Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 70–78
Marseille, 20-25 June 2022

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

Joanna Ut-Seong Sio, Luis Morgado da Costa
Palacký University Olomouc, The Czech Republic
Katedra asijských studií FF UP, tř. Svobody 26, 779 00 Olomouc
joannautseong.sio@upol.cz, lmorgado.dacosta@gmail.com

Abstract

This paper reports on the most recent improvements on the Cantonese Wordnet, a wordnet project started in 2019 (Sio and Morgado da Costa, 2019) with the aim of capturing and organizing lexico-semantic information of Hong Kong Cantonese. The improvements we present here extend both the breadth and depth of the Cantonese Wordnet: increasing the general coverage, adding functional categories, enriching verbal representations, as well as creating the Cantonese Wordnet Corpus – a corpus of handcrafted examples where individual senses are shown in context.

Keywords: wordnet, Cantonese, functional categories, separable compound verbs, example corpus

1. Introduction

1.1. Background

This paper reports on the most recent version of the Cantonese Wordnet, started in 2019 (Sio and Morgado da Costa, 2019). Cantonese is the second most widely known Chinese 'dialect' after Mandarin (Matthews and Yip, 2013).[1] It is spoken in Guangdong Province, Guangxi Province, the Special Administrative Regions of Hong Kong and Macau, as well as throughout diaspora communities in North America, Australia, Malaysia, Singapore, etc. There are about 73 million Cantonese speakers worldwide (*Ethnologue*).[2] Our Cantonese Wordnet is built based on Hong Kong Cantonese (Hong Kong has a population of over 7 million people). Both Cantonese characters and romanization are used to represent the Cantonese lemmas in the Cantonese Wordnet. We adopted *Jyutping* (粵拼) as the romanization system for the Cantonese characters. Jyutping was developed by the Linguistic Society of Hong Kong (LSHK) in 1993. Its formal name is The Linguistic Society of Hong Kong Cantonese Romanization Scheme.[3] Since its inception, it is used widely in academic papers as well as social media. For more details about this romanization system, see Kataoka and Lee (2008) and Sio and Morgado da Costa (2019). 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 character are needed. For example, *sam1* can mean 'heart' (心) or 'dark/deep' (深). Without the character, the Jyutping transcription is ambiguous. The Cantonese Wordnet uses traditional characters.[4] Cantonese is primarily a spoken Chinese variety and has never been subjected to rigorous and formal standardization. The knowledge of written Cantonese among its speakers arises informally through exposure to its pervasive use (Bauer, 2018). The standardization of written Cantonese lexical items exhibits a gradience, ranging from items like the negator *m4* (唔) and 'to see' *tai2* (睇), which are not controversial, to items which are regularly represented phonetically with English letters in its written forms in online forums, e.g., *hea* (pronounced *he3*) meaning '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*, meaning 'to give', can be written with four different characters: 比, 俾, 畀, 被 (Bauer, 2018, 135). We include all options, whenever possible. Similarly, in cases where multiple pronunciations are possible (be it segmental or tonal), all possibilities are listed. This decision was made to future-proof our resource, and to make our wordnet useful in different language contexts. For further details regarding the methodological choices made in building the Cantonese Wordnet, see Sio and Morgado da Costa (2019).

1.2. New additions

At the center of this new version, we present the Cantonese Wordnet Corpus – a corpus of handcrafted examples where individual senses are shown in context. In addition, this paper also discusses other improvements made since our wordnet's first release. These improvements extend both the breadth and depth of the Cantonese Wordnet: increasing the general coverage, adding functional categories, and enriching verbal

[1] There are seven *fāngyán* (方言) ('dialect') groups in Chinese: Mandarin (or Northern Chinese), Xiang, Gan, Wu, Yue (Cantonese), Hakka and Min (Yuan, 1960). 'Dialects' is arguably not the best term to describe these Chinese varieties, as they are mutually unintelligible. Terms such as 'regionalect' (DeFrancis, 1986) and 'topolect' (Mair, 1991) have been proposed to alternatives (Tang, 2017).

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

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

[4] Traditional characters are used in Hong Kong, Macau and Taiwan, as opposed to Mainland China, where simplified characters were adopted since the 1950s.

representations. All these changes are geared towards the goal of making the Cantonese Wordnet a useful resource for linguistic analyses. In this paper, we discuss all the aforementioned updates in detail and also sketch the plan for future work.

2. Enriching the Sense Inventory

The development of the Cantonese Wordnet was and continues to be done manually. This was a conscious choice, embracing a slow development cycle to guarantee consistency and high quality data that would be suitable for empirical linguistic research.

The Cantonese Wordnet was originally built using the expansion approach, leveraging on the existing Chinese Open Wordnet (Wang and Bond, 2013, **COW**), and the Princeton Wordnet's (Fellbaum, 1998, **PWN**) semantic hierarchy. This project has since been slowly stepping away from the expansion approach in order to make this resource more useful to empirical linguistic research.

