Proceedings of the 1st Workshop on Dataset Creation for Lower-Resourced Languages (DCLRL) @LREC2022, pages 53–62
Marseille, 24 June 2022

Words.hk: a Comprehensive Cantonese Dictionary Dataset with Definitions, Translations and Transliterated Examples

Chaak Ming Lau[*], **Grace Wing-yan Chan**[*]
Raymond Ka-wai Tse[†], **Lilian Suet-ying Chan**[‡]
[*]The Education University of Hong Kong, 10 Lo Ping Road, Tai Po, Hong Kong
[†]The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong
[‡]Words.hk
{lchaakming, cwyan}@eduhk.hk, kwtseab@connect.ust.hk, info@words.hk

Abstract

This paper discusses the compilation of the *words.hk* Cantonese dictionary dataset, which was compiled through manual annotation over a period of 7 years. Cantonese is a low-resource language with limited tagged or manually checked resources, especially at the sentential level, and this dataset is an attempt to fill the gap. The dataset contains over 53,000 entries of Cantonese words, which comes with basic lexical information (Jyutping phonemic transcription, part-of-speech tags, usage tags), manually crafted definitions in Written Cantonese, English translations, and Cantonese examples with English translation and Jyutping transliterations. Special attention has been paid to handle character variants, so that unintended "character errors"(equivalent to typos in phonemic writing systems) are filtered out, and intra-speaker variants are handled. Fine details on word segmentation, character variant handling, definition crafting will be discussed. The dataset can be used in a wide range of natural language processing tasks, such as word segmentation, construction of semantic web and training of models for Cantonese transliteration.

Keywords: Cantonese dictionary, diglossia, corpora, Jyutping, parts of speech, word segmentation, character variants, semantic web, crowdsourcing

1. Introduction

This paper discusses the compilation and properties of a Cantonese dictionary dataset which includes basic lexical information, Cantonese and English definitions and transliterated examples.

Cantonese (ISO-639-3: yue) is a linguistic variety spoken in Hong Kong, Macau, Guangzhou, and several cities or towns with Cantonese immigrants over the past centuries. As a member of Yue dialect group of the Sinitic language(s), Cantonese serves as the lingua franca of Hong Kong in both formal and informal settings across the city (Lai, 2013; Bacon-Shone et al., 2015), and is used as one of the medium of instruction in the formal education system.

Cantonese will be of interest to the language resource community due to its status a low-resource language and at the same time the most resourceful non-Mandarin Chinese language. Cantonese is often considered a dialect in the education system despite its dominant usage, and falls under a diglossic division of labor with a localized version of Standard Written Chinese (SWC). The implication of this diglossic situation is that most written resources, including transcripts are customarily translated into SWC by the user at the time of writing, making it extremely difficult to obtain Cantonese data. On the other hand, the spoken language has a huge user base and numerous video resources, making it a good starting point to explore resource development for a non-Mandarin Chinese language. Language resource work done for Cantonese has the potential to be transferred to other nearby Chinese linguistic varieties (e.g. Hakka, Teochew, Shanghainese, etc.).

There is a need to obtain not just naturally occurring text from Hong Kong users of Cantonese, for those texts will be a mixture of SWC and Cantonese. The fact that Cantonese is not standardized (although the spoken form was "codified" by the mass media, with a relatively stable phonology and grammar) can partially account for the lack of unmixed Cantonese resources. The paper presents the compilation of a resource that is in "authentic" Written Cantonese (Snow, 2004) that is a faithful representation of the lexical and grammatical aspects of the spoken language.

1.1. Existing Resources

At the time of writing Cantonese remains to be a low-resource language due to the factors discussed above. Lexicons and corpora are available for the language, yet are still relatively scarce when compared to languages with similar populations like Korean, Italian, or Polish.

Lexicons with basic lexical information (*Cifu*, *Rime-Cantonese*, *CyberCan*) are readily available but there are not enough resources at higher linguistic levels, partly due to licensing issues. There is a lack of high quality corpus resources (*HKCanCor* and *CantoMap* are the only available open-access corpora), which makes it difficult to compile

