The Construction of a Large-Scale Hong Kong Chinese Lexicon with Multilingual Translations for Chinese-as-an-Additional-Language (CAL) Students

Chaak Ming Lau¹, and Wai Sum Sandy Leung¹

¹ The Education University of Hong Kong, 10 Lo Ping Road, Hong Kong
{lchaakming, sandyleung}@eduhk.hk

Abstract. This paper describes the structure and compilation process of a Hong Kong Chinese Lexicon being developed for a forthcoming Cantonese keyboard. The lexicon contains more than 120,000 headwords including common words and high-frequency named entities, and incorporates information such as Jyutping romanization, indicative word frequency, simple translations in five languages (English, Hindi, Indonesian, Nepali, Urdu), disambiguation, POS information, speech register and cross-register synonyms. The lexicon improves the Chinese learning experience of CAL students in two ways, (a) to introduce and promote the effective usage of input methods and dictionaries for learners that facilitate them to locate the correct choice of word or phrase within a particular context, and (b) to provide additional register-specific data that enables advanced learners to further expand their vocabulary. This lexicon can be used in conjunction with wordnet or LIWC (Linguistic Inquiry and Word Count) for a range of natural language processing tasks.

Keywords: Hong Kong Chinese Lexicon, Chinese as an additional language, Cantonese Jyutping keyboard, Multilingual Translation

1 Introduction

This project is a lexicon development project in response to the increasing needs of Hong Kong’s non-sojourner ethnic minorities to learn Chinese as an additional language. These ethnic minorities often have a relatively lower level of Chinese proficiency but are placed in a mainstream Chinese-medium classroom setting. Developing language resources targeting the Hong Kong population can be more complex than similar tasks targeting the Greater Chinese region due to a long-standing diglossic situation in Hong Kong: Cantonese is the spoken vernacular
dominantly used by the Hong Kong population, but school children are required to learn Standard Written Chinese in their formal education. This form of Chinese is often read out in Cantonese but different from either written Cantonese or vernacular Mandarin [9]. To further complicate the issue, Standard Written Chinese used in Hong Kong is different from that used elsewhere in China or other Chinese-speaking countries because of (a) the use of Traditional Han script in Hong Kong (but not in most part of China), (b) significant differences in lexical items pertaining to societal functions [9,10], (c) strong influence from Cantonese and Classical Chinese usage. In addition, with Cantonese being a low-resource language, the construction of a Cantonese-specific lexical input system is believed to benefit linguistic minorities in accessing the language in a much easier way.

1.1 The Project

The current project started as a commissioned project to create an enhanced Jyutping keyboard for ethnic minority students. It was proposed that a phonemic keyboard with English and minority language prompts will allow the target population, i.e., CAL learners with intermediate (CEFR B1) spoken Cantonese proficiency, to conduct computer-mediated written communication even without requiring perfect Chinese character recognition [5]. This lexicon with multilingual translations was constructed to support the development of this keyboard and possible future expansion of the project.

1.2 Limitations with Existing Datasets

Existing linguistic datasets and resources for Cantonese language input methods are limited, despite the popularity of the language and abundance of raw language data. The creation of input methods lacks suitable centralized lexical datasets, and ongoing efforts to improve Cantonese lexicons is dispersed among several projects.

Apart from general lock-up issue (e.g. Bauer’s ABC Cantonese Dictionary), other problems include: (a) lack of normalization in orthography particularly for words without agreed-upon written forms, (b) conflicting pronunciation data mixed with hyper-correction and obscure derived pronunciation, (c) sparse subjective judgement data, and (d) absence of a word-level map that contain short one-word translation for mobile screen display.

This paper aims to address these limitations found in Cantonese, which may be shared by other lower resource languages, by exploring alternative compilation approaches:
(1) How can a reliable list of words with their canonical written forms and pronunciation be extracted from existing resources;
(2) What kind of manual work is required to provide high-quality translation and semantic annotation to the lexicon; and
(3) What measures can be taken to promote integration of lexicon compilation effort.

The sections below outline compilation procedures (#2), the structure of the lexicon (#3), and the significance of contribution of our work (#4).

2 Compilation

2.1 Data Source

The lexicon was constructed by incorporating several datasets, taken from open-source and governmental resources, such as Cifu [4], the Hambaanglaang graded scale (unpublished), the Hong Kong Chinese Lexical Lists for Primary Learning [1], Rime-Cantonese (https://github.com/rime/rime-cantonese), and words.hk [5][7]. Approximately 500,000 entries (either written forms or character-pronunciation tuples) were aggregated from these sources for processing.