The original design of the PWN included only content words and open class words: nouns, verbs, adjectives and adverbs. However, we can see a slow movement towards the addition of new classes of concepts. Some examples of this are efforts concentrated on the inclusion of pronouns, interjections and classifiers (Seah and Bond, 2014; Morgado da Costa and Bond, 2016). This idea has been reinforced by the inclusion of a few new wordnet part-of-speech tags in the most current **WN-LMF** format[5] – notably accepting conjunctions, adpositions (prepositions, postpositions, etc.) and a general class for 'other' (e.g., particles, classifiers, bound morphemes, determiners) (P. McCrae et al., 2021). We follow this trend, and have added new concepts for some important functional categories in Cantonese.

There are at least two reasons why one should include functional categories in the Cantonese Wordnet. Firstly, many functional categories in Cantonese do not exist in English, e.g., sortal classifiers, post-verbal particles, sentence-final-particles, etc. Without including these categories, the representation of Cantonese would not be complete. Secondly, some of these functional elements must be added to a lemma in order to accomplish accurate cross-lingual linking of concepts.

In what follows, we will discuss the background and the procedures in incorporating two functional categories, classifiers and post-verbal particles, into the Cantonese Wordnet.

2.1. Classifiers

In numeral classifier languages, classifiers are obligatory when a noun is used with numerals (Allan, 1977). In Cantonese, classifiers have also been claimed to be related to the expression of definiteness (Cheng and Sybesma, 1999; Sio, 2006). In Cantonese (and in Chinese in general), classifiers, *liàngcí* (量詞), can be roughly divided into two kinds, sortal classifiers and measure words. The former names the unit that is already present in the semantic denotation of the noun (1) while the latter creates the unit (2) (Croft, 1994).

(1)	一	個	蘋果
	jat1	go3	ping4gwo2
	one	CL.sortal	apple

'an apple'

(2)	一	杯	茶
	jat1	bui1	caa4
	one	CL.measure	tea

'a cup of tea'

Sortal classifiers 'categorize' objects by picking out some salient perceptual properties (Tai and Wang, 1990; Del Gobbo, 2014). For instance, the Cantonese classifier *tiu4* (條) is used with long thin objects like ropes, straws, rivers, sausages, etc., while the Cantonese classifier *bun2* (本) is used with books. A count noun is often only paired with one sortal classifier, though occasionally more than one sortal classifier can be compatible with a count noun especially if different salient properties are being focused on, e.g., in Cantonese, *jat1 ceot1 hei3* 'one movie' vs. *jat1 coeng4 hei3* 'one scheduled showing of a movie'.

Measure words are a heterogeneous group (Cheng, 2012; Cheung, 2016). They can be: measuring units (e.g., *bong6* (磅) 'pound'), containers (e.g., *bui1* (杯) 'cup/glass'), collective terms (e.g., *kwan4* (群) 'group' (for animate objects)), terms denoting generic kinds (e.g., *zhǒng* (種) 'kind'), terms denoting the shape in which the objects/stuff can be gathered (e.g., *deoi1* (堆) 'heap' for count nouns like potatoes and *taan1* (灘) 'pool' for mass nouns like water), and terms denoting an indeterminate number/amount, such as *di1* (啲). The list here is not intended to be exhaustive. We consider all non-sortal classifiers to be measure words while being fully aware that measure words as defined constitute a heterogeneous group. Some of these measure words could be argued to have counterparts in the PWN, but they are mostly treated as nouns due to their syntactic behavior in English.

It is important to distinguish sortal classifiers from measure words because even though they generally occupy the same syntactic position (between a numeral and a noun), they differ not only in semantic content (measure words are also often nouns and have more tangible meaning) but also in terms of syntactic behaviors, e.g., adjectival modification of the classifier, insertion of the modification between the classifier and the noun, etc. (Cheng et al., 1998; Her and Hsieh, 2010). Distinguishing sortal classifiers from measure words will facilitate the future use of the Cantonese Wordnet for language processing applications. In view of that, we maintain the simple dichotomy of sortal classifiers vs.

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

measure words in the Cantonese Wordnet, with the plan of refining the classification in the future.

We focused our initial efforts on sortal classifiers as they don't have equivalents in English. We follow the steps of the Chinese Open Wordnet (COW), which has already incorporated many sortal classifiers for Mandarin (Morgado da Costa and Bond, 2016). However, a direct mapping of Cantonese sortal classifiers to their Mandarin counterparts is not possible for several reasons. Firstly, even though Mandarin and Cantonese classifiers do overlap, there are many classifiers that are unique to either dialect. For example, the Mandarin classifier *zhū* (株), used to classify a variety of flora, is generally not used in Cantonese. Furthermore, for classifiers that exist in both dialects, some have different coverage over nouns they are associated with. For example, the Mandarin classifier *zhī* (隻) is used to classify small animals such as birds, cats, rabbits, etc., but its Cantonese counterpart, *zek3* (隻), covers both large and small animals including ants, cows and dinosaurs, among other things (Erbaugh, 2013).