Name	Type	Size	License
Cifu (Lai and Winterstein, 2020)	Lexicon	51,798 words	GPLv3
Rime-Cantonese	Lexicon (Input Method)	185,809 items	CC-BY-4.0
CyberCan (Shen et al., 2021)	Lexicon	133,212 words	CC-BY-4.0
Cantonese WordNet (Sio and da Costa, 2019)	Wordnet	3,500 concepts, 12,000 senses	CC-BY-4.0
HKCanCor (Luke and Wong, 2015)	Corpus	230,000 words	CC-BY-4.0
CantoMap (Winterstein et al., 2020)	Corpus	105,000 words	GPLv3
HKCAC (Leung and Law, 2001)	Corpus	170,000 words	Proprietary
ABC Cantonese-English Comprehensive Dictionary (Bauer, 2020)	Dictionary	15,000 entries	Proprietary
CC-Canto	Dictionary	34,335 words	CC-BY-SA-3.0
CantoDict	Dictionary	60,714 words	Proprietary

Table 1: Selected Cantonese language resources

larger-scale semantic webs or dictionaries.

1.2. The Project

The submitted dataset is a resource developed by the Cantonese dictionary project 粵典 (*words.hk*), founded in 2014 by the first author and a couple of associates (Lau, 2019). It is the only dictionary of comparable scale that contains detailed explanations of Cantonese words in both Cantonese and English, acting as both a monolingual and bilingual dictionary. As of writing, the project contains more than 53,000 dictionary entries, of which more than 11,550 have been thoroughly reviewed and made available to the general public. Other entries are available for research purposes.

The following section explains the compilation process (section 2), the philosophy of word segmentation and entry creation (section 3), linguistic considerations (sections 4, 5, 6), the crafting of new entries (section 7), data format (section 8), licensing issues (section 9) and future work (section 10).

2. Compilation Process

Creating a comprehensive, monolingual dictionary from scratch is a non-trivial task. We adopted a multipronged approach in constructing the dictionary.

2.1. Initial Data

We started with word lists from previous projects, for example a list from an unpublished dictionary prepared by the Department of Linguistics of the University of Hong Kong, Hong Kong Education Bureau *Lexical Items with English Explanations for Fundamental Chinese Learning in Hong Kong Schools* (which contains basic words that learners of SWC are expected to know), and the Dictionary of Cantonese Slang (Hutton and Bolton, 2005), as the initial foundation of *words.hk*. Some data from MoeDict[1] were also imported for editor's reference for words that are shared among SWC and Cantonese.

2.2. Crowd-editing

Words.hk was designed to be a crowd-editing project from the onset, and it fits the description of a crowd-sourcing project summarized by Estellés-Arolas and González-Ladrón-de Guevara (2012): it has a clearly defined goal, participative in nature and internet-based. A wiki-like online system was deployed in 2014 for registered users, most of them volunteers recruited from social media, to contribute to the project.

Since then, about 300 people have made at least one edit, and, as of today, around 20 editors contribute regularly. The cumulative number of revisions made to all dictionary entries totals 150,000. Other than editing existing entries, editors are also granted the right (and encouraged) to create new ones. While no structured process is implemented to tease out new entries, editors are keen to create entries for words they consider sufficiently significant. The search result interface would also prompt editors to create a new entry if the word could not be found in the dictionary. These editor-initiated entries often reflect their diverse background and lived experiences, providing the dictionary with a rich variety of entries, ranging from

[1] MoeDict (萌典), which stands for the Ministry of Education Dictionary, is an online dictionary developed by the open source community in Taiwan.

slangs used by local blue-collar workers, to specialized terms used in medical and scientific fields. This organic process allows the project to capture words that would often be passed over by a dictionary produced in a strictly academic setting that relies on written records.

New editors are trained under an informal system akin to apprenticeship. New editors typically start off with editing existing entries; their first few edits will be reviewed by a more senior editor and feedback would be given. Editors usually gravitate towards working on aspects they felt capable of; some editors focus on Cantonese definitions, some focus on English translations, some on Jyutping romanisation while some other on examples. A specialization system is naturally formed, thereby assuring most of the entries are crafted and cross-checked by multiple editors.

Prior to the ongoing COVID-19 pandemic, we also organized meetups on a regular basis. Editors (especially new ones) were invited to attend and would be assigned to work on particular items (e.g. idiomatic expressions) in small groups. Besides induction of new editors, the meetups allowed editors to discuss issues face to face, complementing the existing online channels. With these activities and communication channels, cohesion, mutual understanding and shared responsibility could be maintained among our team members.

