Multi-Tiered Cantonese Word Segmentation

Charles Lam[1], Chaak Ming Lau[2], Jackson L. Lee[3]
[1]University of Leeds, [2]Education University of Hong Kong, [3]Talkbit LLC

Abstract

Word segmentation for Chinese text data is essential for compiling corpora and any other tasks where the notion of "word" is assumed, since Chinese orthography does not have conventional word boundaries as languages such as English do. A perennial issue, however, is that there is no consensus about the definition of "word" in Chinese, which makes word segmentation challenging. Recent work in Chinese word segmentation has begun to embrace the idea of multiple word segmentation possibilities. In a similar spirit, this paper focuses on Cantonese, another major Chinese variety. We propose a linguistically motivated, multi-tiered word segmentation system for Cantonese, and release a Cantonese corpus of 150,000 characters word-segmented by this proposal. Our work will be of interest to researchers whose work involves Cantonese corpus data.

Keywords: Word Segmentation, Cantonese

1. Introduction

Word segmentation is the process of delimiting the boundaries between words in a text.[1] In English, word boundaries are orthographically available as the space in most cases. However, for languages without orthographic word boundaries, word segmentation must be either conducted before linguistic analysis and natural language processing (NLP) tasks that assumes the availability of words, or incorporated as part of joint learning or an end-to-end NLP system. One such language is Cantonese, for which there has been growing interest in language learning and digital applications (e.g., instant messaging and online forums), thereby giving rise to a surge in the written use of Cantonese (Snow, 2004; Bauer, 2018). To facilitate digital processing of Cantonese textual data, it is therefore necessary to include the word segmentation information. As recent work on word segmentation for Mandarin Chinese, a Chinese variety closely related to Cantonese, has begun to embrace the idea of multiple word segmentation possibilities due to different interpretations of "word" for Chinese (Wu 2003; Gong et al., 2017), the present study pursues a similar idea for Cantonese and describes a multi-tiered word segmentation scheme that can flexibly cater to different needs. The contributions of this work are two-fold:

- A multi-tiered word segmentation scheme for Cantonese, organized by word classes;
- A publicly available Cantonese corpus of 150,000 Chinese characters, segmented by the proposed scheme.

2. Literature Review

Defining wordhood for Chinese is challenging, due to the great number of potential factors ranging across syntactic, semantic, phonological, psycholinguistic, sociological, and other ones (Packard 2000). For Cantonese, because it is closely related to Mandarin Chinese, it is natural to use Mandarin-based work as a starting point, and combine it with Cantonese-specific considerations (Hou & Wu 2017). This strategy of word segmentation practically applies to all Cantonese language sources that require some notion of "word", including word-segmented corpora (e.g., Lee et al., 1994; Yip & Matthews, 2007; Chin, 2013; Luke & Wong, 2015; Wong et al., 2017; Winterstein et al., 2020), wordlists (e.g., Shen et al., 2021; Lau et al., 2022), and word segmenters (e.g., Fung & Bigi, 2015; Lee et al., 2022).

An implicit yet important assumption in all previous works that require Cantonese word segmentation is that there is only one way to segment a given text. This assumption results in the situation where different Cantonese language resources make different decisions about word boundaries. Inconsistencies across resources impede further development of resources and tools for linguistic research and natural language processing. The following examples of such inconsistencies are from HKCanCor (Luke & Wong, 2015), the CHILDES Cantonese Lee/Wong/Leung Corpus (Lee et al. 1994, hereafter LWL), and Universal Dependencies Cantonese HK (Wong et al.2017, hereafter UD):

- Common nouns. HKCanCor occasionally splits diminutive and derivational morphemes (e.g., 金魚 gam1jyu4 + 仔 zai2 "goldfish + DIM", 商業 soeng1jip6 + 化 faa3 "commercial + ize"), a treatment not seen in LWL or UD.
- Proper nouns. HKCanCor splits proper names as much as possible, e.g., between a surname and given name (e.g., 黎 lai4 + 明 ming4, "Leon Lai (Hong Kong pop singer)"), whereas LWL and UD do not.
- Verb-object compounds. HKCanCor and UD keep the verb and noun unsegmented from each other (e.g., 食飯 sik6faan6 "eat rice", 唱歌

[1] All authors contributed equally to this work. They are listed alphabetically by last name.

LREC-COLING 2024, pages 11993–12002

coeng3go1 “sing song”), whereas LWL splits them up as two separate words.