The current version of this wordnet added 41 sortal classifiers. Following the work done by Morgado da Costa and Bond (2016), these concepts received the part-of-speech 'x' (used for non-referential concepts), and have a standardized definition that tries to encapsulate the general class of nouns associated with each classifier, along with notable examples of those categories. For example, the sortal classifier *gaa3* (架) has as definition 'a sortal classifier used for wheeled vehicles such as a car, a motorcycle or a wheelchair'. The concepts are then linked through the concept relation `exemplifies` (used to indicate the usage of this word) to a newly added concept representing sortal classifiers.

In addition to sortal classifiers, we also added 25 measure words. We focused our efforts in including first the subclass known as container measure words. Container measure words can also function as nouns, e.g., *bui1* (杯) 'cup' can appear as a measure word in (3) and as a noun in (4), with its own classifier *zek3*:

(3)	一	杯	水
	jat1	bui1	seoi2
	one	CL.measure	water

'a cup of water'

(4)	一	隻	杯
	jat1	zek3	bui1
	one	CL.sortal	cup

'a cup'

As already mentioned above, many measure words in Cantonese have equivalents in PWN as nominal senses (e.g., kilogram, cup, piece). For the specific case of container measure words, the PWN actually has a small, flat hierarchy under the concept `13756125-n` defined as *the quantity that a container will hold*. Under this concept we find many English senses such as *mouthful*, *cupful*, *bowlful*, *roomful*, *houseful*, etc. The list goes on, but it becomes evident that these concepts were motivated by the presence of the semi-productive suffix '*-ful*', which is able to derive a new meaning denoting the quantity that can be held by a particular noun, when this noun can be interpreted as a container in English.

The existence of this small list of 57 concepts under the PWN's concept `13756125-n` poses a small dilemma on how to approach the analysis of measure words in the Cantonese Wordnet. On the one hand, we want to take advantage of the possibility of linking concepts across languages. On the other hand, treating measure words as regular nouns does not feel legitimate. We enumerate a few reasons why we chose to part from using nominal representations for measure words.

First, the list of 57 concepts defined under *the quantity that a container will hold* quickly become insufficient to represent all measure words. For instance, even though there are words representing concepts like *cupful* or *bowlful*, there are currently no concepts representing senses such as '*wok-ful*' or '*suitcase-ful*' (both 'wok' and 'suitcase' can function as measure words in Cantonese). As mentioned above, it seems that the list defined under the concept `13756125-n` is compiled from the words that were found to use the semi-productive suffix '*-ful*'. However these concepts include lemmas for both the original and affixed forms (e.g. both *bowl* and *bowlful* appear inside concept `13765531-n`, defined as *the quantity contained in a bowl*). This effectively allows imbalanced representations of some very similar words as shown below.

(5) He ate a <u>bowlful</u> of fried rice.

(6) He ate a <u>bowl</u> of fried rice.

(7) He ate a <u>wok</u> of fried rice.

Under the current analysis of the PWN, both example sentences (5) and (6) could use concept `13765531-n` denoting a meaning of quantity. But sentence (7) would not have a similar counterpart denoting a quantity. The only available sense for the lemma *wok* is `04596742-n`, defined as *pan with a convex bottom; used for frying in Chinese cooking*. The absence of the quantity sense for the lemma *wok* is most likely motivated by the absence of instances of the lemma 'wok-ful' – which would arguably sound odd to many, although not as odd as '*suitcase-ful*' most certainly would, and for which the same argument could be made.

In brief, we believe that the PWN has a poor treatment of senses denoting quantities: some container nouns have a quantity sense (e.g., cup, bowl) and some do not (e.g., wok, suitcase). It becomes a problem when linking the PWN to the Cantonese Wordnet, as the quantity reading is available when the Cantonese counterparts of 'wok' or 'suitcase' are used as measure words while 'wok-ful' and 'suitcase-ful' do not exist in PWN.

Second, in addition to missing senses denoting quantities, we also found certain senses defined under 13756125-n which do not seem to have a Cantonese counterpart. Some examples include: *houseful* and *roomful*. Although English has readily available senses to denote the quantity/volume contained in a house or room, Cantonese does not have similar measure words (i.e., the coercion of homologous words *house* (*uk1*, 屋) and *room* (*fong2*, 房) to function as measure words does not sound natural).

And finally, there is phonological evidence to suggest that the nominal and measure word concepts in Cantonese needed to be separated. When the morpheme for 'bag' 袋 is used as a noun (8), it is in tone 2; when it is used as a measure word, it is in tone 6 (9).[6]

(8) 三 個 袋
saam1 go3 doi2
three CL.sortal bag

'three bags'

(9) 三 袋 米
saam1 doi6 mai5
three CL.measure rice

'three bags of rice'

Another example is the morpheme for 'box' (盒): *hap2* is the noun and *hap6* is the measure word. This motivates the splitting of the measure word sense from the nominal sense, even though not all morphemes show different tones when they are used as a noun and as a measure word.

