The Open Cantonese Sense-Tagged Corpus

Joanna Ut-Seong Sio and **Luis Morgado da Costa**

Global Wordnet Conference, 26 Jan 2023

Univerzita Palackého v Olomouci

VRIJE UNIVERSITEIT AMSTERDAM

European Commission

Marie Skłodowska-Curie Actions

Contents

- Intro to the Cantonese Wordnet
- Open Cantonese Sense-Tagged Corpus
  - Goals
  - Methodology
  - Results
- Future Directions

The Cantonese Wordnet

- It’s an **open** project, **started 2019**, and is ongoing (*slow and steady*)
- Mainly 2 people (with occasional paid help when we have money)
- Cantonese is a widely known Chinese regional variant (**mostly spoken**), close to 80 million speakers worldwide
- Cantonese is fairly large yet **under-resourced** language
- The wordnet currently focuses on **Hong Kong** Cantonese (about 7 m. population)
- Our project is **fully hand-checked**, with a focus on some key linguistic features useful for linguistic analysis (*small and high-quality*)

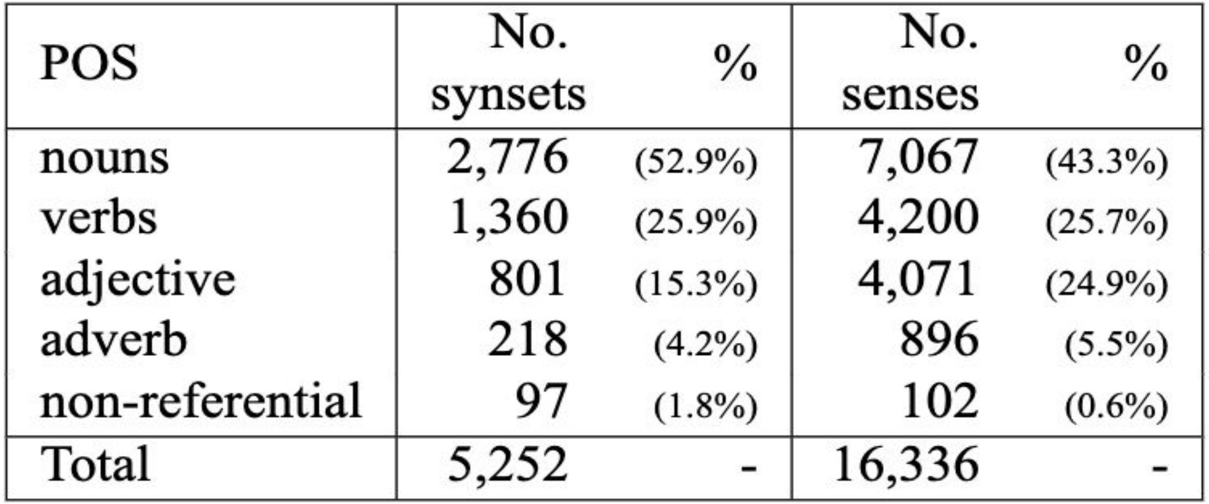
The Cantonese Wordnet

- **STAGE 1 (the first version)**: **characters** (traditional characters); **romanization** (Jyutping, including tonal information), 花, faa1, 'flower'; **expansion approach** using the **Princeton WordNet (PWN) synsets** (provide **definitions**, **semantic hierarchy**, etc.); we try to capture as much **variation** as possible (e.g., multiple characters, lazy pronunciations, etc.) (Sio and Morgado da Costa 2019)

- **STAGE 2**: Initially it was pivoted on the PWN, but is slowly expanding in various directions with added **functional categories** (e.g., classifiers, post-verbal particles) + **example corpus** (Sio and Morgado da Costa 2022)

- **STAGE 3**: Added **sense tagging** (this short paper)

The Cantonese Wordnet



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	-

- The current version of the Cantonese Wordnet contains over 16,000 senses (one sense = character + jyutping, over 32800 forms) distributed over a bit more than 5,000 concepts

Cantonese Wordnet Corpus

Why a new corpus of Cantonese?

- We wanted a corpus to do some types of linguistic research, but:
  - Existing Cantonese corpora are mostly sourced from speech/spoken data ↣ natural (include filler and pauses) but not ideal to extract clean examples
  - Some are made up of texts collected from the Internet, using Cantonese seed words for crawling texts ↣ the text is often a mix of Cantonese and Mandarin
  - Not all are under open license, though some are, e.g., materials in Haambaanglaang.com are all published with a CC-BY licence.
  - Out of the 30 most frequent tokens in the CantoneseWaC (Cantonese Web Corpus), quite a few are strictly non-Cantonese (e.g., 的, 是, 在 and 也)

Cantonese Wordnet Corpus

- **Cantonese Wordnet Corpus**: a corpus of handcrafted examples for individual verb sense, with reference to their compatibility with post-verbal particles

Synset: 00005815-v
Character: 咳

English lemma: cough
Romanization: *kat1*

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

↑

- **3,570** hand-crafted example sentences by two native speakers; begin with sentences with *zo2* (perfective aspectual particle)
    - Each sentence has a sense-tagged verb

Word Segmentation

- Cantonese has characters (e.g., 朱古力 ‘chocolate’); segmentation by space would only give you characters
- one character ≠ one word (e.g., 好食 ‘yummy’ [easy-eat], 朱古力 ‘chocolate’, 解決 ‘to solve’)

↣ A word in Cantonese can be represented by 1 or more characters

- Wordhood in Cantonese is problematic: lack of inflectional morphology, little derivational morphology

↣ Hard to decide what is a word or a phrase (compounds vs. phrases)

- Mandarin Chinese word segmentation tools are not suitable for Cantonese.

Word segmentation: ambiguity

- Structural ambiguity in word-segmentation: homographs
  - **Overlapping (crossing) ambiguity**: ABC, if AB and BC are possible 美国会: 美国 ‘American’+会 ‘will/can’ or 美 ‘American'+ 国会 ‘congress’
  - **Combinatorial ambiguity**: AB, if A, B, and AB 才能: 才能 (noun: “talent”), or as the combination of the 才 (adverb: just now) and 能 (verb: able to)
  - **Mixed type:** ABC, if AB and BC, and A or B or C are possible 太平淡 (too dull), 太平 (peaceful), 平淡 (dull), 太 (too/over), 平 (flat), 淡 (plain) are all possible words,

Word Segmentation

What did we do (Huang, Hsieh and Chen 2017) ?