- Verbal morphology. HKCanCor and LWL split verbal/adjectival suffixes (e.g., 靚 leng3 + 啲 di1, pretty+er, “prettier”) and A-not-A constructions (e.g., 買 maai5 + 唔 m4 + 買 maai5, buy+not+buy, “buy or not”). UD splits the affixes but keeps the A-not-A constructions as one unit.
- Sentence-final particles (SFPs). HKCanCor splits all SFPs into monosyllables, while LWL and UD do not insert a space if the SFP is not separable, e.g., zi1maa3 (LWL), gu3 喎 (LWL), 咖嘛 (UD).

Our view is that there is not a single correct way to segment a Cantonese text, because word segmentation decisions vary depending on both theoretical and practical factors, and that a given resource can (and should) be used by researchers who may not define “word” in the same way; under this view, for the examples of the verb-object compounds above, both splitting the verb and object and keeping them unsegmented are valid segmentation choices.

Inconsistencies in word segmentation across Cantonese resources also give rise to the following view: Rather than attempting to fixate on any theoretical or philosophical definition of wordhood, word segmentation is about what word *boundaries* are. In section 4, we describe a multi-tiered word segmentation scheme for Cantonese, where the tiers are defined by the types of word boundaries. Recent work on Mandarin Chinese word segmentation has also explored multiple segmentation possibilities, e.g., Wu (2003) and Gong et al. (2017) to handle larger named entities in addition to smaller tokens.

3. Methodology

Devising a word segmentation scheme and preparing a sample corpus segmented by the scheme should go hand-in-hand. This is because it is methodologically impossible to flesh out the details of a word segmentation scheme in a vacuum. Segmenting a corpus provides a crucial opportunity for both testing existing specifics of the scheme as well as covering lesser-known situations or anything that a linguist’s introspection might have missed. Moreover, a segmented corpus serves as a reference for how the word segmentation scheme works in practice. For these reasons, both the word segmentation principles, to be discussed in detail in section 4, as well as the associated corpus with our segmentation are the major outputs of our work.

To produce both the segmentation scheme and corpus, we adopted an iterative process. We examined a corpus with pre-existing word segmentation, re-segmented it according to our segmentation scheme, and refined our scheme as issues arose.

The input to our segmentation process was a subset of 150,000 Chinese characters from the Hong Kong Cantonese Corpus (HKCanCor; Luke & Wong, 2015). This corpus was chosen because, to our knowledge, HKCanCor is the only corpus of adult Cantonese conversational text data that has all data files publicly available and is associated with a permissive license (CC BY) that would not interfere with the flexible use of the data in its raw or derived form in any context.

The goal of this work was to collaboratively create a segmentation scheme. To achieve this, all the data files from HKCanCor were distributed among the co-authors of this paper who have strong linguistic training. Each data file was handled independently by two of the authors. We iterated between working on our individual re-segmentation for the assigned data files and discussing discrepancies in our re-segmentation until all discrepancies were resolved. The HKCanCor data segmented by our scheme is publicly available.[2]

Our workflow had the advantage of involving all researchers in discussing and constructing the segmentation scheme rather than silo-ing any individual, as well as having a pre-determined goal of arriving at a single, principled product for release. An added benefit was that as minor course corrections took place throughout the entire process of segmentation work, it made the corpus annotation work more efficient by avoiding late-stage adjustments that would have been more costly.

For inter-annotator agreement, three annotators individually applied the scheme to previously unused HKCanCor data of 3,200 Chinese characters. The pairwise agreement scores were 97.9%, 97.1%, and 97.6%, with an average of 97.5% indicating general robustness of the proposed segmentation scheme.

4. Segmentation Principles

This section discusses our word segmentation principles and exemplifies them by word categories. (For a general overview of Cantonese grammar, see Matthews & Yip (2011) and Tang (2015).)

4.1 The Tiers

Our implementation of the word segmentation has multiple tiers to accommodate the various ways in which wordhood can be defined. Typographically, each tier employs a different symbol to segment words for the tier in question:

- A space ( ) marks a boundary with a word category change (e.g., from a pronoun to a verb). This is the prototypical, clear-cut cases for segmentation (shown as ␣ in this paper).

[2] https://github.com/jacksonllee/multi-tiered-cantonese-word-segmentation

- A dash (-) is used for a morpheme-level boundary (e.g., from a verb to an aspect marker) or boundaries within a compound (e.g., a verb-noun or noun-noun compound). The choice of the dash symbol is not accidental, as this is how morpheme segmentation is shown in glossing in linguistics.
- A pipe (|) is used within a named entity. A pipe connects together words that are otherwise segmented by a space.