In some years, we also hired university students as summer interns to work on the dictionary. Working under direct supervision of the chief editor, the interns served as a cadre to spearhead specific tasks and set the example for fellow editors with their high standard of work. Many of the interns have continued to contribute to *words.hk* well after the end of their contract, eventually becoming an integral part of the community.[2]

2.3. Web-scraping

The relatively small volume of available Cantonese corpora led us to abandon data driven methods in favor of a more manual approach initially, since most NLP techniques applied to Cantonese have been largely ineffective in helping us achieve our goals. For instance, due to diglossia and the small size of existing corpora, frequency data was not a good indicator for entry prioritization.

That said, we recently started inspecting high frequency bigrams and trigrams compiled from data crawled from Hong Kong online forums, resulting in about 1,000 new entries to our word list. This is labor-intensive work as predicted, since most words had already been included in our existing dataset. Editors must sift through false positives of word combinations (e.g. 屋企喺 *uk1kei2 hai2*, "home is at"), typographic errors and proper names (celebrities or user handles).

2.4. Quality Control

We employed technology when applicable to facilitate our workflows. To reduce personal bias and facilitate crowdsourcing, we adopted multiple interaction approaches. When editing and cross-checking entries, editors can leave comments in the column for internal reference. Editing histories are also preserved and made visible to all editors for record-keeping. In addition, public users can report issues on a form on our website or by email.

To ensure correctness in our published entries, all new entries are unpublished when created, and only a handful of senior editors have the rights to publish an entry. In practice the chief editor publishes the vast majority of entries. The chief editor is assisted by a review system where any editor can mark an entry as reviewed. This gives the chief editor an additional assurance vouching for the accuracy of the entry, and allows the system to prioritize entries that have already been reviewed by others for the chief editor's final review before publication.

3. Segmentation

In earlier Cantonese dictionary or corpus projects, significant time was put into deciding what should and what should not be listed as an entry. It is probably impossible to come up with a wordhood test that is universally accepted by linguists. Sproat et al. (1996) showed that Chinese speakers do not have a consensus on where the word boundaries lie, whereas inter-subject agreement on word segmentation was around 70% prior to training. A similar issue is found in the handling of Cantonese data.

The primary consideration here is whether segmentation will affect lexicographic functions (Bergenholtz and Tarp, 2003). This wordhood issue does not constitute a problem as long as we allow multiple ways to segment a string into words. For instance, 女朋友 (*neoi5pang4jau5*, "girlfriend") can be listed as one entry, and the constituents 女 (*neoi5*, "female") and 朋友 (*pang4jau5*, "friend") will be listed as separate entries. The second part is further divided into 朋 and 友, and recorded as separate morphemes in the dataset. This treatment creates some redundancies in the entries, but avoids the need to make difficult decisions about removal or non-removal of entries, which is essential to foster growth of the dictionary as a crowdsourcing project. Important collocations are included, on the same ground. 打 (*daa2*) is literally "to hit" but it is also semantically vacuous when used in conjunction with 官司 (*gun1si1*, "lawsuit"), 比賽 (*bei2coi3*, "match"), 電話 (*din6waa2*,

[2] The third author was an intern in our early years. He is now the chief editor.

"phone"). These collocations are listed as separate entries despite their phrasal nature. The same morpheme may be split into multiple entries (if it is used in more than one parts-of-speech). Expressions which can be broken down into smaller units will be included if the inclusion of the entry will benefit the learners or other downstream processing tasks.

The dataset is not meant to be a reliable way to measure the size of the Cantonese vocabulary, and the only drawback is slightly higher maintenance cost due to more entries.

4. Orthographic Representation

Words.hk aims to record Cantonese data as used in Hong Kong, which means the Traditional Chinese script (ISO 15924: *Hant*) is the natural choice for the entries and the definitions. We do not insist on having all morphemes assigned a Traditional Chinese character due to impracticality. There is a fair amount of code-mixing in Cantonese speech of speakers at all education levels. The project maintains a pragmatic approach and includes all words of enough significance regardless of etymology. English loanwords are represented as is (or in a localized form) in the Latin alphabet to faithfully represent how Cantonese words are written. Other Cantonese words that are traditionally represented using letters include compounds of a mixed etymology, onomatopoeic words and native words with obscure origins.

4.1. Character Choice

The treatment of language resources for a language without explicit standardization is a matter that calls for thorough documentation. The descriptive nature of the dictionary does not mean that all unconventional written forms will be recorded. In fact, using a purely frequency-based account will rule out well-justified orthographic forms and include a long list of forms that are considered incorrect by most users. The project is descriptive to the level that we try to describe the preferred or perceived-to-be-correct forms (if they exist) of educated users of the language.