> **Freedom of parts**: If a character can function alone, it is a word; If at least one of a multi-character unit is bound, then it is a word (e.g., 消防員 'fireman').
>
> **Non-compositionality (semantic/structural)**: If the meaning of a multi-character unit is not compositional, it is a word; If the grammatical category of a multi-character unit is different from what one expects (e.g., 擔心 [carry-heart] 'worry', 好食 [good-eat] 'yummy')

There are many other criteria proposed in the literature (some English examples below, Duanmu 2017), but it is not possible to do all the tests while doing sense-tagging.

New York and New Orlean, *New York and Orlean (**Conjunction reduction**)

He lives in the white house, I live in the green one. (**Lexical integrity**)

??He lives in the **White House**, I live in the green one.

Word Segmentation

一日飲八杯水先啱啱好滿足咗人體需要。

“Drinking 8 glasses of water a day exactly satisfies the need of the human body.”

一 // 日 // 飲// 八// 杯// 水// 先// 啱啱好// 滿足// 咗// 人體// 需要

人體: ‘human body’

人 ‘human’: free

體 ‘body’: bound (cannot be used alone)

啱啱好: ‘just right’, ‘exactly’

啱 ‘match’, ‘correct’: free

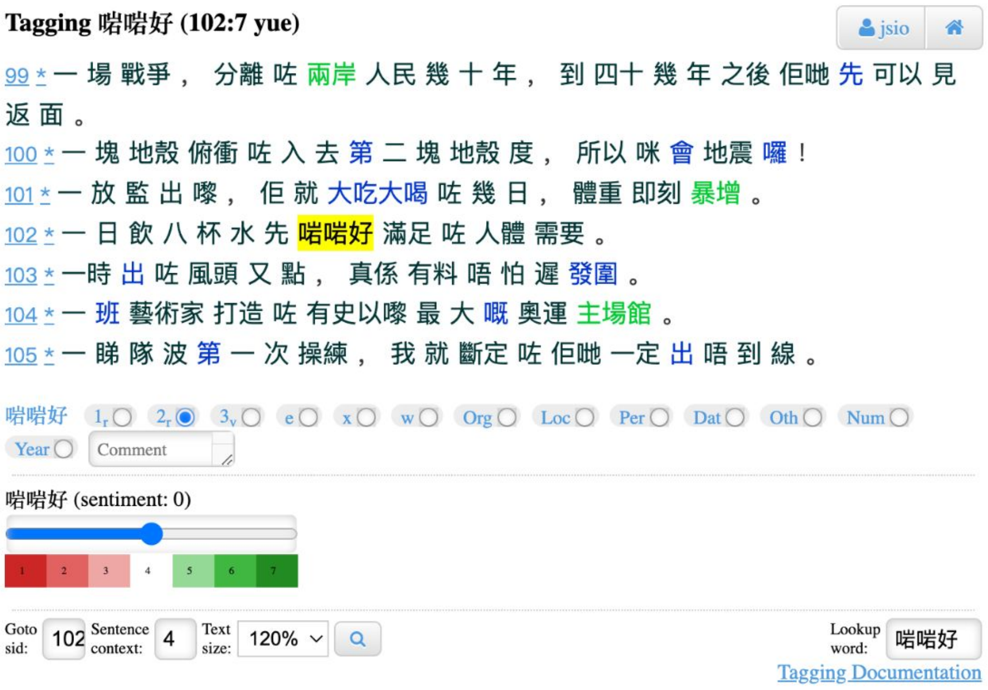
啱啱 ‘exact’: bound

好: ‘good’

Why sense tag?

- There is a **long tradition**, starting with SemCor (10+ languages)
- Helps improve **both the coverage and the precision** of the lexicons being used in the annotation (Miller et al. 1993)
- It helps with a variety of problems/goals:
    - Finding **missing senses**
    - Identifying **indistinguishable definitions**
    - Providing **example sentences**
        - Sense attestation
        - Material for language education
        - Data for WSD and other NLP tasks
- We have only done 300 sentences using **IMI (Bond et al., 2015)**

IMI — A Multilingual Semantic Annotation Environment



Tagging 啱啱好 (102:7 yue)

99 * 一 場 戰爭 ， 分離 咗 兩岸 人民 幾 十 年 ， 到 四十 幾 年 之後 佢哋 先 可以 見 返 面 。

100 * 一 塊 地殼 俯衝 咗 入 去 第 二 塊 地殼 度 ， 所以 咪 會 地震 囉 ！

101 * 一 放 監 出 嚟 ， 佢 就 大吃大喝 咗 幾 日 ， 體重 即刻 暴增 。

102 * 一 日 飲 八 杯 水 先 啱啱好 滿足 咗 人體 需要 。

103 * 一時 出 咗 風頭 又 點 ， 真係 有料 唔 怕 遲 發圍 。

104 * 一 班 藝術家 打造 咗 有史以嚟 最 大 嘅 奧運 主場館 。

105 * 一 睇 隊 波 第 一 次 操練 ， 我 就 斷定 咗 佢哋 一定 出 唔 到 線 。

啱啱好 1_r ○ 2_r ● 3_v ○ e ○ x ○ w ○ Org ○ Loc ○ Per ○ Dat ○ Oth ○ Num ○ Year ○ Comment