Using the tiers and symbols above, the re-segmentation of HKCanCor can be categorized into merging and splitting, compared to the original segmentation in Luke & Wong (2015). Table 1 shows the cases of the merging changes. "Tokens" below refer specifically to the number of instances where the kind of change was made. "Types" refer to the number of unique examples of the changes.

Changes	Tokens (Types)
␣ → ø (non-compositional) 好␣肉 → 好肉 (good␣meat, "meaty") 郵␣chop → 郵 chop (stamp␣chop, "chop stamp")	620 (71)
␣ → - (bound morphemes) 玩␣咗 → 玩-咗 (play-Perf, "played") 遲␣啲 → 遲-啲 (late-Comp, "later")	7,014 (2,282)
␣ → | (named entities) 李␣麗珊 → 李|麗珊 ("Lee Lai Shan") 公關␣部 → 公關|部 ("Dept. of Public Relations")	191 (90)

Table 1: Merging re-segmentations

The first type of mergers is non-compositional, where the two elements form their idiomatic units that are not strictly compositional (e.g., 好肉 means full of meat, rather than "good meat"). The second type of mergers includes bound morphemes that cannot occur on their own (e.g., perfective 咗 and comparative 啲). The third type includes proper names of people, events and (e.g., Lee Lai Shan) or bound morphemes as part of a name. (e.g., 部 in the second example), which were separated by a space in Luke & Wong (2015).

Changes	Tokens (Types)
ø → ␣ compositional 唔係 → 唔 係 (not be "is not / are not") 排隊 → 排 隊 (line.up queue "to stand in line")	1,466 (251)
ø → - (bound morphemes) 嗰個 → 嗰-個 (that classifier "that (one)") 嗰陣時 → 嗰-陣時 ("that time")	4,259 (601)
␣ → | (named entities) 青馬大橋 → 青馬|大橋 ("Tsing Ma Bridge")	44 (32)

Table 2: Splitting re-segmentations

Table 2 shows the different cases of splitting. Whenever two elements are compositional, they are split by a space. For example, 唔 "not" and 係 "be" are independent of each other, where both can stand alone and should not be analyzed as one word. In the second type, the separated units are seen as independent morphemes (e.g., 嗰 "that" contributes meaning compositionally), so they should be segmented by a dash (-) to indicate the status of bound morpheme. The pipe (|) is added to indicate units within a proper name. For example, 大橋 in 青馬大橋 carries the meaning "big bridge".

The HKCanCor subset used in this paper has 120,532 words with its original word segmentation. After re-segmentation by our proposed scheme, there are 113,111 words with the space as the only delimiter, 113,410 words with both the space and pipe as delimiters, and 125,718 words with both the space and dash as delimiters.

In the following where word segmentation is illustrated, numbered examples are in the style of the Leipzig Glossing Rules, where the tiers are (i) Chinese characters, segmented by the proposed scheme, (ii) Jyutping romanization, (iii) optionally, by-morpheme glossing if the internal structure is of interest, and (iv) idiomatic translation in English. In-line examples are formatted as <Chinese, Jyutping, gloss (optional), translation>.

4.2 Nominals

A nominal string, if considered to be a standalone noun, should be separated by a space, and morphemic boundaries within a word are marked by a dash: <男-朋友(␣嘅)␣屋企, naam4-pang4jau5 (␣ge3)␣uk1kei2, male-friend(␣POSS)␣home, "boyfriend's home">.

The actual treatment is less trivial due to the linguistic nature of nominals, that nouns can be strung together without overt morphological markers (for example, without 嘅 *ge3* in the previous example), and can form larger units (compounds or complex noun phrases) that are somewhat syntactically indistinguishable from a noun. To illustrate this point, consider the English phrase "buttermilk ice cream", which consists of three orthographic words. "Ice cream" is often listed as one entry in the dictionary, and prosodically behaves as if the two syllables were one word. On the other hand, "buttermilk" clearly contains both "butter" and "milk" and one may wonder why they are not written separately like "ice cream". How a space is added in English is a convention built up over time, and may or may not coincide with true lexical or syntactic boundaries. Depending on one's analysis, one can also argue that the same string of nouns may be segmented in multiple correct ways. Below we discuss this issue for Cantonese and provide practical solutions for word segmentation.

4.2.1 Common Nouns

The following discussion can be summarized by a simple rule: "group the twos and break the threes", which is motivated by evidence showing disyllabic tendency in Chinese varieties (Duanmu, 2007; Myers, 2022) and the disyllabic preference in Cantonese loanword truncation for nouns (Luke & Lau, 2008). In

general, disyllabic common nouns are treated as words and should not be further segmented morphologically, e.g., <朋友, pang4jau5, friend, “friend”>, <雪糕, syut3gou1, ice_cream, “ice cream”>.