The principles we follow include:

1. If a word is of Classical Chinese origin or it is shared with SWC, and there is no dispute in their standardized form as used in Hong Kong, this form will be used.
2. If a word is of English origin, the Standard English spelling (British) will be used.
3. If a word is a native Cantonese word with unsettled etymological dispute, or is onomatopoeic which does not have a conventional written form, forms listed in other paper-based dictionaries and/or attested forms will be included at the discretion of the editorial team.
4. If there are no available Han characters or conventional English spellings for the morpheme, its Jyutping spelling (with the tone numbers removed) will be used as its default representation.

4.2. The Equivalent of "Misspelling" in Cantonese

Online Written Cantonese data is known to contain incorrect (as judged by the majority of native speakers) characters, due to the spontaneous nature of online text. It is common to see homophones being used in the representation of Written Cantonese. Some of these homophones might be more frequent than the perceived correct usage. These "incorrect characters" will generally be excluded from the dataset, although a separate list has been compiled for the purpose of identification of incorrect characters. If the editors judge that the incorrect forms are frequently used and must be included, a separate entry (with the "typo" label) will be created to redirect the users to the main entry with the conventional orthographic form.

4.3. Character Variants

Due to the ideographic nature of Han characters, various localities using Han characters have developed specific preferences in representing essentially the same character. To facilitate such preferences, different Unicode code points were created to represent some of these variants. For example, "溫" (0x6EAB) and "温" (0x6E29) are represented by different code points in Unicode, but in the context of Cantonese users, they both represent the same character.

Without an authority to define the use of traditional Han characters, Cantonese users tend to use variants interchangeably. For example, "温" is taught in schools in Hong Kong as the canonical variant (Lee, 2000); however, "溫" is often used on the Internet due to its existence in older encoding methods such as Big5, which did not include "温". Even government websites in Hong Kong (such as the Hong Kong Observatory) tend to use the Big-5-compatible "溫" character instead of the "correct" character "温".[3]

Given that Cantonese users treat such character variants interchangeably, a Cantonese dictionary must be able to recognize common variants of characters. For straightforward cases, we opted to pick one variant and normalize the characters into that variant. We started by using a mapping from

[3] However, a simple online search reveals that this preference is far from being consistent. Occasionally "温" is used on some government websites.

Type	**Etymology**	**Morpheme**	**Headword**
Cantonese/ SWC-shared	刀	*dou1* “knife”	刀
English	CID	*si1aai1di1* “CID”	CID
English	download	*daang1lou1* “download”	Download / 單撈
English	sorry	*so1* “sorry”	sor / 梳
Mixed	P + 牌	*pi1paai4* “probationary driving license”	P 牌
Mixed	calculator + 機	*keu1gei1* “calculator”	cal 機
Mixed	有 + point	*jau5pon1* “sensible, making good judgements”	有 point
Onomatopoeia	(unknown)	*bi1li1baa1laa1*	嗶哩叭啦 / bi li ba la
Onomatopoeia	(unknown)	*tiu4tiu2fing6*	條條 fing / 條條掬
Disputed	奇離	*ke4le4* “weird”	騎呢 / 奇離
Disputed	(unknown)	*liu1lang1* “uncommon; complicated”	撩 lung / 吖嚨

Table 2: Examples of orthographic representations

OpenCC[4]. We removed most mappings with characters that are structurally different (as opposed to minor variants). We then applied the mapping onto our dictionary database, and checked for characters that are not on the List of Graphemes of Commonly-used Chinese Characters (Lee, 2000). Then, for each of these characters, we either manually added it to the list of accepted characters, or we added an entry to the variants mapping, or we made a conscious decision to leave it as-is as a special case of a rare or unusual character.

In the end, we produced a list of canonical characters, and a mapping of variants to canonical characters. We include these two items in our repository.

Note that, despite variants being usually interchangeable, there are exceptions. Names are one common case. For example, the name of the HSBC bank “滙豐 *wui6fung1*”[5] is not supposed to be written as “匯豐”, in spite of the second form being the conventional character for this morpheme. For a dictionary, the use of proper names in the definition and explanation texts is rare, so we have not handled them explicitly. We would imagine that to process variants in corpora that include proper names, an additional step of identifying proper names would be required to keep them intact.

To avoid over-aggressive normalization, we err on the side of caution. We actually maintain two variant maps, the first we call a “safe” map, so called because we believe the variants can be safely used interchangeably. This safe map mostly contains glyph-level “interchangeable variants” (異寫字), which involve variation in minor stroke display (e.g. 說 vs 説) or configuration (e.g. 啟 vs 啓) rather than differences in structural components (e.g. 惗 vs 諗, 綫 vs 線). We use this “safe” map to automatically normalize variant characters in our database (and in this public dataset).