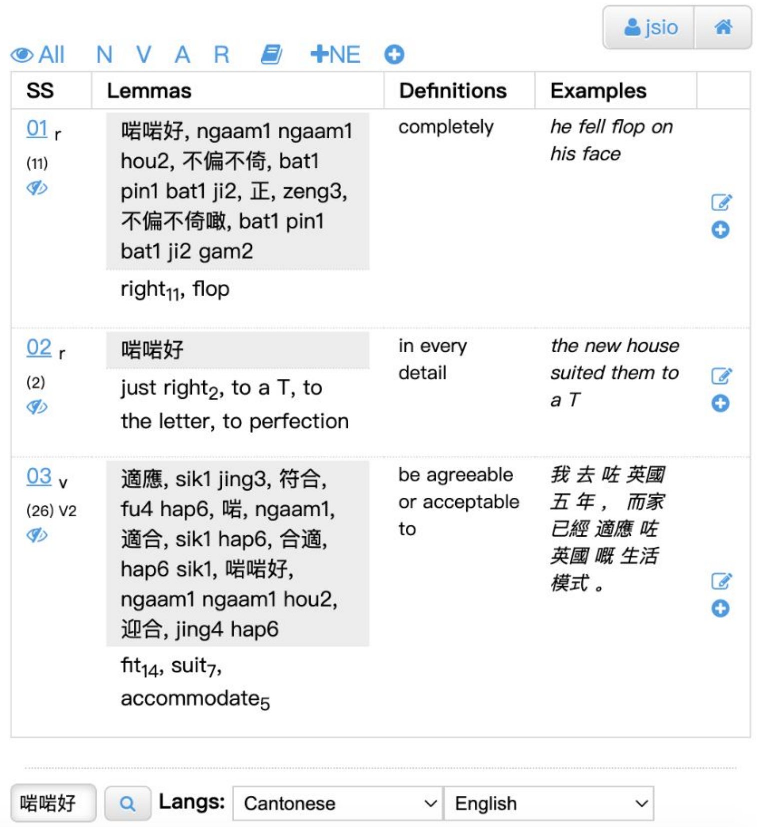
啱啱好 (sentiment: 0)

Goto sid: 102 Sentence context: 4 Text size: 120%

Lookup word: 啱啱好

Tagging Documentation

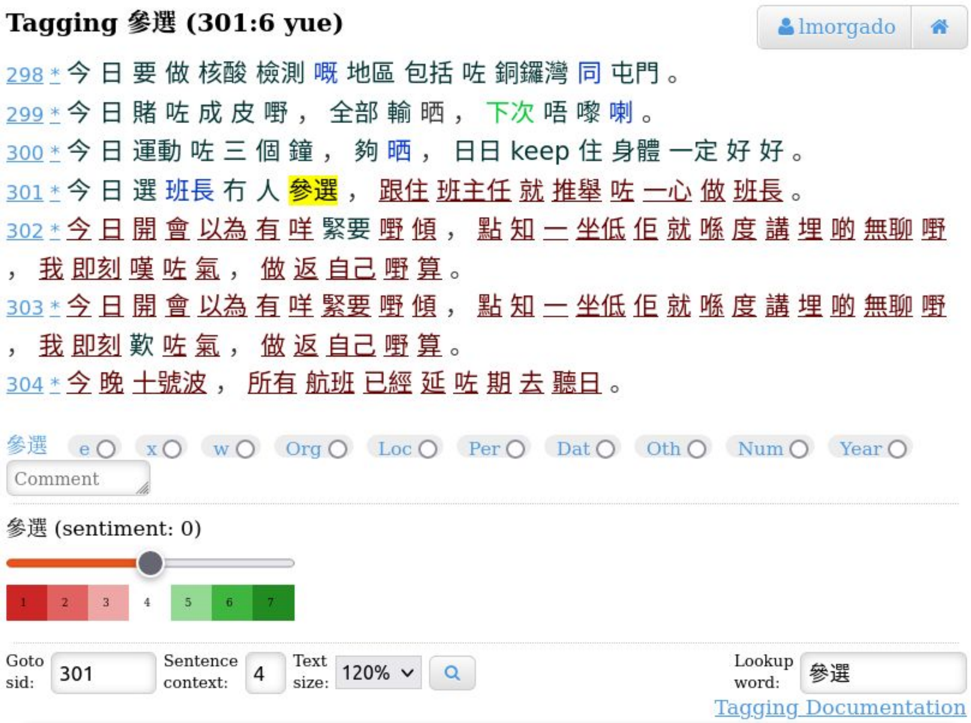
Senses **not** ordered by frequency



SS	Lemmas	Definitions	Examples	
01 r (11)	啱啱好, ngaam1 ngaam1 hou2, 不偏不倚, bat1 pin1 bat1 ji2, 正, zeng3, 不偏不倚噉, bat1 pin1 bat1 ji2 gam2 right$_{11}$, flop	completely	*he fell flop on his face*	
02 r (2)	啱啱好 just right$_{2}$, to a T, to the letter, to perfection	in every detail	*the new house suited them to a T*	
03 v (26) V2	適應, sik1 jing3, 符合, fu4 hap6, 啱, ngaam1, 適合, sik1 hap6, 合適, hap6 sik1, 啱啱好, ngaam1 ngaam1 hou2, 迎合, jing4 hap6 fit$_{14}$, suit$_{7}$, accommodate$_{5}$	be agreeable or acceptable to	*我 去 咗 英國 五 年 ， 而家 已經 適應 咗 英國 嘅 生活 模式 。*	

啱啱好 Langs: Cantonese English

IMI — A Multilingual Semantic Annotation Environment



Tagging 參選 (301:6 yue)

298 * 今 日 要 做 核酸 檢測 嘅 地區 包括 咗 銅鑼灣 同 屯門 。

299 * 今 日 賭 咗 成 皮 嘢 ， 全部 輸 晒 ， 下次 唔 嚟 喇 。

300 * 今 日 運動 咗 三 個 鐘 ， 夠 晒 ， 日日 keep 住 身體 一定 好 好 。

301 * 今 日 選 班長 冇 人 參選 ， 跟住 班主任 就 推舉 咗 一心 做 班長 。

302 * 今 日 開 會 以為 有 咩 緊要 嘢 傾 ， 點 知 一 坐低 佢 就 喺 度 講 埋 啲 無聊 嘢 ， 我 即刻 嘆 咗 氣 ， 做 返 自己 嘢 算 。

303 * 今 日 開 會 以為 有 咩 緊要 嘢 傾 ， 點 知 一 坐低 佢 就 喺 度 講 埋 啲 無聊 嘢 ， 我 即刻 歎 咗 氣 ， 做 返 自己 嘢 算 。

