# The Open Cantonese Sense-Tagged Corpus

**Joanna Ut-Seong Sio**
Palacký University Olomouc
The Czech Republic
joannautseong.sio@upol.cz

**Luis Morgado da Costa**
Vrije Universiteit Amsterdam
The Netherlands
lmorgado.dacosta@gmail.com

## Abstract

This paper introduces the Open Cantonese Sense-Tagged Corpus, a new and ongoing project to serve as the companion to the development of the Cantonese Wordnet. This corpus is built on top of the Cantonese Wordnet Corpus, which currently provides example sentences for most verbs in this wordnet. This paper motivates the choice of starting a sense-tagged corpus from both linguistic and educational perspectives, and discusses the current solutions to issues arisen from the sense-tagging exercise. In total, we have tagged over 5,000 concepts, with more than 3,700 direct links to the Cantonese Wordnet.

## 1 Introduction

This paper presents the first sense-tagged corpus for Cantonese, an open corpus being built with and alongside the development of the Cantonese Wordnet (Sio and Morgado da Costa, 2019).

Sense annotation is the task of pairing a corpus with a semantic lexicon, by linking every *substantive* word in the corpus to its correct sense (as represented in the lexicon). This kind of annotation can help identify a variety of problems in the lexicon, such as missing senses or indistinguishable definitions, and hence helps improve both the coverage and the precision of the lexicon being used in the annotation (Miller et al., 1993). And it can also contribute to the concept of attestation, which is becoming a common requirement in most large lexicographic projects. [1]

While building sense annotated corpora is an extremely time-consuming task, building better language resources (both corpora and lexicons) addresses some of the ever-increasing needs required to solve complex Natural Language Processing problems such as information retrieval, machine translation, and automatic summarization.

The earliest project attempting to do sense annotation with wordnets was SemCor (Landes et al., 1998), a companion corpus to the Princeton Wordnet (PWN, Fellbaum, 1998) – the first wordnet, and the first sense-tagged corpus. Since then, a large number of wordnets started to emerge, alongside similar sense-tagged corpora. A good summary of the existing work in this field can be found in Petrolito and Bond (2014), which reports finding more than 20 sense-annotated corpora using wordnets, in more than 10 different languages.

In addition to the reasons stated above, which would already be sufficient, our project is also motivated from an educational standpoint. Despite being widely spoken, many scholarly efforts often seem to forgo Cantonese in preference to other varieties of Chinese (e.g., Mandarin). This project is one more contribution to support this language's maintenance and preservation. We believe that, if planned properly, sense annotated corpora can serve as excellent resources for language education – especially if the data being sense-tagged is suitable to be used in educational contexts. This is also why we chose to start the annotation using the Cantonese Wordnet Corpus (Sio and Morgado da Costa, 2022) – which comprises hand-crafted examples from a variety of day-to-day, modern and culturally-appropriate contexts.

## 2 Methodology

This paper reports an experiment that sense-tagged 300 random sentences extracted from the Cantonese Wordnet Corpus. These sentences were segmented manually by a native Cantonese speaker studying linguistics, and revised by a second native speaker who is a senior linguist. We are aware that the notion of 'word' is a contentious issue in Chinese languages (including Cantonese) (Packard, 2000). The native speakers were instructed to segment sentences (into words) based on their intuition, while taking into consideration both on-

[1] See, e.g., https://en.wiktionary.org/wiki/Wiktionary:Criteria_for_inclusion#Attestation

going linguistic discussion on Chinese wordhood (e.g., freedom-of-parts, semantic and structural non-compositionality, etc., Chu-Ren et al., 2017), and previous decisions made in the process of building the Cantonese Wordnet.

The tagging is being carried out by a single native Cantonese speaker lexicographer, but annotation issues and solutions are frequently discussed with the maintainers of the Cantonese Wordnet.

We are currently using IMI – a multilingual semantic annotation environment (Bond et al., 2015)[2]. IMI was designed for multilingual sense annotation. But in addition to sense-tagging, it provides multiple layers of annotation that include lemmatization, POS tagging, sentiment annotation and interlingual-mapping. This annotation tool has been tested for a wide selection of languages (i.e., English, Mandarin, Japanese and Indonesian) while tagging the NTU Multilingual Corpus (Tan and Bond, 2014; Bond et al., 2013) – a project that heavily influenced our corpus.