Our current working solution to deal with measure words is very similar to how we are dealing with sortal classifiers. These concepts are also using the part-of-speech 'x' (used for non-referential concepts), and have a standardized definition that tries to encapsulate the type of classifiers. For example, the measure word *bui1* (杯) has, as definition, '*a container measure word used to denote the quantity a cup or glass can hold*'. We once again use the concept relation exemplifies to link each measure word to a newly added concept representing measure words. We are aware that the part-of-speech 'x' might not be the best fit for measure words as they are intuitively more 'referential' than other functional categories. This is something we plan to look into in the future.

In addition, if a suitable nominal counterpart (e.g., a concept like *cupful* for *cup*) is available in the PWN, these two concepts are linked with the concept relation eq_synonym. This concept relation is defined by the Global Wordnet Association as '*a relation between two concepts where A and B are equivalent concepts but their nature requires that they remain separate (e.g., Exemplifies)*', and it has been used by the COW to link idiomatic phrases to their respective meanings while preserving the ability to classify these idiomatic phrases through the relation exemplifies – just like what we are doing with measure words. Using this relation allows us to create a synonymy relation across concepts belonging to different parts of speech (nouns and classifiers) and ensure a minimal level of cross-linking is available between the Cantonese Wordnet and the PWN.

In the near future we plan to follow the work shown in Morgado da Costa and Bond (2016) and make extensive use of newly available semantic relations in the WN-LMF (such as classified_by and classifies) to link classifiers to the nominal concepts they are generally associated with.

2.2. Post-verbal Particles

Cantonese has a very rich inventory of post-verbal particles. They can be subsumed under 4 types: aspectual, directional, resultative and quantifying (Matthews and Yip, 2013):

(i) Aspectual particles, such as *gan2* (緊) for the progressive aspect or *zo2* (咗) for the perfective aspect;

(ii) Directional particles, such as *dai1* (低) meaning 'down';

(iii) Resultative particles, such as *bao2* (飽) meaning 'full', *dou2* (到) meaning 'arrive'/'attainment', or *sei2* (死) meaning 'die'; and

(iv) Quantifying particles, such as *saai3* (晒) meaning 'all';

There are two kinds of aspect: 'viewpoint aspect' and 'situation aspect'. The former focuses on the temporal perspective of the situation. In a lot of languages, including English, 'viewpoint aspect' can be indicated by inflectional affixes. For example, the progressive *-ing* suffix in English expresses viewpoint aspect. Situation aspect, also known as Aktionsart or lexical aspect, is concerned with the internal structure of the situation, encoded by the verb phrase. Common situation classes include states, activities, accomplishments, achievements and semelfactives (Rothstein, 2008). The choice of aspectual particles in Cantonese is sensitive to both kinds of aspects. Both aspectual and quantifying particles lack lexical content and cannot be used separated from the verb. Directional and resultative particles can be used in positions other than post-verbally.

Post-verbal particles are needed in order to accomplish accurate cross-lingual linking of concepts. We present a few examples here for illustration. It has been observed that achievement verbs in Mandarin Chinese are often compound verbs (Sybesma, 1997). This also applies to Cantonese. In Cantonese, the equivalent of the English achievement verb 'to find' is *wan2 dou2* (搵到), with

[6] See Alan (2007) for other cases of morphological tone change in Cantonese.

the first character, *wan2* (搵), meaning 'to look for' and the second character, *dou2* (到), which is a post-verbal particle of the resultative type, denoting that the action has been brought to a successful end.[7] Another example is the Cantonese equivalent of the English verb 'to remember', in the sense of 'to recall knowledge from memory'. Its Cantonese counterpart is *nam2 hei2* (諗起): *nam2* (諗) means 'to think' and *hei2* (起) is a directional particle meaning 'up'. The concept 'to remember' in Cantonese must be formed from both 'think' and 'up'.

Not all verbs are compatible with all post-verbal particles. Their compatibility is determined by their respective semantic and syntactic classes. As a language resource, it would be valuable if such compatibility information is available. To achieve this goal, we have started by testing the compatibility of our current coverage of verbs with the perfective aspectual particle *zo2*. We will report our finding in section 2.4.

The current version of the Cantonese Wordnet includes 32 post-verbal particles. We added a nominal concept that introduces 'post-verbal particles', and added the four sub-types of post-verbal particles discussed above (also as nominal concepts). As non-referential concepts, post-verbal particles take 'x' as part-of-speech, and have a definition that alludes both to their category and their meaning.

We hope to keep building this hierarchy of important functional categories in Cantonese, which we believe can be instrumental for many future lines of research in Cantonese Linguistics.