Other known non-canonical variant characters have been added to a much longer “unsafe” map. These characters should be avoided unless the use is justified, e.g. in proper names or certain combinations. A warning message will be displayed when an editor tries to use one of these characters in an entry, so that they can replace them with the canonical variant.

5. Pronunciation

The dictionary uses the Linguistic Society of Hong Kong Cantonese romanisation scheme (also known as “Jyutping” or “LSHK Jyutping”). This is the most common Cantonese romanisation scheme in education and research contexts, which is also employed in *HKCanCor*, *Cifu*, *Cantomap* and the *Unihan* database. Our system allows all combinations of initials and rhymes listed in LSHK 1993 and its 2018 expanded rhyme set, as well as nucleus-coda combinations that are attested but not officially recognized, e.g. -oem.

Pronunciations for the vast majority of the entries follow the LSHK schema. The small number of loanwords that cannot be represented in traditional Cantonese phonology will be recorded using an augmented version of the original LSHK sys-

[4] The OpenCC Hong Kong variant map is available at `https://github.com/BYVoid/OpenCC/blob/master/data/dictionary/HKVariants.txt`

[5] The Hongkong and Shanghai Banking Corporation, a bank with a major presence in Hong Kong.

tem, and these violating transliterations (e.g. the use of –s in the coda position) will be indicated by manually adding a "!" in front of the pronunciation.

To cater for phonological variation in the population of Hong Kong, certain compromises have been made. The pronunciation listed will be more conservative than actual usage. The mergers of the coda pairs {-n, -ng} and {-t, -k}, the onset pairs {ng-, ø-}, {n-, l-}, the tonal pairs {3, 6}, {2, 5}, {3, 5} have been reported (Fung and Lee, 2019; JyutJyuSi (JJS) Work Group, The Linguistic Society of Hong Kong, 2019), and the onset mergers are almost complete. However, the traditional forms will be listed in the dictionary, since the pre-merger pronunciation is considered the proper pronunciation and is expected in text-to-speech systems. However, the difference between the high-falling and high-level tones and other earlier phonological changes will not be represented in our data.

If there are multiple pronunciations for the same lexical item and are unrelated to recent phonological changes, all of them will be recorded in the entry.

6. Part-of-speech Tagging

The dictionary data is part-of-speech tagged, following the POS system in Tang (2015), and can be roughly mapped to the Universal Dependencies (UD) Cantonese-HK tag-set (Wong et al., 2017). By default, each entry should contain only one part-of-speech, with the exception of the following, which can be listed with multiple parts-of-speech in the same entry[6]:

1. verbal nouns, e.g. 默書 (*mak6syu1*, "dictation"), 尊稱 (*zyun1cing1*, "honorable title"), where the nominal usage is similar to that of a gerund;

2. some "好 (*hou2*)-nominal" constructions in the attribute-head form as adjectives when referring to qualities and as nouns when referring to nominals, e.g. 好人 (*hou2jan4*, "good person; kind, generous"), 好嘢 (*hou2je5*, "good stuff; excellent") and 好朋友 (*hou2pang4jau5*, "good friend; in deep, close friendship");

3. words that can be analyzed as either POS category and the choice is purely theory-driven.

Additional usage-related labels have also been provided in Table 4.

POS	English Translation	UD
名詞	nouns	NOUN
區別詞	distinguishing words	ADJ
數詞	numerals	NUM
量詞	quantifiers	NOUN
代詞	pronouns	PRON, DET
動詞	verbs	VERB
形容詞	adjectives	ADJ
副詞	adverbs	ADV
介詞	prepositions	ADP
連詞	conjunctions	CCONJ, SCONJ
助詞	particles	PART
擬聲詞	onomatopoeia	INTJ
感嘆詞	interjection	INTJ
詞綴	affixes	PART, AUX
語素	morpheme	N/A
語句	expressions	N/A

Table 3: Part-of-speech tags and their corresponding POS tags in UD-Cantonese

7. Definition Crafting

Word entries from the initial data contain only basic information (a written form, Jyutping pronunciation and sometimes reference text from other online resources). The definition needs to be crafted manually by our editors. Instead of preparing templates for all possible entries, our decision was to choose efficiency over consistency. These are some guiding instructions that we give to new editors.