IMI uses an interface to the Open Multilingual Wordnet (OMW, Bond and Foster, 2013) to show candidate senses for concepts in the corpus. Fig. 1 shows an example of how our corpus is being created. In addition to data from the Cantonese Wordnet, we also rely on data from PWN and the Chinese Open Wordnet (COW, Wang and Bond, 2013) to find the right concepts.

Because we considered this preliminary work an exercise to fool-proof future annotation efforts, we decided to tag concepts sequentially, as they appear in a sentence, instead of relying on more efficient annotation methods such as tagging all instances of the same concept all at once (see Wang and Bond, 2014).

Clicking on a word in the corpus generates a web form upon which the lexicographer can make a decision based on existing senses in the wordnet. In the example shown in Fig. 1 we see an attempt to tag the word '會' wui5 (highlighted in yellow, around the middle of the figure). This word could be tagged as any of three concepts currently in the Cantonese Wordnet (numbered from 1 to 3, on the right side). In this case, the correct tag is the concept number 3, which is shown by the selection of the appropriate bullet on the left side, below the main text. In addition to the senses provided by the wordnet, the annotator has a few other options to choose from:

- the tag **e** notes that there is some sort of error in the corpus. This can be a segmentation or orthographic mistake, or an idiomatic but separable multi-word expression – which failed to generate automatically;
- the tag **x** is used for words that should not be sense-tagged. Currently, this is only being used for punctuation. In previous projects this tag was used, e.g., to tag determiners or auxiliary verbs in English. However, with the move to adding more and more parts-of-speech to wordnets such as pronouns, interjections and classifiers (see: Seah and Bond, 2014; Morgado da Costa and Bond, 2016), this tag is used less and less;
- the tag **w** notes that the wordnet is missing the right concept to tag the word in question. In cases where the OMW hierarchy has the right concept but the Cantonese Wordnet was missing a sense, we add the missing sense using OMWEdit (Morgado da Costa and Bond, 2015) – a tool integrated into IMI which allows editing a wordnet on the fly. However, even though this tool also allows adding new concepts to the semantic hierarchy, we decided not to use this feature for the moment (see Section 4);
- the tags **Org**, **Loc**, **Per**, **Dat**, **Oth**, **Num**, and **Year** are used to tag named entities and other productive expressions (e.g., dates, time expressions) that cannot be found in the wordnet;[3]