2.3. Increasing Sense Coverage

In addition to adding new classes of concepts to the wordnet, we have also continued expanding the sense repository of our wordnet.

As first described in Sio and Morgado da Costa (2019), the creation of our sense repository is currently done mainly through a process of validation, using data generated automatically through COW, along with our lexicographer's input on missing words and jyutping readings for each sense — see the original paper for details about this validation process.

The first version of the Cantonese Wordnet contained 12,092 senses, distributed across 3,533 concepts, and covered 35.81% of the 'core' PWN concepts.

The current version of the Cantonese Wordnet contains over 16,000 senses distributed over more than 5,000 concepts. Table 1 shows detailed numbers of the current size of our wordnet, along with information concerning part-of-speech distribution of concepts and senses.

Each individual sense in the Cantonese Wordnet comes with human quality jyutping romanisation. This information is stored inside each sense as a form, and hence it does not count for the number of senses. In total, the latest version of the Cantonese Wordnet has more than 32,800 forms (including character-based lemmas and jyutping forms), and it now covers 52.8% of PWN's 'core' concepts.

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	-

Table 1: Cantonese Wordnet Statistics

2.4. Enriching Verbal Representations: Separability of Verbal Compounds

Many meaning components that are lexicalised in English are represented by compound verbs in Cantonese. Typologically, Chinese is considered more analytic than English. In Chinese, it is generally the case that one character corresponds to one syllable and one morpheme, with some exceptions (e.g., *pútao* 葡萄 'grape', has two syllables/characters but represents just one morpheme). We mentioned in the previous section that in order to ensure accurate cross-lingual linking of concepts between English and Cantonese, Cantonese verbal lemmas might contain a verb and a post-verbal particle. There are other kinds of compound verbs as well. For example, the verb 'to run' in English is *paau2 bou6* (跑步) in Cantonese, with *paau2* (跑) meaning 'run' and *bou6* (步) meaning '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. For example, *lei4 fan1* (離婚) in Cantonese is a verb-object compound verb that corresponds to the lemma 'divorce' in English. The first morpheme *lei4* (離) means 'separate' and the second morpheme *fan1* means 'marriage' (婚). The compound verb allows post-verbal particles, (10), as well as frequency adverbs, (11), to be inserted in between:

(10) 離 緊 婚
lei4 gan2 fan1
separate gan.prog marriage

'in the processing of divorcing'

(11) 離 咗 三 次 婚
lei4 zo2 saam1 ci3 fan1
separate gan.perf three time marriage

'divorced three times'

To the extent that the Cantonese Wordnet will be used as a linguistic resource, whether a compound verb is separable or not is an important piece of information.

[7] *Dào* (到), the Mandarin counterpart of the Cantonese *dou2* (到), is also referred to as a phase-complement (Chao, 1965).

Compound verbs in Cantonese can be of different types, depending on the morpho-syntactic relationship between the morphemes in the compound (Chan and Cheung, 2020): verb-object, subject-predicate, coordinate, subordinative and verb-resultative. Verb-object compounds are observed to be the most separable (but not always), and the rest of the compound types vary (Chan and Cheung, 2020). Since separability is not totally predictable, at least according to our current understanding, and there can be different kinds of separating elements leading to different results, we started with just one separating element, the perfective aspectual particle *zo2*. *Zo2* is chosen because it is one of the most common aspectual particles. Testing the placement of *zo2* on verbs would also provide information on the compatibility between *zo2* and different kinds of verbs.

We tested every verb sense individually (including monosyllabic and multi-syllabic verbal lemmas). For each verb sense, we checked if it was compatible with *zo2*, and if so, we indicated where *zo2* was placed (it is trivial for monosyllabic verbs, as in such cases there is only one post-verbal position). In addition, we have also added information on transitivity, indicating whether each verbal sense accepts a complement (regardless of its type: e.g., nominal or clausal). We hired a graduate student who is a native Hong Kong Cantonese speaker to help with this task. The student's work was monitored by the authors by having weekly meetings to discuss and review problematic cases. About 15% of all the cases had been checked twice by both the graduate student and one of the authors, who is also a native Hong Kong Cantonese speaker.

Overall we have 4,200 verbal senses, with the following verb length distribution: 14.2% monosyllabic verb senses, 80.4% disyllabic verb senses and 5.4% verb senses with more than 2 syllables. It is not surprising that the majority of verbs are disyllabic. Modern Chinese shows a strong tendency to form disyllabic words, it is estimated that over 80% of the words in modern Chinese are disyllabic (Shi, 2002; Basciano, 2017).

Our work yielded the following results: the large majority, approximately 82.4%, of the verb senses is transitive. 11.5% of all the verb senses in our wordnet do not take *zo2* at all. The large majority of verbs, 79.5%, take *zo2* only at the end. About 8.9% of verbs are separable, allowing the placement of *zo2* in the middle of the lemma. And only 0.3% (only 12 verbs) seem to be able to take *zo2* both in between and after the verb.