304 * 今 晚 十號波 ， 所有 航班 已經 延 咗 期 去 聽日 。

參選 e ○ x ○ w ○ Org ○ Loc ○ Per ○ Dat ○ Oth ○ Num ○ Year ○

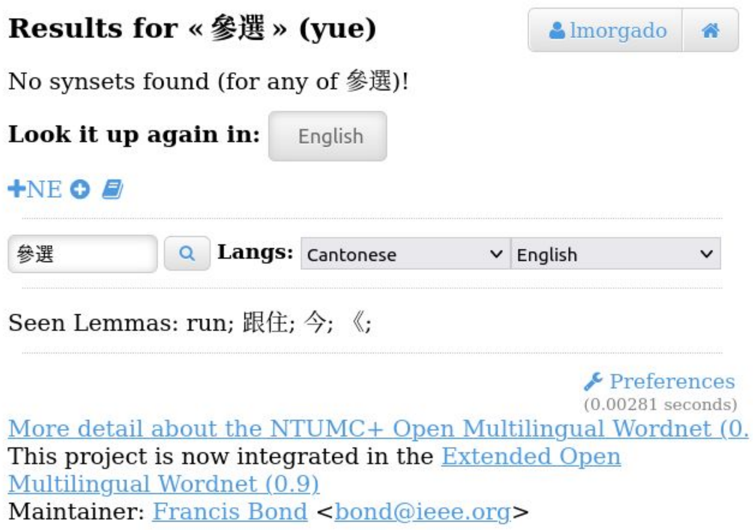
Comment

參選 (sentiment: 0)

Goto sid: 301 Sentence context: 4 Text size: 120%

Lookup word: 參選

Tagging Documentation



Results for « 參選 » (yue)

No synsets found (for any of 參選)!

Look it up again in: English

參選 **Langs:** Cantonese English

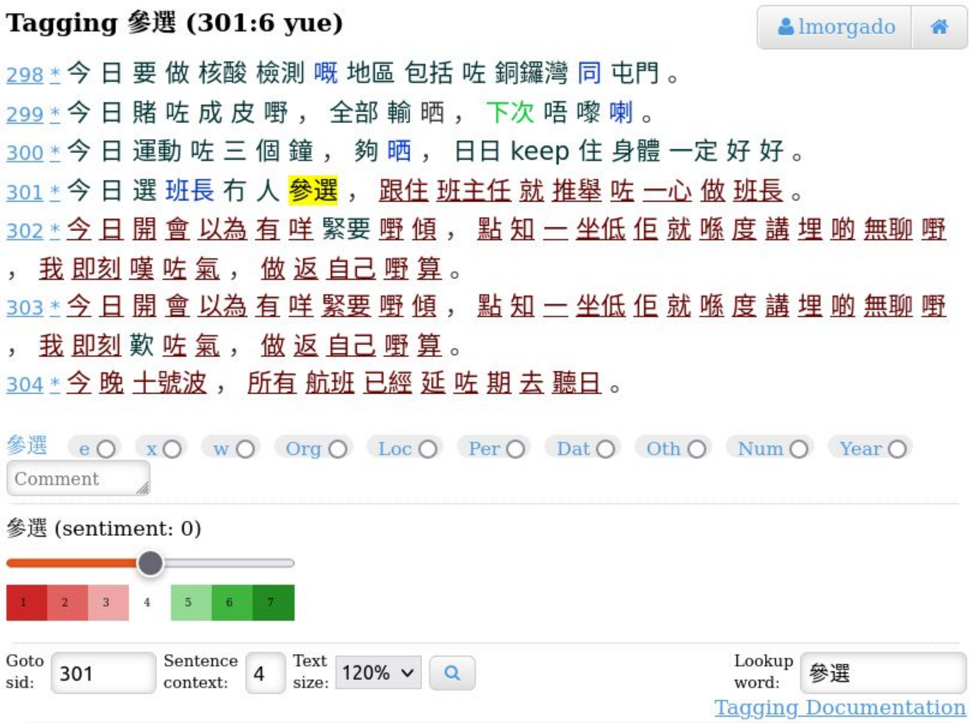
Seen Lemmas: run; 跟住; 今; 《;

Preferences
(0.00281 seconds)

More detail about the NTUMC+ Open Multilingual Wordnet (0.
This project is now integrated in the Extended Open Multilingual Wordnet (0.9)
Maintainer: Francis Bond <bond@ieee.org>

Adding new senses on the fly

IMI — A Multilingual Semantic Annotation Environment



Tagging 參選 (301:6 yue)

298 * 今 日 要 做 核酸 檢測 嘅 地區 包括 咗 銅鑼灣 同 屯門 。
299 * 今 日 賭 咗 成 皮 嘢 ， 全部 輸 晒 ， 下次 唔 嚟 喇 。
300 * 今 日 運動 咗 三 個 鐘 ， 夠 晒 ， 日日 keep 住 身體 一定 好 好 。
301 * 今 日 選 班長 冇 人 參選 ， 跟住 班主任 就 推舉 咗 一心 做 班長 。
302 * 今 日 開 會 以為 有 咩 緊要 嘢 傾 ， 點 知 一 坐低 佢 就 喺 度 講 埋 啲 無聊 嘢 ， 我 即刻 嘆 咗 氣 ， 做 返 自己 嘢 算 。
303 * 今 日 開 會 以為 有 咩 緊要 嘢 傾 ， 點 知 一 坐低 佢 就 喺 度 講 埋 啲 無聊 嘢 ， 我 即刻 歎 咗 氣 ， 做 返 自己 嘢 算 。
304 * 今 晚 十號波 ， 所有 航班 已經 延 咗 期 去 聽日 。