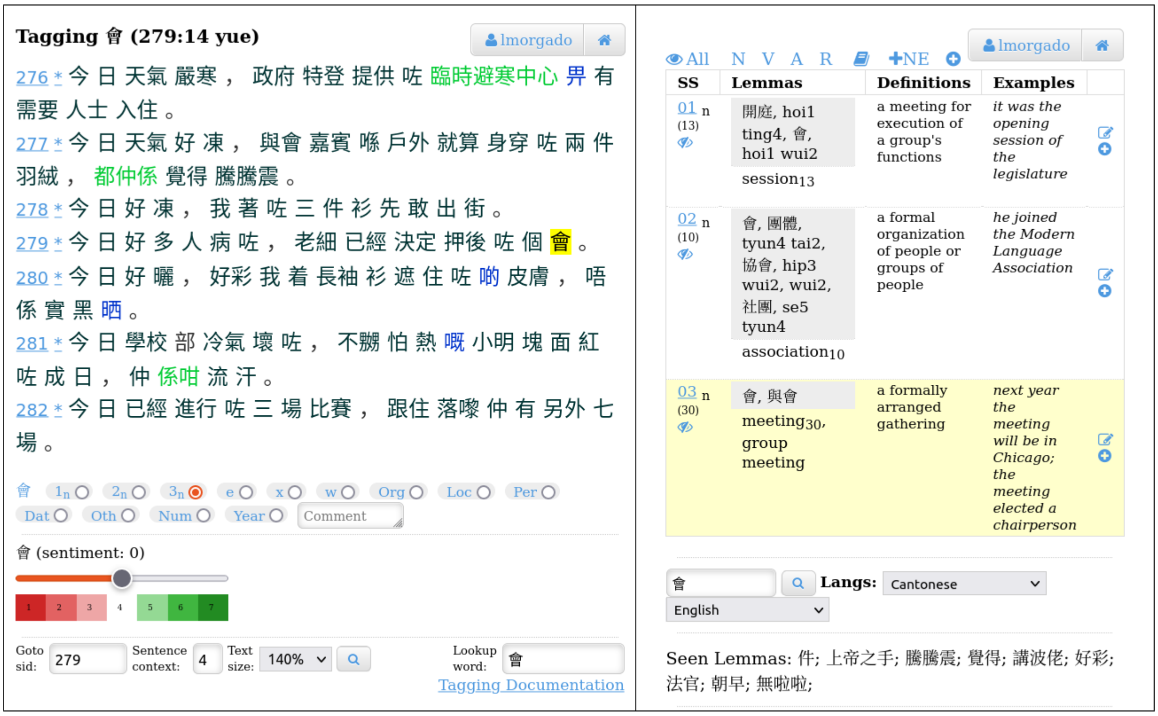
## 3 Tagging Results and Release

The results from the tagging exercise are summarized in Table 1. In total, the 300 sentences discussed in the section above generated a total of 5,279 candidate concepts. This number closely reflects the work done for segmentation, where each word was considered a possible concept. The tagged corpus contains 3,728 concepts linked to the Cantonese Wordnet.

The remaining lines in Table 1 should be interpreted with reference to the discussion of Fig. 1, above. In summary, the lexicographer identified 196 errors in the corpus – comprising segmentation errors, orthographic mistakes, and instances of separable idiomatic expressions – all of which will

[2] https://github.com/bond-lab/IMI

[3] A fuller guide on how to use these tags can be found in: https://github.com/bond-lab/IMI/blob/develop/docs/tagdoc.html



**Tagging 會 (279:14 yue)**

276 * 今 日 天氣 嚴寒 ， 政府 特登 提供 咗 臨時避寒中心 畀 有 需要 人士 入住 。

277 * 今 日 天氣 好 凍 ， 與會 嘉賓 喺 戶外 就算 身穿 咗 兩 件 羽絨 ， 都仲係 覺得 騰騰震 。

278 * 今 日 好 凍 ， 我 著 咗 三 件 衫 先 敢 出 街 。

279 * 今 日 好 多 人 病 咗 ， 老細 已經 決定 押後 咗 個 會 。

280 * 今 日 好 曬 ， 好彩 我 着 長袖 衫 遮 住 咗 啲 皮膚 ， 唔 係 實 黑 晒 。

281 * 今 日 學校 部 冷氣 壞 咗 ， 不嬲 怕 熱 嘅 小明 塊 面 紅 咗 成 日 ， 仲 係咁 流 汗 。

282 * 今 日 已經 進行 咗 三 場 比賽 ， 跟住 落嚟 仲 有 另外 七 場 。

會 ○ $1_n$ ○ $2_n$ ● $3_n$ e ○ x ○ w ○ Org ○ Loc ○ Per ○ Dat ○ Oth ○ Num ○ Year ○ Comment

會 (sentiment: 0)

Goto sid: 279 Sentence context: 4 Text size: 140%

Lookup word: 會

| SS | Lemmas | Definitions | Examples |
|---|---|---|---|
| 01 n (13) | 開庭, hoi1 ting4, 會, hoi1 wui2<br>session13 | a meeting for execution of a group's functions | *it was the opening session of the legislature* |
| 02 n (10) | 會, 團體, tyun4 tai2, 協會, hip3 wui2, wui2, 社團, se5 tyun4<br>association10 | a formal organization of people or groups of people | *he joined the Modern Language Association* |
| 03 n (30) | 會, 與會<br>meeting30, group meeting | a formally arranged gathering | *next year the meeting will be in Chicago; the meeting elected a chairperson* |

會 Langs: Cantonese English

Seen Lemmas: 件; 上帝之手; 騰騰震; 覺得; 講波佬; 好彩; 法官; 朝早; 無啦啦;

Figure 1: IMI's "Sentence Tagger" mode, in Cantonese

be further discussed in the section below. There were 658 instances where the concept was not contentful (currently only punctuation is tagged with *x*). And our corpus identified 461 instances of a missing concept in the OMW hierarchy (provided by PWN). This number excludes cases where only a sense was missing from and added to the Cantonese Wordnet – which happened 709 times.