- Is this a common, mid-range or rare word, in terms of perceived frequency in speech?
  - For a *common* word, list out different senses of the word with ample collocations and examples.
  - For a *mid-range* word, explain the word in plain language, and give one or two example sentences.
  - For a *rare* word, explain the word in a way that can describe its precise sense without using any other rare words.
- If it is an abstract concept, how would you explain it to a five year-old child?
- Is your definition too broad or restrictive for the word?

[6] Cases such as the extended usage of onomatopoeia are not considered as exceptions as they may not always share the exact meaning.

Label	English
粗俗	vulgar
俚語	colloquial / slang
爭議	controversial
潮語	meme
專名	common name / proper noun
術語	jargon
舊式	obsolete
香港	hongkong
大陸	mainland
台灣	taiwan
澳門	macau
日本	japan
外來語	loanword
書面語	written
口語	verbal / spoken
錯字	wrong
文言	classical
黃賭毒	nsfw
民間傳說	folk etymology

Table 4: Usage-related labels

Editors will need to decide what plain language and rare words refer to, and there is no need for a predefined controlled vocabulary, since there is not yet sufficient resources to compile one.

Certain categories, e.g. chemical elements, constellations, place names, names of languages and ethnic groups, are crafted based on a template. Entries created before the implementation of a template can be corrected afterwards. It is up to the editors to decide how the entries can be improved, through systematic checking or refining of a chosen categories initiated by individual editors.

8. Data Format

The dataset and other supporting files can be downloaded from this link: *https://github.com/wordshk/data2021*

The CSV with the latest dictionary data comprises of the written form of entries, their pronunciations, explanations and examples. The CSV comes in five columns; the content of each column and a sample entry are shown in Table 5 and Table 6 respectively[7].

Col1	Index
Col2	Orthographic representation & Jyutping
Col3	Entry-data (POS, Label, Synonyms, Antonyms, Explanation, and Examples)
Col4	Character variations
Col5	Review status

Table 5: CSV Columns

Col1	76359
Col2	一般來說:jat1 bun1 loi4 syut3
Col3	(pos: 語句)(label: 書面語)(sim: 一般而言) <explanation> yue: 用嚟引起下文，表示只係睇普遍情況，唔考慮個別例子 eng:in general, in most situations <eg> zho: 一般來說，男生都喜歡漂亮的女孩子。(jat1 bun1 loi4 syut3, naam4 sang1 dou1 hei2 fun1 piu3 loeng6 dik1 neoi5 haai4 zi2.) yue: 一般嚟講，男仔都鍾意靚嘅女仔。(jat1 bun1 lai4 gong2, naam4 zai2 dou1 zung1 ji3 leng3 ge3 neoi5 zai2.) eng:In general, boys like beautiful girls.
Col4	一般來說
Col5	OK

Table 6: A Sample Entry

9. Licensing

In the exploratory phase of this project, we discovered that many institutions and people had attempted to create Cantonese dictionaries before us. Unlike languages that have established lexicography traditions and institutions supporting them, Cantonese dictionary projects have a tendency to become abandoned by their original owners. We suspect one reason is that, before the popularization of modern database technologies, and before the Internet made reference materials easily accessible, compilation of dictionaries from scratch required multiple years of dedication and highly focused attention, which is often beyond the capability of a single person or team.

We therefore made the assumption that even if the project is successful beyond our expectations, it will still benefit from arrangements to ensure the dictionary can continue to be developed even after

[7] The Entry-data (Column 3) can be parsed by an open source tool (`https://crates.io/crates/wordshk_tools`)

the original team has moved on.

Our license[8] is designed to do exactly that. Specifically:

1. Most non-commercial uses are allowed and do not require additional licensing.
2. Most copyright restrictions (including commercial use) expire in 10 years after publication[9]
3. Permission is given by default if the copyright owner does not respond to licensing requests.
4. Fair use and personal use exemptions are unambiguously defined.

We retained the right to license commercial use of the dictionary for two reasons: Firstly, we were funded exclusively by small donations from private individuals. While profit has never been a goal, reserving commercial rights may help sustain the project financially. Secondly, these restrictions discourage “forks” of the dictionary, preventing our work from being adapted to promote ideas that run counter to our tenets, in particular, folk etymology and fringe linguistic theories, which are unfortunately a common phenomenon with regard to Cantonese where there is no official body ready to make authoritative statements on the subject matter.

