Chinese. This resulted in two lines produced (the top two lines in Table 3). The human validation task comprised:

1. asserting if the candidate sense in each line provided was a correct Cantonese sense – incorrect senses would be deleted (see Table 3, line 1);
2. adding Jyutping romanization for each correct sense – senses with more than one pronunciation required the line to be copied and the corresponding romanization added to the new line (see Table 3, lines 2-3);
3. adding any missing senses that were not suggested by the conversion of the Mandarin lemmas. This was a non-exhaustive search, and it depended on the lexicographer's ability to recall missing senses (see Table 3, lines 4-5);

At this moment, our lexicographer has hand-checked 18,168 (38.25%) of the total set of candidate senses (i.e. 47,499 senses). Out of the total number of candidate senses checked, 8,295 (45.7%) were kept (i.e. the conversion of Mandarin lemmas was correct), which is in line with Snow's (2004, 49) predictions. In addition to these converted senses, a total of 3,797 new senses were added by the lexicographer (i.e. that were not suggested by the conversion from simplified Mandarin Chinese) – this comprises about 31.4% of the total number of senses we currently have in our wordnet, and which is in line with Ouyang's (1993, 23) predictions concerning the ratio of exclusive Cantonese senses. In total, our wordnet currently has 12,092 senses (a summary of this release's statistics is provided in Section 5).

4 Issues

4.1 Separated and Intervening Lexemes

What is represented by one lemma in English sometimes requires two lexemes separated from each other with an intervening lexeme in Cantonese. For example, 'to punch' in the sense of 'to deliver a quick blow' is expressed as [daai2...jat1kyun4], literally 'hit...one punch' (打... 一拳), where ... is the slot for the recipient of the punch, the object of the verb. Another example is 'to fire' in the sense of 'terminate the employment of', which can be expressed as [gaak3...zik1] (one of the many options in Cantonese), 'remove...duty' (革... 職), where ... is the slot of the person being fired, the object. This is essentially different from the English ‘pick up’ and ‘pick...up’ cases (where ... is the object) as [daai2jat1kyun4 + ‘object’] and [gaak3zik1 + ‘object’] are both ungrammatical – the separation is obligatory. In view of this, we have used separated lexemes (with ...) whenever it is necessary to be faithful to the English concept, a practice also adopted by COW.

4.2 Compositionality of Telic Verbs

In many cases, the translated term in Cantonese is compositional. For example ‘to remember’ in the sense of ‘recall knowledge from memory’, is nam2hei2 諗起 in Cantonese, where a post-verbal particle hei2 meaning ‘up’ is needed. Sybesma (1997) points out that Mandarin does not have monomorphemic counterparts for English verbs like ‘see’, ‘hear’, and ‘find’, which qualify as achievements (indeed he claims that Chinese has no inherently telic verbs at all). The Mandarin counterparts of these verbs are compound verbs, where the second constituent expresses the attainment of the result (‘phase complement’ in Chao (1968); see also Li and Thompson (1981)), e.g., kàndào 看到 ‘look-arrive > see’; kànjiàn 看見 ‘look-see > see’; tīngjiàn 聽見 ‘listen-see > hear’; zhǎodào 找到 ‘look for-arrive > find’. The situation is the same in Cantonese. To ensure we have high quality translation equivalents, these particles are included in the lemmas (the same procedure is adopted by COW). The consequence is that such entries can be analyzed as compositional.

4.3 The Lack of Standardization in Written Cantonese

Cantonese is primarily a spoken dialect. Cantonese has never been subjected to rigorous and formal standardization, despite efforts of lexicographers which resulted in a few Cantonese-standard Chinese dictionaries and Cantonese word lists (Li, 2000). Cantonese school children are not taught how to read or write Cantonese. The knowledge of written Cantonese among its speakers arises informally through exposure to its pervasive use (Bauer, 2018).

Written Cantonese is mainly used for informal or less serious kind of communication (Snow, 2004, 18), but is not uncommon. It is used regularly in advertising (e.g. signs, posters, novels) as well as newspapers (e.g. Apple Daily, a popular newspaper in Hong Kong). Written Cantonese conveys a greater degree (compare with standard Chinese) of ‘informality, directness, intimacy, friendliness, casualness, freedom, modernity and authenticity’ (Bauer, 2018, 4). At least partly due to the special situation in Hong Kong for a long time, where children speak Cantonese but write in standard Chinese (the situation has changed since the handover in 1997), written Cantonese ranges over a continuum. On the one end, there are texts that are essentially standard Chinese but with a few Cantonese items, on the other end are texts that are written entirely in Cantonese (Snow, 2004, 60-61).