The dataset was then filtered by removing duplicate entries, obvious typos, fictitious entities, and long strings. This elimination process resulted in about 120,000 good entries, which were manually reviewed and edited by linguists and other native speakers. The team followed an iterative approach with ongoing refactoring to improve the quality of the dataset.

The following measures were subsequently implemented to enhance the data accuracy and consistency across entries in this dataset: (a) data representation strategies which include handling glyph variation (2.2), loose word segmentation (2.3), and separate tables for grammatical constructions (2.4); (b) a pipeline for updating named entities (2.5); (c) the reduction of translation inconsistency by using disyllabic words as translation anchors (2.6); and (d) incorporation of subjective frequency and familiarity data (2.7) for ordering or translation decisions.

2.2 Glyph Selection

Glyph choice was carefully considered to tackle written variation due to (a) script variation (traditional vs simplified vs non-Chinese), (b) regional variation (e.g. Hong
Kong vs Taiwan), (c) pre-existing historical variation, and (d) pre-conventionalized Cantonese morphemes.

For script and regional variation, only one normalized form for each word is kept in the dataset. This dataset follows the usage recommended by the Education Bureau [1] and refers to the conventional choice by Hong Kong educated users, if necessary, which should resolve most inconsistencies found in existing lexicons.

The convention from and to this normalization was handled by the OpenCC library, by which all entries in our data are converted into the Hong Kong variant. The resulting written form can be slightly different from actual computer-mediated usage in Hong Kong. For example, the codepoint for the character 溫 (study; warm) is common in digital text from Hong Kong due to the legacy of Big-5 encoding. This character is normalized as 温 in the lexicon, which is the same as how it is hand-written. If there is an actual need to use other regional variants, downstream applications can make use of OpenCC to revert the normalization.

Another source of variation in our data can be attributed to the historical development of orthography that is shared among Chinese varieties. These entries are often free variants without much regional preference. We have chosen to list each of these variations as separate entries and select one written form as the normalized version, which is added to the Normalized field of all variant entries.

Apart from variation in orthography, there is also a list of pre-conventionalized Cantonese morphemes [3], which is written in multiple ways. There are signs of convergence as reflected in forum data [8], but this information cannot be obtained easily. We have normalized these morphemes based on frequency-based metrics and survey data with manual adjustments.

The reason we adhere to a strict elimination of alternative orthographic representations is not solely because of the educational nature of our project. We believe this consistent approach is crucial for other normalization tasks and helps to maintain data integrity.

### 2.3 Word Segmentation

Instead of imposing a strict definition of “words” as the basic unit, this project introduces a more flexible approach by including long polysyllabic entries sourced from external datasets. Such entries may include longer proper names and set phrases, which would otherwise be analysed as “phrases” linguistically, but will be
kept so long as they meet a specified frequency of occurrence. This strategy mitigates the likelihood of mistakenly omitting named entities and collocations that are non-compositional and hence require separate translation.

2.4 Separate Tables for Grammatical Constructions

Certain combinations, such as *numeral expressions* and *conjugated verbs*, are often entered as one chunk and therefore needed in the lexicon, but are highly repetitive in nature. These entries can be derived by rules and then added to the final table through an auto-compilation pipeline. For example, numerals-classifier (e.g. 一本 jat1 bun2 “one book of”), classifier-noun (e.g. 本書 bun2 syu1 “the book”), verb-aspect (e.g. 做緊 zou6 gan2 “do-progressive”) can be generated from the table.

2.5 Pipeline for Named Entities

Unlike conventional dictionaries, the lexicon includes a list of key proper names with the justification of inclusion, which is designed to be updated on a regular basis. The rationale behind this is that the frequency and/or the importance of a word in any language is dependent on its connection to the user’s daily life. It is therefore essential to include words connected to local places, companies, people and events, and other internationally renowned equivalents in our lexicon. The lexicon also compiles a list of Cantonese or Chinese surnames and first names, as recognizing people’s names sets the scene for any kind of communication. Proper names were initially collected manually with predetermined inclusion and exclusion criteria, and this standardization makes future modification or expansion possible, if we were to set up a pipeline for automatic named-entity extraction from news or forums.