Note that the particular copy/version of our submitted data to this conference is also licensed under the Creative Commons Non Commercial license (CC-BY-NC 4.0). Although we believe our tailor-made license is superior for our purposes and goals (and we encourage data owners to consider adopting similar ideas into their licensing schemes), we nonetheless include a more commonly understood alternative for this particular version to avoid confusion and to facilitate sharing and collaboration.

10. Conclusion

This paper presents the design and compilation process of the first comprehensive dictionary for Cantonese that provides both Cantonese and English definitions. The project started as a lexicographical endeavor, which was later expanded into a language resource that serves both language teaching and natural language processing purposes.

Immediate use cases include simplistic (longest string matching) word segmentation and training of text-to-speech models with verified pronunciation mapping data. This project can fill the gap of the lack of written materials for the language due to its diglossic tradition by providing manually crafted example sentences for both common and rarer words, as well as Jyutping transcription for sentences. The size of the dictionary and its accompanying language materials have already surpassed existing openly available spoken corpora, and the project team continues to work on expanding the content of the project. Since all definitions are written in Cantonese (as opposed to other resources which normally provide definitions only in SWC or English), the dataset can be used in the construction or expansion of any semantic web projects or knowledge base. Similar techniques can also be applied to minority languages in the vicinity that use Han characters and may be facing similar issues. Future plans for the project include developing labels to store grammatical (morphological composition and syntactic properties) information and the conversion of the current format to follow the TEI Lex-0 standard.

11. Acknowledgments

The team would like to thank all contributors to the project, and all anonymous reviewers for their invaluable comments and suggestions.

12. Bibliographical References

Bacon-Shone, J., Bolton, K. R., and Luke, K. K. (2015). *Language use, proficiency and attitudes in Hong Kong*. Social Sciences Research Centre, the University of Hong Kong, Hong Kong.

Bergenholtz, H. and Tarp, S. (2003). Two opposing theories. On H.E. Wiegand’s recent discovery of lexicographic functions. *HERMES - Journal of Language and Communication in Business*, 16(31):171–196.

Estellés-Arolas, E. and González-Ladrón-de Guevara, F. (2012). Towards an integrated crowdsourcing definition. *Journal of Information Science*, 38(2):189–200.

Fung, R. and Lee, C. (2019). Tone mergers in Hong Kong Cantonese: An asymmetry of production and perception. *The Journal of the Acoustical Society of America*, 146(5):EL424–EL430.

JyutJyuSi (JJS) Work Group, The Linguistic Society of Hong Kong. (2019). The recognition and acceptance of phonological variations in Cantonese and its justification. In *The 25th International Conference on Yue Dialects, Guangdong-Hong Kong-Macao University Alliance for Chinese*.

Lai, M. L. (2013). The linguistics landscape of Hong Kong after the change of sovereignty. *International Journal of Multilingualism*, 10(3):251–272.

Lau, C.-m. (2019). Building Cantonese dictionaries using crowdsourcing strategies: The words.hk project. In Tso A., editor, *Digital Humanities*

[8] `https://words.hk/base/hoifong/`

[9] as opposed to lifetime plus 50 years by default

and New Ways of Teaching, Digital Culture and Humanities, vol. 1. Springer, Singapore.

Lee, H.-m. (2000). *List of graphemes of commonly-used Chinese characters (revised version in 2000).* The Hong Kong Institute of Education, Hong Kong.

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

Sproat, R., Gale, W., Shih, C., and Chang, N. (1996). A stochastic finite-state word-segmentation algorithm for Chinese. *Computational Linguistics*, 22(3).

Tang, S. W. (2015). *Lectures on Cantonese grammar.* Commercial Press, Hong Kong.

13. Language Resource References

Bauer, R. S. (2020). *ABC Cantonese-English Comprehensive Dictionary.* University of Hawaii Press.

CantoDict. (n.d.). *CantoDict.* `http://www.cantonese.sheik.co.uk/dictionary`.

Cantonese Computational Linguistics Infrastructure Development Group (CanCLID). (n.d.). *Rime-Cantonese Input Method.* `https://github.com/rime/rime-cantonese`.

CCCanto. (n.d.). *CCCanto.* `https://cantonese.org`.

Hutton, C. and Bolton, K. (2005). *A dictionary of Cantonese slang: The language of Hong Kong movies, street gangs and city life.* University of Hawaii Press.

Lai, R. and Winterstein, G. (2020). *Cifu: A frequency lexicon of Hong Kong Cantonese.* In Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pp. 3069-3077, 1.0, ISLRN 321-291-722-262-7.