There is substantial overlap between Mandarin and Cantonese vocabulary. For shared vocabulary items, e.g., 飯 ‘rice’, fàn in Mandarin and faan6 in Cantonese, the traditional version of the same character is used, and with a different pronunciation.

It is estimated that about one-third of the lexical items in Cantonese are not shared with Mandarin (Ouyang, 1993, 23). This also includes some very basic vocabulary, such as the negator, which is 不 bù in Mandarin and 唔 m4 in Cantonese, or very basic content words like ‘see’, which is 看 kàn in Mandarin, but 睇 tai2 in Cantonese. For Cantonese-specific lexical items, the choice of the characters is not always obvious due to the lack of standardization.

The standardization of written Cantonese lexical items exhibits a gradience, ranging from items like the negator 唔 m4 and ‘see’ 睇 tai2, which are not controversial, to items which are regularly represented phonetically with English letters in its written forms in online forums, e.g. *hea* he3 ‘to laze around’. In-between the two extremes, there are many cases where two or more characters are used to represent the same lexical item. For example the word bei2 ‘to give’ can be written with 4 different characters, 比, 俾, 畀, 被 (Bauer, 2018, 135). For this first version of the Cantonese Wordnet, items which are only represented by English letters are not listed. For cases where multiple characters are used, all options will be given whenever possible. For discussion on strategies on how Cantonese characters are formed, see Li (2000) and Bauer (2018).

4.4 Alternation in Pronunciation

In the Cantonese Wordnet, there are many cases where a particular character is given multiple pronunciations. The two common causes for alternation is *pinjam* 變音 ‘changed tone’ and laan5jam1 懶音 ‘lazy pronunciation’.

Many morphological constructions in Cantonese are expressed solely or partly by tone change (Yu, 2009). Traditional descriptive linguistic literature of Cantonese refers to this *pinjam* 變音 process. Table 4 shows some examples of tone change cases in deverbal nominalization (Yu, 2009).

character	verb	noun
掃	‘to sweep’ sou3	‘broom’ sou2
磅	‘to weight’ bong6	‘scale’ bong2
油	‘to grease’ jau4	‘oil’ jau4

Table 4: Cantonese Tone Change (I)

The term laan5jam1 (‘lazy pronunciation’, lǎnyīn in Mandarin, literally meaning ‘lazy pronunciation’) has been used in recent years to refer to ongoing sound changes in Hong Kong Cantonese. This term designates the use of a variety of consonant variants in the speech of younger native speakers of Hong Kong Cantonese (Ding, 2010). One example is syllable-initial /n/ and /l/ merger (/n/ > /l/), a phenomenon that started around the 70s. This is shown in the Table 5. There are many other examples of ‘lazy pronunciation’ (e.g., /ng/ > /m/) in Cantonese.

character	meaning	jyutping
男	‘male’	naam4 **or** laam4
女	‘female’	neoi5 **or** leoi5
呢度	‘here’	nei1 dou6 **or** lei1 dou6

Table 5: Cantonese Tone Change (II)

In addition to *pinjam* 變音 ‘changed tone’ and laan5jam1 懶音 ‘lazy pronunciation’, there are also cases of tone change, which are not clear what the motivation is. Nevertheless, whenever possible, all options were captured by our wordnet.

4.5 The Continuum between Spoken and Written Cantonese

Cantonese has different registers (e.g., everyday conversation vs. news report). A lot of words which are too formal to use in regular conversation might appear in TV broadcast, or formal speeches and thus some more formal versions of such terms (as long as they are deem possible in Cantonese) are also included in our wordnet with the aim of covering the range of registers. The consequence is that the boundary is not always clear. When in doubt, the decision was always to include such items.

The question as to what to include can be determined in a more objective way in the future. We would like to experiment with Cantonese texts of various registers, using both the Cantonese and Mandarin wordnets in parallel to help better understand and identify words that were not included as part of the Cantonese Wordnet. In time, we hope to establish the extent of shared vocabulary items between Mandarin and Cantonese, as well as to identify uniquely Cantonese items.

5 Statistics

Table 6 provides a summary of the current state of the Cantonese wordnet.