參選 e ○ x ○ w ○ Org ○ Loc ○ Per ○ Dat ○ Oth ○ Num ○ Year ○

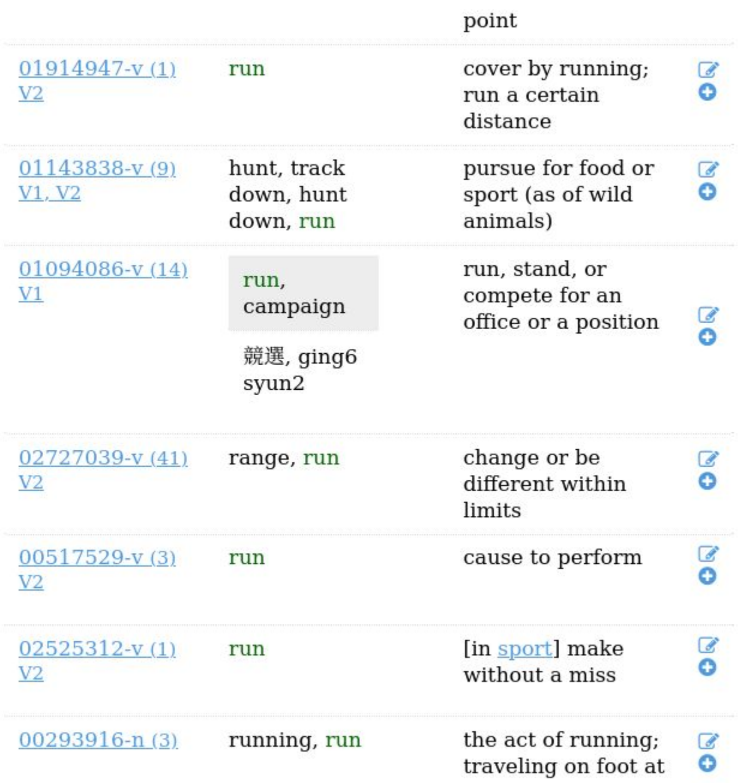
Comment

參選 (sentiment: 0)

Goto sid: 301 Sentence context: 4 Text size: 120%

Lookup word: 參選

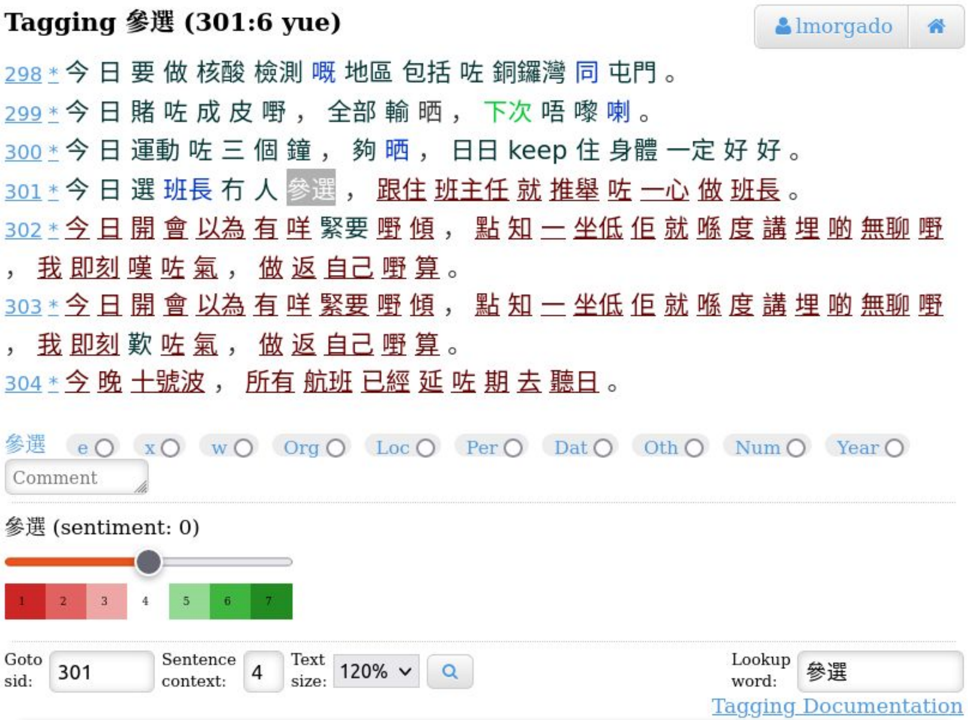
Tagging Documentation



		point	
01914947-v (1) V2	run	cover by running; run a certain distance	
01143838-v (9) V1, V2	hunt, track down, hunt down, run	pursue for food or sport (as of wild animals)	
01094086-v (14) V1	run, campaign 競選, ging6 syun2	run, stand, or compete for an office or a position	
02727039-v (41) V2	range, run	change or be different within limits	
00517529-v (3) V2	run	cause to perform	
02525312-v (1) V2	run	[in sport] make without a miss	
00293916-n (3)	running, run	the act of running; traveling on foot at	

Adding new senses on the fly

IMI — A Multilingual Semantic Annotation Environment



Tagging 參選 (301:6 yue)

298 * 今 日 要 做 核酸 檢測 嘅 地區 包括 咗 銅鑼灣 同 屯門 。

299 * 今 日 賭 咗 成 皮 嘢 ， 全部 輸 晒 ， 下次 唔 嚟 喇 。

300 * 今 日 運動 咗 三 個 鐘 ， 夠 晒 ， 日日 keep 住 身體 一定 好 好 。

301 * 今 日 選 班長 冇 人 參選 ， 跟住 班主任 就 推舉 咗 一心 做 班長 。

302 * 今 日 開 會 以為 有 咩 緊要 嘢 傾 ， 點 知 一 坐低 佢 就 喺 度 講 埋 啲 無聊 嘢 ， 我 即刻 嘆 咗 氣 ， 做 返 自己 嘢 算 。

303 * 今 日 開 會 以為 有 咩 緊要 嘢 傾 ， 點 知 一 坐低 佢 就 喺 度 講 埋 啲 無聊 嘢 ， 我 即刻 歎 咗 氣 ， 做 返 自己 嘢 算 。

304 * 今 晚 十號波 ， 所有 航班 已經 延 咗 期 去 聽日 。

參選 e ○ x ○ w ○ Org ○ Loc ○ Per ○ Dat ○ Oth ○ Num ○ Year ○

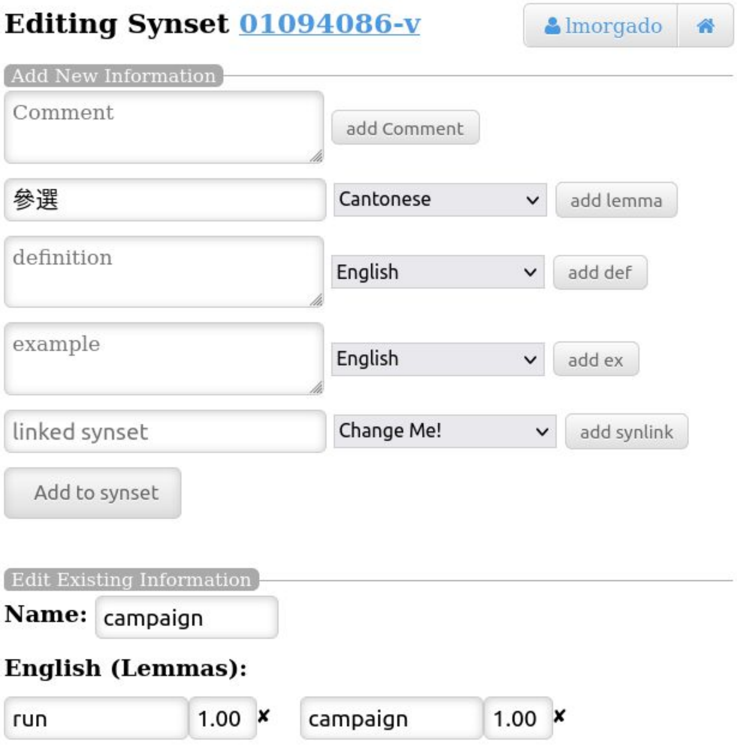
Comment

參選 (sentiment: 0)