The remainder of Table 1 shows the number of named entities found in the corpus, as well as a small amount of tags under *Other* which are currently being used to capture the use of foreign words within the corpus. Problems surrounding the use of 'foreign words', which are a mix of code-switching and loanwords, will be further discussed in the section below.

Finally, Table 1 also shows that 1,239 distinct concepts were used to tag the 3,729 contentful concepts in the corpus. This is a useful measure to show that there is a considerable semantic overlap between example sentences.

This sense-tagged corpus will be released as part of the Cantonese Wordnet Corpus, which will be released in the Cantonese Wordnet's main Github repository.[4] New senses added to the Cantonese Wordnet will be included in following releases.

| Tag Type | No. of Concepts |
|---|---|
| Cantonese Wordnet | 3,728 |
| Errors in the corpus (*e*) | 196 |
| No need to tag (*x*) | 658 |
| Missing Concepts (*w*) | 461 |
| Named Organization (*org*) | 79 |
| Named Location (*loc*) | 24 |
| Named Person (*per*) | 40 |
| Number (*num*) | 18 |
| Other (*oth*) | 75 |
| Total | 5,279 |
| Distinct Concepts | 1,239 |

Table 1: Summary of Annotation

## 4 Discussion and Future Work

We have encountered several noteworthy issues during the segmentation process: (i) missing concepts in the PWN; (ii) lack of distinction of senses in Cantonese; (iii) separable verbs; (iv) errors in segmentation; and (v) Other

There are many concepts that are unique to Hong Kong culture, which are (understandably) missing in the Princeton WordNet.[5] For example, '籤' cim1 refers to a piece of paper with an arbi-

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

[5] The Cantonese Wordnet is currently built based on Hong Kong Cantonese, though in the future we plan to include also variations in different varieties of Cantonese.

trary fortune prediction written on it, something you receive in a temple by first shaking a cylindrical tube of sticks. Each stick has a unique number and depending on which stick comes out, a different prediction is given. Another example is '利是' lai6 si6, which is a monetary gift given to unmarried people by married people, during Chinese New Year and in other special occasions to anyone (married or otherwise). The same goes for typical Cantonese dishes, such as '乾炒牛河' gon1 caau2 ngau4 ho2. Even though the dish name can be decomposed into smaller meaningful units (i.e., dry-fried-beef-rice noodles), it is not just any dish that stir-fries beef with rice noodles. There is a region-based expectation as to how the dish should look like. Thus, the term is somewhat idiomatic and should be listed. There are also names for common products in Hong Kong which need to be added, e.g., '八達通' baat3 daat6 tung1, of which the official English name is 'Octopus Card' in Hong Kong. It is a reusable stored-value smart card that can be used for all kinds of electronic payment. All these concepts should and will soon be added to the Cantonese Wordnet. As mentioned above, we decided to hold off on adding new concepts for now. This decision was based on the upcoming release 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. We would like the creation of the new concepts to happen already within CILI's context, in order to avoid having to redo this work later.

There are also many concepts which are not culturally/societally bounded, but are unique to the language. For example, '成' sing4 is the equivalent of 10%, a concept that is missing in the PWN. Other more common instances are Cantonese functional elements, such as classifiers, post-verbal particles, sentence-final particles, conjunction, prepositons, etc. The current version of the Cantonese Wordnet already has concepts for 32 post-verbal particles and 41 sortal classifiers, but more are needed.