POS	No. synsets	%	No. senses	%
nouns	1,830	(0.52)	5,114	(0.42)
verbs	975	(0.28)	3,227	(0.27)
adjective	565	(0.16)	3,044	(0.25)
adverb	163	(0.05)	707	(0.06)
Total	3,533	-	12,092	-

Table 6: WN Statistics

In total, the first version of our wordnet covers a bit over 3,500 concepts using over 12,000 senses. The part-of-speech distribution is generally in sync with other projects, such as the PWN – with perhaps a weaker dominance of nominal senses and concepts to a slight heavier presence of their verbal counterparts. Our current version covers 35.81% (n = 1,776) of the ‘core’ PWN concepts.

Since our wordnet is currently pivoting on the hierarchy provided by PWN, through COW, we have no information about semantic relations to report. In further stages of our project, however, we might revise this position and consider taking advantage of CILI to adapt our wordnet's semantic hierarchy to better fit the assumptions of Cantonese native speakers.

As mentioned above, in Section 3.1, the process of human validation is still ongoing, and we expect to provide an update to these statistics in the camera-ready version of this paper.

6 Release

This Cantonese Wordnet will be released under a Creative Commons Attribution 4.0 International License (CC BY 4.0)[4].

Keeping up with the recent changes and requirements of the OMW, the Cantonese Wordnet will be primary released and supported for the recent WN-LMF format,[5] developed and maintained by the Global WordNet Association. The use of WN-LMF is not only required by the most recent version of the OMW, but is also an essential vehicle to access the new Collaborative Interlingual Index (Bond et al., 2016, CILI). Once linked to CILI, our wordnet will be able to contribute with new concepts, present only in Cantonese such as sap1jit6 濕熱, an adjective with the literal meaning of 'hot wet' (it describes a general negative health condition resulting from an unhealthy lifestyle, e.g. smoking, sleep deprivation, etc.), or gung1zyu2beng6 公主病, a noun that literally means 'princess disease'. It describes girls who are over-confident, over-reliant and demand princess-like treatment.

In addition, this release will also include the tab-separated-value format used by the original OMW specifications. These files are still very useful for their size, simplicity, and legacy compatibilities with existing systems. One such example is the use of this data through NLTK: Python Natural Language Toolkit (Bird et al., 2009) – which currently still uses this legacy format. However, the simplicity of this format doesn't come without a cost. Due to the flatter nature of this format, the Jyutping romanization of Cantonese lemmas will be added as separate lemmas (i.e. effectively doubling the number of words and senses within this format).

The data for this wordnet is available on Github[6].

7 Conclusions and Future Work

This paper presented the ongoing efforts to build a Cantonese Wordnet. We have motivated this project with the lack of digital resources available for Cantonese – a major Chinese dialect. We have introduced our methodology, which is to use existing Mandarin wordnets to project Cantonese candidate senses. So far our wordnet includes over 3,500 concepts and over 12,000 senses. We have discussed some specific challenges encountered while building our wordnet and how we addressed them. We hope that this new open resource will promote a variety of future uses, including language processing tasks and linguistic research.

We would like to continue our efforts to improve the coverage and quality of our Cantonese Wordnet. This would include:

- finish validating and revising the list of candidate senses generated through the methods explained in Section 3 (so far we have completed 38.25% of this validation);
- add example sentences for each sense, which would be the start of an open, sense-tagged Cantonese corpus;
- given that Cantonese is predominantly used in speech, we would also like to add audio recording for each pronunciation of each lemma;

Once the Cantonese wordnet reaches a sufficient coverage, we would like to use it to research a variety of topics, including:

- study the amount of Mandarin words that have entered common Cantonese speech and writing and, conversely, when and why some Mandarin words are never used Cantonese;
- study the morphologically conditioned tone changes in Cantonese such as *pinjam* and other less understood phenomena; and
- shed some light on the potential relation between register (formal register is often tied to written Chinese, which is based

[4] https://creativecommons.org/licenses/by/4.0/

[5] https://github.com/globalwordnet/schemas

[6] https://github.com/lmorgadodacosta/CantoneseWN

on Mandarin) and tone change (a speech phenomenon);

Acknowledgments

We thank the reviewers for their comments. Joanna Ut-Seong Sio gratefully acknowledges the research funding received from the Faculty of Arts, Palacky University in Olomouc through its Research Grant Scheme (funding cycle 2019-2021).

References

Robert Bauer. 2018. Cantonese as written language in hong kong. *Global Chinese*, 4(1):103–142.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1352–1362.

Francis Bond, Piek Vossen, John P McCrae, and Christiane Fellbaum. 2016. CILI: the Collaborative Interlingual Index. In *Proceedings of the Global WordNet Conference*.