1	2	3	4	5	6	7

Goto sid: 301 Sentence context: 4 Text size: 120% Lookup word: 參選

Tagging Documentation



Editing Synset 01094086-v

Add New Information

Comment — add Comment

參選 — Cantonese — add lemma

definition — English — add def

example — English — add ex

linked synset — Change Me! — add synlink

Add to synset

Edit Existing Information

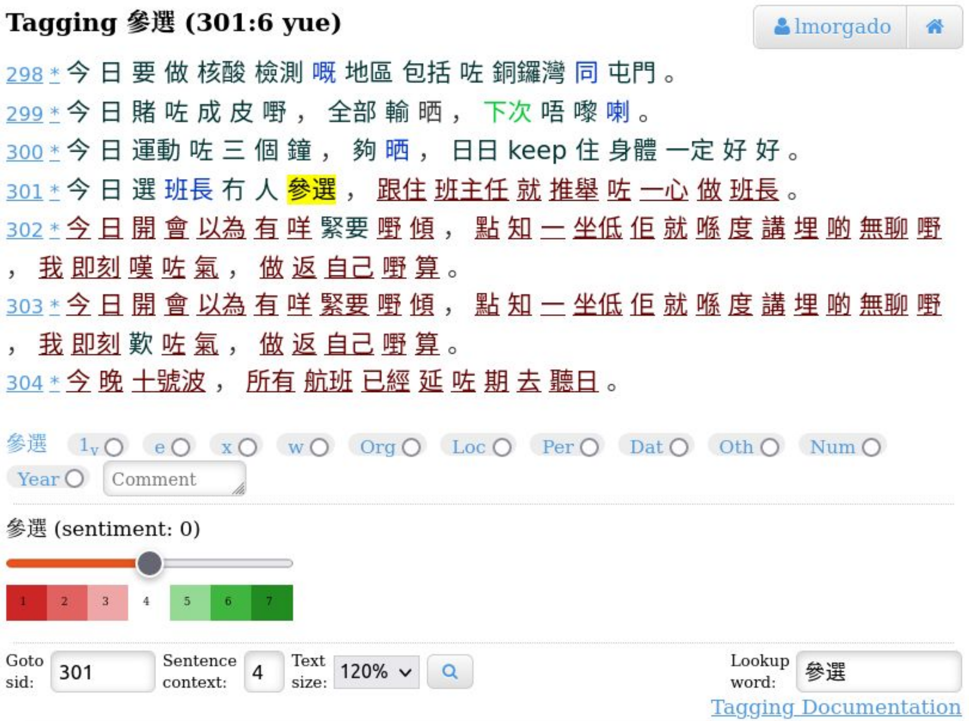
Name: campaign

English (Lemmas):

run 1.00 ✗ campaign 1.00 ✗

Adding new senses on the fly

IMI — A Multilingual Semantic Annotation Environment



Tagging 參選 (301:6 yue)

298 * 今 日 要 做 核酸 檢測 嘅 地區 包括 咗 銅鑼灣 同 屯門 。

299 * 今 日 賭 咗 成 皮 嘢 ， 全部 輸 晒 ， 下次 唔 嚟 喇 。

300 * 今 日 運動 咗 三 個 鐘 ， 夠 晒 ， 日日 keep 住 身體 一定 好 好 。

301 * 今 日 選 班長 冇 人 參選 ， 跟住 班主任 就 推舉 咗 一心 做 班長 。

302 * 今 日 開 會 以為 有 咩 緊要 嘢 傾 ， 點 知 一 坐低 佢 就 喺 度 講 埋 啲 無聊 嘢 ， 我 即刻 嘆 咗 氣 ， 做 返 自己 嘢 算 。

303 * 今 日 開 會 以為 有 咩 緊要 嘢 傾 ， 點 知 一 坐低 佢 就 喺 度 講 埋 啲 無聊 嘢 ， 我 即刻 歎 咗 氣 ， 做 返 自己 嘢 算 。

304 * 今 晚 十號波 ， 所有 航班 已經 延 咗 期 去 聽日 。

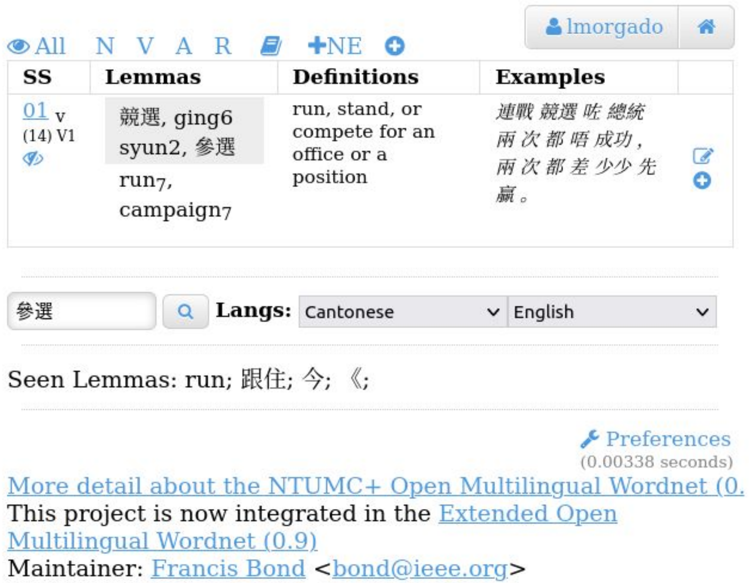
參選 1_v ○ e ○ x ○ w ○ Org ○ Loc ○ Per ○ Dat ○ Oth ○ Num ○ Year ○ Comment