Regarding the subset of verbal senses that are not compatible with *zo2* at all, some of them belong to a more formal register. Even though such verbs can be used in Cantonese, it is very unnatural to combine them with *zo2*. One such example is *hok6 zap6* (學習) 'to learn', which is not compatible with *zo2*. However its monosyllabic counterpart *hok6* (學) is. It has been observed that many Chinese words (or morphemes) can have a long (disyllabic) version or a short (monosyllabic) version, where the former contains the latter (Duanmu, 2017). Duanmu (2017) also mentions that in some pairs, the long form is more formal, more abstract, or of a larger quantity. In this particular case, the disyllabic version seems to be more formal and abstract. We speculate that this would be related to its incompatibility with *zo2*. Another reason of incompatibility is due to aspectual properties. *Zo2* is not compatible with states (Sybesma, 2004). Verbal compounds expressing states, e.g., *seoi1 jiu3* (需要) 'to need', are not compatible with *zo2*. For cases among this subset of verb senses with multi-syllabic forms, we could not ascertain whether they are separable or not, and would require further testing with different separating elements.

Through this exercise, we have enriched the Cantonese Wordnet with a variety of linguistic information concerning: transitivity, compatibility with the perfective aspectual particle *zo2* and separability of verb compounds. For verbs that are separable, we have marked them as such with the `exemplifies` relation. With the other information, we haven't yet come upon a proper representation. Thus, as of now, such information will be provided as external resources of the Cantonese Wordnet.

3. Building an Example Corpus

Despite being a fairly under-resourced Chinese variety, it is still possible to find publicly available corpora of Cantonese. Some examples of this include: **CantoneseWaC** – the Cantonese Web Corpus[8]; the **CantoMap** – a Hong Kong Cantonese MapTask Corpus[9] (Winterstein et al., 2020); the **HKCanCor** – Hong Kong Cantonese Corpus[10] (Luke and Wong, 2015); and the Corpus of Mid-Twentieth-Century Hong Kong Cantonese[11] (Chin, 2019).

Different kinds of corpora are best suited for different kinds of research. A general trend we find in the above mentioned corpora is that they are mostly sourced from speech/spoken data – this is true for all the above, except CantoneseWaC.

While speech corpora can be useful for a variety of research questions and computational tasks, they often come transcribed and annotated to preserve speech artifacts like fillers and pauses, and don't often contain data that is ideal to extract clean example sentences from.

Even though it is also a spoken corpus (collected by transcribing dialogue of Cantonese movies), the Corpus of Mid-Twentieth-Century Hong Kong Cantonese does actually have valuable data that could be useful for our purpose. Unfortunately, despite being openly accessible, this corpus does not seem to be published under an open license – effectively preventing us from freely

[8] `https://www.sketchengine.eu/cantonesewac-corpus/`

[9] `https://github.com/gwinterstein/CantoMap`

[10] `http://compling.hss.ntu.edu.sg/hkcancor/`

[11] `https://hkcc.eduhk.hk/v2`

reusing and redistributing parts of it. From the above-mentioned corpora, only CantoMap and the HKCanCor are published under open licenses.

Finally, concerning CantoneseWaC, in addition to not using an open license, this corpus shows problems that are shared among many online Cantonese written sources: CantoneseWaC is made up of texts collected from the Internet, using Cantonese seed words for crawling texts (Kilgarriff et al., 2010). This, unfortunately, produces extremely noisy data. The main reason is that Cantonese is a spoken variety (as most other Chinese dialects). In Hong Kong, people speak Cantonese but write in standard Chinese. Written Cantonese (mainly for non-formal purposes) ranges over a continuum. On the one end, there are texts that are essentially standard Mandarin Chinese but with a few Cantonese items, on the other end are texts that are written entirely in Cantonese (Snow, 2004). To exemplify this problem, out of the 30 most frequent tokens in the CantoneseWaC, quite a few are strictly non-Cantonese (e.g., 的, 是, 在 and 也). The majority of instances of these lemmas are most definitely extracted from Mandarin text or Mandarin-Cantonese mixed text. The continuum (or shift) between Mandarin and Cantonese in written discourse makes online written Cantonese data unsuitable for research focusing only on Cantonese.

As an answer to the problems raised above, we have started collecting an example corpus that will be updated and released alongside the Cantonese Wordnet.

The Cantonese Wordnet Corpus currently contains 3,570 example sentences. These example sentences are hand-crafted, by two native speakers, and are uniquely created for this corpus. Each sentence is created with a particular sense in mind, and each example sentence is crafted taking into consideration the common usage of that sense (e.g., the context provided for the sentence should be as natural as possible).