Jordan Boyd-Graber, Christiane Fellbaum, Daniel Osherson, and Robert Schapire. 2006. Adding dense, weighted connections to wordnet. In *Proceedings of the third international WordNet conference*, pages 29–36.

Yuen Ren Chao. 1968. *Language and symbolic systems*, volume 457. CUP Archive.

Siu-Pong Cheng and Sze-Wing Tang. 2016. *Cantonese Romanizaton*. Routledge.

Picus Sizhi Ding. 2010. Phonological change in hong kong cantonese through language contact with chinese topolects and english over the past century. *Marginal dialects: Scotland, Ireland and beyond*, pages 198–218.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Zev Handel. 2015. The classification of chinese. *The Oxford handbook of Chinese linguistics*, page 34.

Chu-Ren Huang, Ru-Yng Chang, and Hshiang-Pin Lee. 2004. Sinica bow (bilingual ontological wordnet): Integration of bilingual wordnet and sumo. In *LREC*.

Chu-Ren Huang, Shu-Kai Hsieh, Jia-Fei Hong, Yun-Zhu Chen, I-Li Su, Yong-Xiang Chen, and Sheng-Wei Huang. 2010. Chinese wordnet: Design, implementation, and application of an infrastructure for cross-lingual knowledge processing. *Journal of Chinese Information Processing*, 24(2):14–23.

Chu-Ren Huang. 2003. Sinica bow: integrating bilingual wordnet and sumo ontology. In *International Conference on Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003*, pages 825–826. IEEE.

Shin Kataoka and Cream Yin-Ping Lee. 2008. A system without a system: Cantonese romanization used in hong kong place and personal names. *Hong Kong Journal of Applied Linguistics*, 11(1):79–98.

Charles Li and Sandra Thompson. 1981. A functional reference grammar of mandarin chinese. *Berkeley, CA: University of California Press. Find this author on.*

David CS Li. 2000. Phonetic borrowing: Key to the vitality of written cantonese in hong kong. *Written Language & Literacy*, 3(2):199–233.

Victor H Mair. 1991. *What is a Chinese" dialect/topolect"?: Reflections on some key Sino-English Linguistic Terms*.

Stephen Matthews and Virginia Yip. 1994. *Cantonese*. Routledge.

Jerry Norman. 1988. *Chinese*. Cambridge University Press.

Jueya Ouyang. 1993. Putonghua guangzhouhua de bijiao yu xuexi [comparison and study of putonghua and cantonese].

Yu Jie Seah and Francis Bond. 2014. Annotation of pronouns in a multilingual corpus of Mandarin Chinese, English and Japanese. In *Proceedings 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, page 82.

Don Snow. 2004. *Cantonese as written language: The growth of a written Chinese vernacular*, volume 1. Hong Kong University Press.

Rint Sybesma. 1997. Why chinese verb-le is a resultative predicate. *Journal of East Asian Linguistics*, 6(3):215–261.

Liling Tan and Francis Bond. 2014. NTU-MC toolkit: Annotating a linguistically diverse corpus. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 86–89.

Piek Vossen. 1998. A multilingual database with lexical semantic networks. *Dordrecht: Kluwer Academic Publishers. doi*, 10:978–94.

Shan Wang and Francis Bond. 2013. Building the chinese open wordnet (cow): Starting from core synsets. In *Proceedings of the 11th Workshop on Asian Language Resources*, pages 10–18.

Shan Wang and Francis Bond. 2014. Building the sense-tagged multilingual parallel corpus. In *LREC*, pages 2403–2409.

Renjie Xu, Zhiqiang Gao, Yingji Pan, Yuzhong Qu, and Zhisheng Huang. 2008. An integrated approach for automatic construction of bilingual chinese-english wordnet. In John Domingue and Chutiporn Anutariya, editors, *The Semantic Web*, pages 302–314, Berlin, Heidelberg. Springer Berlin Heidelberg.

Alan Yu. 2009. Tonal mapping in cantonese vocative reduplication. In *Annual Meeting of the Berkeley Linguistics Society*, volume 35, pages 341–352.

Anne Yue-Hashimoto. 1991. The yue dialect. *Journal of Chinese Linguistics Monograph Series*, (3):292–322.

Yingjie Zhang, Bin Li, Xiaoyu Wang, Xueyang Liu, and Jiajun Chen. 2014. Mapping word senses of middle ancient Chinese to WordNet. In *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 1, pages 446–450. IEEE.

Yingjie Zhang, Bin Li, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2017. Pqac-wn: constructing a wordnet for pre-qin ancient chinese. *Language Resources and Evaluation*, 51(2):525–545.