There are cases where OMW/PWN has a much-finer sense distinction than in Cantonese – e.g., the 3rd person singular pronoun is 佢 keoi5 in Cantonese, which is not specified for gender. It is now mapped three times to the OMW[6]: to 'he/him' (77000046-n), 'she/her' (77000041-n) and 'it' (77000053-n). Another example is '多' do1, which can mean both 'numerous' (01552419-a) and 'much' (01553629-a). In other words, the count/mass distinction is not reflected in the Cantonese '多' do1. As of now, we attempt to keep this semantic distinction by tagging '多' with one of the two synsets, depending on the context. A potential solution to explore in the future is to merge synsets for senses that are not distinguished.

Many verbs in Cantonese contain two parts/characters, and they are separable in the sense that a post-verbal particle can be inserted in-between the two characters. And since the two parts are non-consecutive in the corpus (with a particle in-between), they couldn't easily be tagged as one concept without manually creating a multi-word expression. For example, '跳舞' tiu3 mou5 means 'dance' (or literally 'dance a dance') should probably be mapped to the synset for 'dance' (01894649-v) but, in the corpus, the two characters were separated by the Cantonese perfective particle '咗' (zo2). Our current solution is to tag each of characters by its literal meaning if there is some level of compositionality (even if not very strong). In this case, '跳' is tagged as the verb 'dance' (01894649-v) and '舞' is tagged as the noun 'dance' (00428270-n), functioning like a cognate object. In the future, when we add these multi-word expressions as concepts, we would like to explore keeping the two levels of annotation (with the example '跳舞' tiu3 mou5, it would be mapped as a multi-word expression to the synset of 'dance' as well as decompositonally as 'dance a dance'), since this could end up being useful for future research.

Examples where two or more characters of an idiomatic separable expression could not preserve any of its meaning if tagged literally include '挖角' waat3 gok3, which means 'headhunt'. Literally, the first character means 'dig' and the second character means 'horn'. The meaning of 'headhunt' is idiomatic. In such cases, we have marked both characters as 'errors' (as in the corpus, the two characters are not consecutive and are separated by an aspectual particle) while noting that as a whole it has an idiomatic reading. In the future, we would like to tag these cases as multi-word expressions.

Our corpus also contained some segmentation errors where the already segmented unit should be

[6] These three synsets are not officially part of the PWN, but are introduced by the OMW's pronoun expansion introduced by Seah and Bond (2014)

further segmented. The expression '今次' gam1 ci3 'this time', for example, can be further segmented into '今' (a proximal demonstrative used in classical Chinese but still appears with various nouns bearing the same meaning) and '次', which means 'time' as in 'an instance or single occasion for some event'. Given their frequency, we plan to fix many of these errors semi-automatically.

Another less common error type found in our corpus were orthographic mistakes. These are cases where a wrong character has been used when the corpus was crafted. These will have to be hand-corrected.

One final note worthy of discussion is the fact that Hong Kong Cantonese, in natural speech, contains a lot of English loanwords and instances of code switching. This is easy to understand since Hong Kong was under British rule for more than 150 years and because it still preserves English as one of its official languages. This is also reflected in our corpus (e.g., 'meet 到 target', with '到' dou2 as a post-verbal particle expressing accomplishment or successful completion of an action; the selected segment means 'succeed in meeting the target'). In such cases, the English words are tagged as 'Other', and a comment marks them as foreign words ('FW'). In the future we will need to take a closer look at these cases and decide whether there is enough reason to include some of these words as part of the Cantonese Wordnet (other examples in our corpus include 'boxing', 'sem' as in 'semester', 'app', among many others), or if we should continue to consider them as foreign words. Deciding whether specific cases are instances of loanwords or code-switching will ultimately determine the treatment these words deserve in our project. If deemed as instances of code-switching, words can most probably be either ignored or should be tagged using the a wordnet for the code-switched language (e.g., PWN, for English). However, whenever deemed as loanwords, these words should be considered as an intrinsic part of the Cantonese lexicon, and must be included in the Cantonese Wordnet (e.g., similar to how 'kindergarten' is part of the PWN, even though it is clear from its orthography that it was borrowed from German).