參選 (sentiment: 0)

1 2 3 4 5 6 7

Goto sid: 301 Sentence context: 4 Text size: 120% Lookup word: 參選

Tagging Documentation



SS	Lemmas	Definitions	Examples	
01 v (14) V1	競選, ging6 syun2, 參選 run7, campaign7	run, stand, or compete for an office or a position	*連戰 競選 咗 總統 兩 次 都 唔 成功 , 兩 次 都 差 少少 先 贏 。*	

參選 Langs: Cantonese English

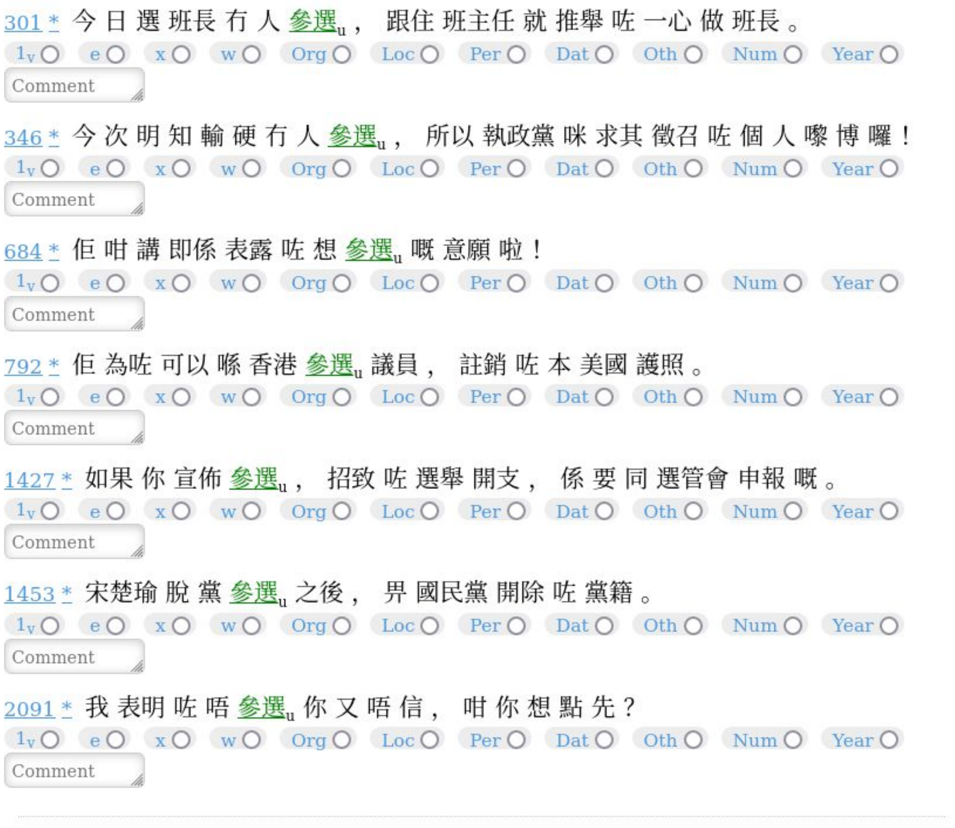
Seen Lemmas: run; 跟住; 今; 《;

Preferences
(0.00338 seconds)

More detail about the NTUMC+ Open Multilingual Wordnet (0.
This project is now integrated in the Extended Open Multilingual Wordnet (0.9)
Maintainer: Francis Bond <bond@ieee.org>

Adding new senses on the fly

IMI — A Multilingual Semantic Annotation Environment



301 * 今 日 選 班長 冇 人 參選$_u$， 跟住 班主任 就 推舉 咗 一心 做 班長 。

1$_v$ ○ e ○ x ○ w ○ Org ○ Loc ○ Per ○ Dat ○ Oth ○ Num ○ Year ○

Comment

346 * 今 次 明 知 輸 硬 冇 人 參選$_u$， 所以 執政黨 咪 求其 徵召 咗 個 人 嚟 博 囉 !

1$_v$ ○ e ○ x ○ w ○ Org ○ Loc ○ Per ○ Dat ○ Oth ○ Num ○ Year ○

Comment

684 * 佢 咁 講 即係 表露 咗 想 參選$_u$ 嘅 意願 啦 !

1$_v$ ○ e ○ x ○ w ○ Org ○ Loc ○ Per ○ Dat ○ Oth ○ Num ○ Year ○

Comment

792 * 佢 為咗 可以 喺 香港 參選$_u$ 議員， 註銷 咗 本 美國 護照 。

1$_v$ ○ e ○ x ○ w ○ Org ○ Loc ○ Per ○ Dat ○ Oth ○ Num ○ Year ○

Comment

1427 * 如果 你 宣佈 參選$_u$， 招致 咗 選舉 開支， 係 要 同 選管會 申報 嘅 。

1$_v$ ○ e ○ x ○ w ○ Org ○ Loc ○ Per ○ Dat ○ Oth ○ Num ○ Year ○

Comment

1453 * 宋楚瑜 脫 黨 參選$_u$ 之後， 畀 國民黨 開除 咗 黨籍 。

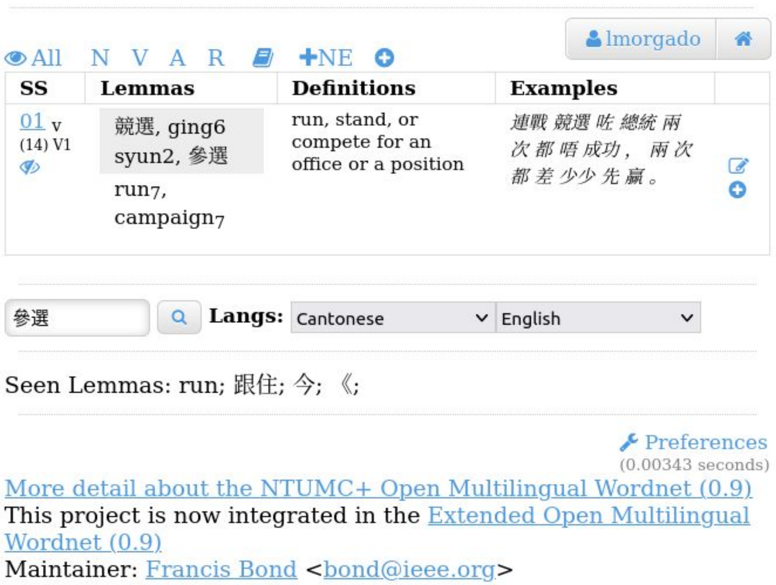
1$_v$ ○ e ○ x ○ w ○ Org ○ Loc ○ Per ○ Dat ○ Oth ○ Num ○ Year ○

Comment

2091 * 我 表明 咗 唔 參選$_u$ 你 又 唔 信， 咁 你 想 點 先 ?