Leung, M. and Law, S. (2001). *HKCAC: The Hong Kong Cantonese Adult Language Corpus.* International Journal of Corpus Linguistics, 6(2): 305-325.

Luke, K. K. and Wong, M. L. (2015). *The Hong Kong Cantonese corpus: design and use.* Journal of Chinese Linguistics, 25(2015), 309-330.

Shen, F., Yu, W., Min, C., Ye, Q., Xia, C., Wang, T., and Wu, Y. (2021). *CyberCan: A new dictionary for Cantonese social media text segmentation.* https://doi.org/10.31235/osf.io/tyjr7.

Sio, J. U.-S. and da Costa, L. M. (2019). *Building the Cantonese Wordnet.* In Proceeding of the 10th Global Wordnet Conference, pp. 206-215, Wroclaw, Poland, July, Global Wordnet Association.

Winterstein, G., Tang, C., and Lai, R. (2020). *CantoMap: a Hong Kong Cantonese MapTask Corpus.* In Proceeding of the 12th Language Resources and Evaluation Conference, pp. 2899-2906, Marseille, European Language Resources Association, ISLRN 167-857-138-471-9.

Wong, T.-s., Gerdes, K., Leung, H., and Lee, J. (2017). *Quantitative comparative syntax on the Cantonese-Mandarin parallel dependency treebank.* In Proceeding of the Fourth International Conference on Dependency Linguistics (Depling 2017), pp. 266-275, Pisa, Italy, September, Linköping University Electronic Press.

A. Appendix

A.1. Data Card

Dataset Name: words.hk Cantonese Dictionary
Dataset Developer: words.hk
Dataset License: CC-BY-NC 4.0
Link to Dataset: `https://github.com/wordshk/data2021`
Project website: `https://words.hk`

A.2. Ethical Considerations and Broader Impact

Words.hk is a dictionary created by the users, for the users of the Cantonese language. We aim to document the usage of Cantonese under a descriptive principle, and reject creation of a rigid dichotomy between "researchers" and "subjects"; all editors are equally empowered to present their experience with the language in this community-based model.

The vast majority of work was completed in a voluntary basis; editors are required to express consent before submitting their work. The consent form is integrated into the editing interface for visibility and written in simple Cantonese; as all of our editors are literate and familiar with the Internet, the consent form is considered sufficient for obtaining informed consent. Also included are clauses requiring the editor to transfer any and all applicable copyright to *words.hk*, in exchange for *words.hk* to release said work under our open data principles (see section 9) to avert copyright-related ambiguities and disputes.

Personal information collected from volunteers is limited to name and e-mail address for registration; editors are welcome to register and participate under a pseudonym. While the edits made by each editor are traceable to their account through the version history and activity log systems, such information will not be released to the public and is not included in the dataset. Contrarily, editors seeking due recognition for their contribution to the project may choose to be publicly acknowledged on the "About Us" page[10]. The displayed name can be customized by the editor entirely separate from the account username to prevent breach of privacy.

[10] `https://words.hk/base/about/`

As the use of real names is not considered necessary for illustrating the typical usage of the words, all sentences included in the dataset as examples are anonymized through replacing names with fictitious ones where applicable. Sentences selected or formed for examples are also carefully considered to avoid propagation of biases, harmful stereotypes and bigotry in general; words that are unavoidably offensive and derogatory in their usage, including slurs or pejoratives, will be specifically labelled as such (see Table 4).

In terms of broader impact, we hope this project can encourage more people to write in their own native language. Prior to the advent of *words.hk*, the use of written Cantonese was for the most part limited to corners of the Internet, and only for informal chatting; existing dictionaries contain only explanations in English or Standard Written Chinese. *words.hk* is the first Cantonese dictionary with explanations provided in both Cantonese and English, which proved the viability of using Cantonese in educational settings and paved the way for widespread use of written Cantonese. As of today, Cantonese is commonly used in a variety of ways, ranging from literature to government publications.

This project could also set an example on how non-expert community members can contribute to a monolingual dictionary. As discussed in section 2.2, editors need not to be proficient in every aspect and capable of crafting an entire entry by themselves before they can contribute to the project. Additionally, we also maintain several communication channels where associative members of the community, who are not directly involved with the editing process for one reason or another, could offer their views and comments. Editors would often raise questions and ask for opinions when they encounter uncertainties in the editing process. In this way, we can involve a greater share of the community beyond those who have the technical skills to work with the online system.