In addition to further researching and addressing the points raised above, we have plans to continue expanding the Cantonese Wordnet corpus by incorporating freely available data useful for educational purposes. Two such projects include Hambaanglaang,[7] a collection of open Cantonese resources created by volunteers and Tatoeba,[8] a multilingual collection of freely available sentences compiled specifically for second language learners. More specifically, we would like to adapt experiments such as the one presented in Bond et al. (2021). In this work, sense tagging is used as a tool to teach lexical semantics, and we believe similar experiments could be set for second language learners – e.g., by inviting learners of Cantonese to tag very basic texts in an attempt to help them recognize multiple senses of individual words.

## 5 Conclusion

This paper presented the Open Cantonese Sense-Tagged Corpus, an ongoing project seeking to improve the Cantonese Wordnet and the digital viability of Cantonese through the creation of a sense-tagged corpus.

The sense tagging process is demanding and yet useful in building linguistic sensitivity to lexical meaning and to discover interesting linguistic phenomena. We hope the work in our corpus will inspire further linguistic research for Cantonese.

In this preliminary experiment, we have tagged more than 5,000 concepts and, with it, we have raised our awareness for some key-issues that must be addressed before proceeding further. We are determined to continue pursuing this project and, with it, also continue to improve the Cantonese Wordnet.

## Acknowledgments

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. We would also like to thank Dennis Ngai Wa Lam for helping with the segmentation work.

## References

Francis Bond and Ryan Foster. 2013. Linking and extending an Open Multilingual Wordnet. In *51st An-*

[7] https://hambaanglaang.hk/
[8] https://tatoeba.org/en/

*nual Meeting of the Association for Computational Linguistics: ACL-2013, Sofia*, pages 1352–1362.

Francis Bond, Andrew Kirkrose Devadason, Rui Lin Melissa Teo, and Luis Morgado Da Costa. 2021. Teaching through tagging —interactive lexical semantics. In *Proceedings of the 11th Global WordNet Conference (GWC 2021)*, Pretoria, South Africa. Global Wordnet Association.

Francis Bond, Luís Morgado da Costa, and Tuan Anh Le. 2015. IMI – A multilingual semantic annotation environment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, System Demonstrations (ACL 2015)*, pages 7–12, Beijing, China.

Francis Bond, Piek Vossen, John Philip McCrae, and Christiane Fellbaum. 2016. Cili: the collaborative interlingual index. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 50–57.

Francis Bond, Shan Wang, Eshley Huini Gao, Hazel Shuwen Mok, and Jeanette Yiwen Tan. 2013. Developing parallel sense-tagged corpora with wordnets. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 149–158.

Huang Chu-Ren, Hsieh Shu-Kai, and Chen Keh-Jiann. 2017. *Mandarin Chinese words and parts of speech: A corpus-based study*. Routledge.

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

Shari Landes, Claudia Leacock, and Randee I Tengi. 1998. Building semantic concordances. *WordNet: An electronic lexical database*, 199(216):199–216.

George A Miller, Claudia Leacock, Randee Tengi, and Ross T Bunker. 1993. A semantic concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Luís Morgado da Costa and Francis Bond. 2015. Omwedit - the integrated open multilingual wordnet editing system. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, System Demonstrations (ACL 2015)*, pages 73–78, Beijing, China.

Luís Morgado da Costa and Francis Bond. 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*.

Jerome L Packard. 2000. *The morphology of Chinese: A linguistic and cognitive approach*. Cambridge University Press.

Tommaso Petrolito and Francis Bond. 2014. A survey of wordnet annotated corpora. In *Proceedings of the Seventh Global WordNet Conference*, pages 236–245.

Yu Jie Seah and Francis Bond. 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*.

Joanna Ut-Seong Sio and Luis Morgado da Costa. 2019. Building the cantonese wordnet. In *Proceedings of the 10th Global Wordnet Conference (GWC 2019), Wroclaw, Poland*.

Joanna Ut-Seong Sio and Luis Morgado da Costa. 2022. Enriching linguistic representation in the cantonese wordnet and building the new cantonese wordnet corpus. In *Proceedings of the 13th Conference on Language Resources and Evaluation*, Marseille, France. European Language Resources Association (ELRA).

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*, page 86–89, Dublin.

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, a Workshop at IJCNLP-2013*, pages 10–18, Nagoya.

Shan Wang and Francis Bond. 2014. Building the sense-tagged multilingual parallel corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).