2.6 Disyllabic Words as Translation Anchor

During the initial phase of the project, the team was aware that the translation of monosyllables is more time-consuming than other items. The reason is that most characters are often morphemic rather than lexical, and therefore it is difficult to determine the most precise translation. Here the team leverages the property of *elastic word length* in Chinese [2]. Almost all monosyllabic words in Chinese can be replaced by a disyllabic synonym to satisfy prosodic or register requirements, e.g. 唱 coeng3 – 唱歌 coeng3 go1 (sing), 房 fong2 – 房間 fong4 gaan1 (room), 全 cyun4 – 全部 cyun4 bou6 (all). This disyllabic form is often less ambiguous, sometimes with a narrower or more specific sense, and can thus be translated into other
languages more directly. Monosyllables, except for the more frequent ones that already have a readily available one-to-one English translation, were first mapped to a synonymous disyllabic word. This also applies to morphemes, e.g. (i) 菜 ung3 -> 菜菜 ung3 coi3 (water spinach) (ii) 蝶 dip6 -> 蝴蝶 wu6 dip2 (butterfly). This can facilitate the translation process and improve consistency across translation.

2.7 The Use of Subjective Judgement

One challenging aspect of Cantonese lexical data is the difficulty in extracting actual usage from dictionary data, and the former is often not found in existing sources. For instance, consider the character 丁, which has three pronunciations according to our data sources: ding1, zang1, zaang1. While the first pronunciation is common, the latter two are rare onomatopoeic variants that appear only in one poem from the The Book of Poetry. Judgement data was collected to provide extra information as to whether a pronunciation for a character or a word is familiar to the average speaker. This is reflected in the table as a pronunciation rank field. Pronunciations that are unknown to native speakers can be filtered out or given a lower ranking in downstream applications.

3 Structure

The dataset compiled from separate working files is released as a flat table in the structure outlined in Table 1. The dataset can be found at https://github.com/type-duck/lexicon upon release.

Each row of the table contains one word (a pair of orthography and pronunciation), accompanied by fields divided into four categories. Linguistic information columns are normative fields of lexical properties: written form, pronunciation, etc. Other fields are informative fields. Usage information fields are statistical information that can inform ordering of keyboard candidates and other language processing tasks. Relational information is provided to establish essential links between entries. Unlike semantic webs which may provide a large range of lexical semantic relations, the current lexicon deliberately limits the selection to register and normalization relations. Translation fields are multilingual translations of the entries for education purposes.
### Table 1. Field and description of the dataset

<table>
<thead>
<tr>
<th>Type</th>
<th>Field</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linguistic information</td>
<td>Honzi</td>
<td>The word entry in Traditional Han characters (with the exception of lettered words).</td>
</tr>
<tr>
<td></td>
<td>Jyutping</td>
<td>Cantonese pronunciation of the word in LSHK Jyutping romanization. Syllable boundaries are marked by space.</td>
</tr>
<tr>
<td></td>
<td>PronOrder</td>
<td>Order of pronunciations if multiple pronunciations are available for the same Honzi (1 is the highest priority).</td>
</tr>
<tr>
<td></td>
<td>Sandhi</td>
<td>Lexical tone-sandhi requires syllables (usually the final one) to be read in a different tone. Most words are read with the original tone (without tone-sandhi). 1 if the entry is a case of tone-sandhi. 0 if unmarked.</td>
</tr>
<tr>
<td></td>
<td>LitColReading</td>
<td>A small number of items use different pronunciations for literary and colloquial readings. If such contrast exists, use -1 for literary and 1 for colloquial. 0 if otherwise.</td>
</tr>
<tr>
<td></td>
<td>Part of Speech</td>
<td>A highly simplified part-of-speech tagging, mainly for automatic generation of affixed forms.</td>
</tr>
<tr>
<td></td>
<td>Register</td>
<td>Labels for lexical items that can only be used in a restricted register: colloquial words (col), written words (wri), and formal spoken Cantonese words (for).</td>
</tr>
<tr>
<td></td>
<td>Label</td>
<td>Additional labels, such as place, other names or rare.</td>
</tr>
<tr>
<td>Usage information</td>
<td>Freq</td>
<td>Corpus-based frequency.</td>
</tr>
<tr>
<td></td>
<td>Freq2</td>
<td>Subjective frequency collected through psycholinguistic judgment tasks.</td>
</tr>
<tr>
<td>Relational information</td>
<td>Written</td>
<td>Corresponding Written entry(-ies) for colloquial words that are not accepted in Standard Written Chinese.</td>
</tr>
<tr>
<td></td>
<td>Colloquial</td>
<td>Corresponding Colloquial entry(-ies) for words that cannot be used in spoken Cantonese.</td>
</tr>
<tr>
<td></td>
<td>Normalized</td>
<td>The normalized form, if Honzi is deemed non-standard.</td>
</tr>
<tr>
<td>Translation</td>
<td>English</td>
<td>A simple translation of the main sense(s) of the words.</td>
</tr>
<tr>
<td></td>
<td>Disambiguation</td>
<td>Further disambiguation if English is ambiguous.</td>
</tr>
<tr>
<td></td>
<td>Urdu</td>
<td>Urdu translation of the word in Urdu script.</td>
</tr>
<tr>
<td></td>
<td>Nep</td>
<td>Nepali translation of the word in the Devanagari script.</td>
</tr>
<tr>
<td></td>
<td>Hin</td>
<td>Hindi translation of the word in the Devanagari script.</td>
</tr>
<tr>
<td></td>
<td>Ind</td>
<td>Indonesian translation of the word in the Latin script.</td>
</tr>
</tbody>
</table>