In addition, we predict that the majority of sentences in the Cantonese Wordnet Corpus will be created to test or illustrate particular syntactic phenomena. For example, the large majority of examples sentences released in the current version of this corpus are related to the task described in Section 2.4. While testing the interaction with the particle *zo2*, each verb sense that could interact with this particle was given an example showcasing this interaction. Following this trend, we expect that each individual sense in the Cantonese Corpus to acquire multiple example sentences over time, with each example sentence exploring one or more key syntactic phenomena in Cantonese. The large majority of example sentences contained in this new corpus are based on verb senses, with a small part distributed over the new concept classes we've introduced above: sortal classifiers, measure words and post-verbal particles.

The Cantonese Wordnet Corpus currently comprises 82,500 characters. Individual sentences have an average of 23.8 characters (with the shortest having 8 characters and the longest having 74 characters).

This corpus will be released by linking each individual sentence to the individual sense they were created for, using the specifications of the WN-LMF schema.

In the future, we plan to use this corpus to bootstrap the development of the Cantonese Wordnet, by providing a fully sense-annotated corpus – similar to other sense annotated corpora such as the English SemCor and many similar corpora.[12]

4. Release

The Cantonese Wordnet is released under a Creative Commons Attribution 4.0 International License (CC BY 4.0)[13]. This new version of the Cantonese Wordnet will soon be made available on its Github repository.[14] Similar to previous releases, this new version of the Cantonese Wordnet will be supported both in the WN-LMF format,[15] developed and maintained by the Global WordNet Association, as well as the legacy tab-separated-value format used by the original Open Multilingual Wordnet (Bond and Foster, 2013) specifications. This legacy format is still very useful due to its reduced size, simplicity, and legacy compatibility with existing systems.

4.1. Linking to CILI

There are a few new concepts in the Cantonese Wordnet that are not present in the Princeton Wordnet. This effectively deviates from the expansion approach (i.e., purely translating concepts), and introduces future problems for cross-linking with other wordnets.

The solution to this problem is CILI – the Collaborative Interlingual Index (Bond et al., 2016), currently under development by the Global Wordnet Association. When this infrastructure becomes ready, all new concepts in the Cantonese Wordnet will be proposed as new CILI concepts, which will facilitate the cross-lingual link with other wordnets that may have similar concepts to the ones we have added.

CILI will also allow us to add other open class concepts that are unique to Cantonese. To give an exmaple, *gung1 zyu2 beng6* (公主病) is a noun that literally means 'princess disease' and is a derogatory term used to describe females who are arrogant, narcissistic, over-reliant and who demand princess-like treatment. Thus far, we have actively avoided developing too much this area of our wordnet since the CILI infrastructure is not yet fully in place and we are not sure what changes will be necessary to satisfy the CILI requirements for proposing new concepts. Despite this, when the time arrives, we hope that the Cantonese Wordnet will become an active contributor to CILI.

[12] `http://globalwordnet.org/resources/wordnet-annotated-corpora/`

[13] `https://creativecommons.org/licenses/by/4.0/`

[14] `https://github.com/lmorgadodacosta/CantoneseWN`

[15] `https://github.com/globalwordnet/schemas`

5. Discussion and Conclusion

This paper reports on the ongoing efforts in building the Cantonese Wordnet. This wordnet was initially built by leveraging on the existing wordnets but has since slowly expanded its structure beyond previous projects – exploring new classes of concepts such as classifiers and post-verbal particles.

Even though there is considerable overlap in lexical items between Mandarin Chinese and Cantonese, their phonological differences, the lack of standardization in written Cantonese as well as the complex tonal alternation phenomena in Cantonese have provided a new set of challenges and methodological insights which we hope could be applied to build wordnets of other Chinese dialects.

Currently, our wordnet includes a little over 5,200 concepts and 16,300 senses – which is a fair improvement over the first version of this resource. In addition, this new version also releases a new open corpus for Cantonese, providing example sentences for more than 3,500 senses in this resource.

We are committed to continue to improve its coverage and quality. As such, we have identified key areas of improvement where we hope to focus next:

- finish validating and revising the list of candidate senses generated through the methods described in Sio and Morgado da Costa (2019). So far we have completed approximately 55% of this validation;
- continue adding sentences to the Cantonese Wordnet Corpus. We strongly believe this will be a very useful resource for linguistic research, and hope similar projects would follow our footsteps. In addition, we would also like to start a fuller sense annotation of this corpus;
- given that Cantonese is predominantly used in speech, we would also like to add audio recording for the pronunciation of each lemma. This would make the Cantonese Wordnet a useful resource to be used for learning Cantonese;
- incorporate other functional categories, such as sentence-final-particles, idioms and interjections.

6. Acknowledgements

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.

7. Bibliographical References

Alan, C. (2007). Understanding near mergers: The case of morphological tone in cantonese. *Phonology*, 24(1):187–214.

Allan, K. (1977). Classifiers. *Language*, 53(2):285–311.