1$_v$ ○ e ○ x ○ w ○ Org ○ Loc ○ Per ○ Dat ○ Oth ○ Num ○ Year ○

Comment



SS	Lemmas	Definitions	Examples
01 v (14) V1	競選, ging6 syun2, 參選 run$_7$, campaign$_7$	run, stand, or compete for an office or a position	連戰 競選 咗 總統 兩 次 都 唔 成功， 兩 次 都 差 少少 先 贏 。

參選 Langs: Cantonese English

Seen Lemmas: run; 跟住; 今; 《;

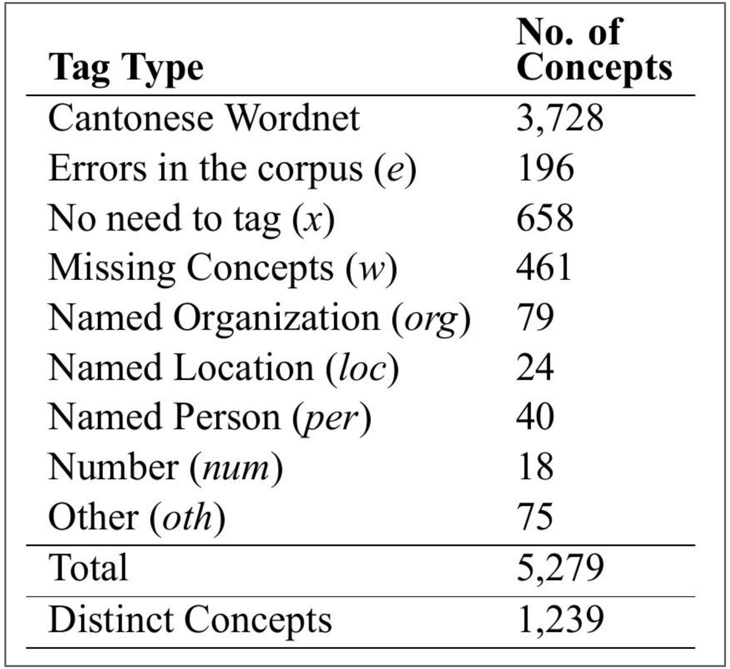
Preferences
(0.00343 seconds)

More detail about the NTUMC+ Open Multilingual Wordnet (0.9)
This project is now integrated in the Extended Open Multilingual Wordnet (0.9)
Maintainer: Francis Bond <bond@ieee.org>

Tagging all instances of a lemma is possible – future work

Summary of Results

- 300 sentences
- 5,279 candidate senses
  - 3,728 senses linked to the CantoneseWN
  - 1,239 distinct concepts
    - 709 new senses
- 162 Named Entities/Numbers
- 461 instances of missing concepts
- 196 instances of segmentation problems/MWEs
- 75 *Other*/FW (in 300 sentences)



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

Issues

- **Missing concepts**: 籤 ‘fortune telling stick’, 利是 ‘red pocket’, 成 sing4 10%
- **Different levels of specification**: 多 ‘much’ or ‘many’; ‘佢’ 3sG ‘he/she/it’
- **Separable verbs**:

  跳舞 ‘dance a dance/dance’ + 咗: perfective marker ➳ 跳咗舞 (compositional)

  挖角 ‘headhunt’ + 咗: perfective marker ➳ 挖咗角 (non-compositional)
- **Mixed code**: ‘Meet 到 target’; 到: succeed in reaching the destination; O 唔 OK ‘is it okay or not?’
- **Errors in segmentation**: ‘今日’ today ➳ ‘今’ + ‘日’ this + day

  (evidence: 今個月 ‘this month’, [this-classifier-month])

Future Work

- Multiple suitable senses for a single concept
  - Allowing ambiguity
  - Finding coarser senses
  - This requires some changes in the annotation tools (allow tagging multiple senses at least temporarily)
- Expand the corpus with other genres
  - We are working with *Hambaanglaang* and *Tatoeba* (graded educational materials)
- Add missing Cantonese-specific concepts (waiting for CILI)

Future Work

- Collaboration with colleagues in Hong Kong (Hong Kong Education University, the Chinese University of Hong Kong) and Vancouver (University of British Columbia)
    - Crowdsourcing data
    - Focus on Cantonese Education

- Use sense-tagging in Cantonese language classrooms (Bond et al., 2021)

Releasing our Data

- The Cantonese Wordnet, Example Corpus, and Sense Annotation will all be released under a Creative Commons Attribution 4.0 International License (CC BY).
- You will be able to download this data from its Github repository:

https://github.com/lmorgadodacosta/CantoneseWN

  - Wordnet LMF (OMW 2.0)
  - TSV (OMW 1.0)

Contacts:

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

Luis Morgado da Costa lmorgado.dacosta@gmail.com

Selected References

- Sio, Joanna Ut-Seong and Morgado da Costa, Luis (2022). Enriching Linguistic Representation in the Cantonese Wordnet and Building the New Cantonese Wordnet Corpus. Proceedings of the 13th Conference on Language Resources and Evaluation. European Language Resources Association (ELRA). Marseille, France.
- Sio, Joanna Ut-Seong and Morgado Da Costa, Luis (2019). Building the Cantonese Wordnet. Proceedings of the 10th Global WordNet Conference (GWC 2019). Wroclaw, Poland.
- 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.
- 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.

Acknowledgements

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 'Chinese Intelligent Language Learning (**CHILL**): Using Grammar Engineering Models for Automated Error Detection' –No.101028782.

thank you