4 Discussion

The paper aims to outline the methodology employed to overcome challenges associated with the compilation task of the Cantonese input method lexicon, that
would also enhance future language resource compilation for low-resource languages.

4.1 Written Forms and Pronunciation

Knowing the issues with authoritative dictionaries, the project relies on both linguistic expertise of the authors and speakers’ knowledge. Subjective frequency and familiarity judgements for character-pronunciation pairs were collected to provide additional variables for entry filtering and prioritizing translation tasks. Strict glyph selection is also enforced to control the use of variants while preserving the mapping information through relevant fields.

4.2 High Quality Translation and Semantic Annotation Accuracy

Loose word segmentation [5] keeps variants at different lengths, whereas the anchor strategy ensures that disyllables will be the focus of translation word. Shorter or longer entries can be linked to disyllabic entries to ensure translation consistencies and will be manually translated if necessary. The focus on disyllabic words for Chinese lexicons, motivated by linguistic observations about elastic words, will save unnecessary translation work, and as a result improves consistency across entries.

4.3 Integration across Projects

The purpose of using a proper name database and suffixal auto compilation is to modulize the compilation of Cantonese lexicon. Breaking down these tables allows a subset of the lexicon to be incorporated into other projects as an upstream resource. This will be a smaller step to improve cross-project integration.

4.4 Potential Use Cases of the Lexicon

With the recent advancements made in Large Language Models (LLMs), one might argue against any manual compilation of small lexicons. However, such manual compilations are still essential for smaller varieties that are underrepresented on the internet. The process of compiling these lexicons also facilitates a more precise understanding of language usage, which can be further distinguished by lesser-discussed registers. This information can then be used to validate or supplement existing language resources, particularly for genre-specific data.
The lexicon will also facilitate social sciences research that make use of online language data from Hong Kong or Cantonese-speaking regions. This addition set of translation data allows us to create more accurate mapping of existing vocabularies, such as LIWC (Linguistic Inquiry and Word Count) to Cantonese keywords.

It is hoped that this systematic approach will also shorten the development cycle for resources dedicated to Cantonese resources, such as learners' dictionaries and WordNet, as well as low-resource Chinese varieties that are closer to Cantonese than to Standard Chinese.

4.5 Development of Learning Resources

Moreover, the lexicon provides several language pairs that were specifically created for the Hong Kong context. The subsequent release of the Cantonese input method, which is in fact the reason why the lexicon was originally compiled, helps students of the Chinese as an Additional Language (CAL) program improve their Chinese language proficiency and integrate more effectively in mainstream Hong Kong schools. On a broader vision, we believe that this project can bring a positive impact to ethnic integration. Creating achievable pathways to local language proficiency is essential to our effort to foster Hong Kong as a culturally diverse city.

Acknowledgments. This project is supported by Standing Committee on Language Education and Research of Hong Kong (EDB(LE)/P&R/CL/104). We would like to thank the CanCLID team for their ongoing support.

References

1. Chinese Language Education Section, Curriculum Development Institute, Education Bureau: Hong Kong Chinese Lexical Lists for Primary Learning. HKSAR (2007)