Basciano, B. (2017). Brand names” . In Rint Sybesma, editor, *Encyclopedia of Chinese Language and Linguistics*, vol. 1. Brill, Leiden.

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

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

Bond, F., Vossen, P., McCrae, J. P., and Fellbaum, C. (2016). Cili: the collaborative interlingual index. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 50–57.

Chan, S. S. and Cheung, L. Y. (2020). Morpho-syntax of non-vo separable compound verbs in cantonese. *Studies in Chinese Linguistics*, 41(2):185–206.

Chao, Y. R. (1965). *A grammar of spoken Chinese*. ERIC.

Cheng, L. L.-S. and Sybesma, R. (1999). Bare and not-so-bare nouns and the structure of np. *Linguistic Inquiry*, 30(4):509–542.

Cheng, L. L.-S., Sybesma, R., et al. (1998). Yi-wan tang, yi-ge tang: Classifiers and massifiers. *Tsing Hua journal of Chinese studies*, 28(3):385–412.

Cheng, L. L.-S. (2012). Counting and classifiers. In Diane Massam, editor, *Count and Mass*, pages 199–219. Oxford University Press, Oxford.

Cheung, C. C.-H. (2016). *Parts of speech in Mandarin: The state of the art*. Springer, Singapore.

Chin, A. C.-o. (2019). Initiatives of digital humanities in cantonese studies: A corpus of mid-twentieth-century hong kong cantonese. In *Digital Humanities and New Ways of Teaching*, pages 71–88. Springer.

Croft, W. (1994). Semantic universals in classifier systems. *Word*, 45(2):145–171.

DeFrancis, J. (1986). *The Chinese language: Fact and fantasy*. University of Hawaii Press.

Del Gobbo, F. (2014). Classifiers. *The handbook of Chinese linguistics*, pages 26–48.

Duanmu, S. (2017). Word and wordhood, modern. In *Encyclopedia of Chinese language and linguistics*, volume 4, pages 543–49. Brill.

Erbaugh, M. (2013). Classifier choices in discourse across the seven main chinese dialects. *Increased empiricism: Recent advances in Chinese linguistics*, pages 101–126.

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

Her, O.-S. and Hsieh, C.-T. (2010). On the semantic distinction between classifiers and measure words in chinese. *Language and linguistics*, 11(3):527–551.

Kataoka, S. and Lee, C. (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.

Kilgarriff, A., Reddy, S., Pomikálek, J., and Avinesh, P. (2010). A corpus factory for many languages. In *LREC*.

Luke, K. K. and Wong, M. L. (2015). The hong kong cantonese corpus: design and uses. *Journal of Chinese Linguistics*, 25(2015):309–330.

Mair, V. H. (1991). *What is a Chinese ”dialect/topolect”? Reflections on some key Sino-English linguistic terms*. Department of Oriental Studies, University of Pennsylvania Philadelphia.

Matthews, S. and Yip, V. (2013). *Cantonese: A comprehensive grammar*. Routledge.

Morgado da Costa, L. and Bond, F. (2016). Wow! what a useful extension! introducing non-referential concepts to wordnet. In *Proceedings of the 10th edition of the International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia*.

P. McCrae, J., Wayne Goodman, M., Bond, F., Rademaker, A., Rudnicka, E., and Morgado Da Costa, L. (2021). The globalwordnet formats: Updates for 2020. In *Proceedings of the 11th Global WordNet Conference (GWC 2021)*, Pretoria, South Africa, January. Global Wordnet Association.

Rothstein, S. (2008). Telicity, atomicity and the vendler classification of verbs. *Theoretical and crosslinguistic approaches to aspect*, pages 43–77.

Seah, Y. J. and Bond, F. (2014). Annotation of pronouns in a multilingual corpus of mandarin chinese, english and japanese. In *Proceedings of the 10th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*.

Shi, Y. (2002). *The establishment of modern Chinese grammar: The formation of the resultative construction and its effects*, volume 59. John Benjamins Publishing.

Sio, J. U.-S. and Morgado da Costa, L. (2019). Building the cantonese wordnet. In *Proceedings of the 10th Global Wordnet Conference (GWC 2019), Wroclaw, Poland*.

Sio, J. U.-S. (2006). *Modification and reference in the Chinese nominal*. LOT Publications.

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

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

Sybesma, R. (2004). Exploring cantonese tense. *Linguistics in the Netherlands*, 21(1):169–180.

Tai, J. and Wang, L. (1990). A semantic study of the classifier tiao. *Journal of the Chinese Language Teachers Association*, 25(1):35–56.

Tang, C. (2017). Dialects of chinese. In *The Handbook of Dialectology*, pages 547–558. Wiley Online Library.

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

Winterstein, G., Tang, C., and Lai, R. (2020). Cantomap: a hong kong cantonese maptask corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2906–2913.

Yuan, J. (1960). *An outline of Chinese dialects*. Wenzi Gaige Chubanshe.