Speakers familiar with Classical Chinese may be tempted to further segment a disyllabic noun, as the morphological composition is often transparent. The nature of Han characters allows us to trace individual etymons in historical texts, which means almost all non-loan words can be decomposed into monosyllabic etymons. For instance, the word for “friend” consists of the etymons 朋 *pang4* (historically, fellows, colleagues) and 友 *jau5* (historically, comrades). Yet this level of tracing defeats the purpose of word segmentation, as this morphological composition, although historically justified, does not reflect the representation of native speakers’ mental lexicon.[3] The same applies to the word 雪糕 *syut3 gou1* “ice cream”, which literally translates to “snow-cake”. This is considered one noun and will not be further decomposed. The exception to this rule is disyllables that are in a modifier-head structure. If both the modifier and the head are free morphemes, and there is no distortion of meaning after adding 嘅 *ge3* between the two elements, then they should be considered two words, separated by a space, e.g., <勁␣人, ging6␣jan4, capable␣person, “capable person”>, <爛␣梨, laan6␣lei2, rotten␣pear, “rotten pear”>. If the insertion of 嘅 *ge3* distorts the meaning of the phrase, then it should remain unseparated, e.g. <大人, daai6jan4, big.person, “adult” (not a person who is big)>, <靚仔, leng3zai2, pretty.son, “handsome guy” (not a son who is pretty)>.

Within a disyllabic unit, the dash is only used between the noun and a monosyllabic (bound) locative marker. A locative suffix, like all affixes in the scheme, is attached to the root. The boundary between a noun and a locative suffix (邊 *bin1*/*bin6* or 面 *min6*) is marked by a dash. Other monosyllabic locatives, such as 中 *zung1* “middle”, 內 *noi6* “inside”, 上 *soeng6* “above”, 下 haa6 “below”, 間 *gaan1* “between”, 側 *zak1* “side”, etc., although not attested in the HKCanCor data, should be treated in the same way: <左-邊, zo2-bin1/6, left, “left”>, <下-邊, haa6-bin1/6, below, “below”>, <開-邊, hoi1-bin6, outer.side, “outer side”>, <入-面, jap6-min6, inside, “inside”>. In contrast, disyllabic locatives, e.g., 隔籬 *gaak3lei4* “next”, are segmented like other common nouns. Nominals that are longer than two syllables are likely to contain at least one boundary, i.e., a dash or space is needed.

Trisyllabic words are almost always analyzable as a 1+2 or 2+1 structure. The singled-out syllable can either be the head or modifier of the noun, free or bound morpheme, root or affixal. The morphological boundary in a trisyllabic structure is indicated by a dash regardless of the status of this element: <警員, ging2jyun4, police.officer, “police officer”>, <職員, zik1jyun4, work.officer, “staff member”>, <公務-員, gung1mou6-jyun4, public.work-er, “civil servant”>; <特性, dak6sing, special.ness, “properties”>, <共通-性, gung6tung1-sing3, shared-ness, “similarities”>; <男人, naam4jan2, male.person, “man”>, <男-朋友, naam4-pang4jau5, male-friend, “boyfriend”>. This is similar to the treatment of Winterstein et al. (2020), which considers productive derivative suffixes to be a separate unit, but different from HKCanCor, which does not separate them from the head noun.

Longer nouns can almost always be analyzed as a modifier-head or a coordination structure with two or more smaller units. If a nominal phrase has four or more syllables, it is likely that it can be divided into two words, i.e., separated by a space. Exceptions to the break-the-three rule include long transliterations of foreign words which became common nouns: <他媽哥池, taa1maa1go1ci4, “Tamagotchi” (electronic pet, a Japanese loanword)>.

There is also a class of tetrasyllabic nouns that consist of multiple coordinated morphemes, in a 1+1+1+1 or 2+1+1 structure. These nouns are used to denote genericity, and sometimes one or more of the components are bound. Morpheme boundaries should be marked with the dash, e.g., <親-朋-戚-友, can1-pang4-cik1-jau5, parents-friend-relative-comrade, “friends and relatives”>, <家爺-仔-嫲, gaa1je4-zai2-naa2, father(obsolete)-son-woman, “the whole family”>, <粥-粉-麵-飯, zuk1-fan2-min6-faan6, congee-vermicelli-noodle-rice, “staple food”>.

Other nominals which are four or more syllables long should be separated into two or more words: <種族␣歧視, zung2zuk6␣kei4si6, race␣discriminate, “racial discrimination”>, <無名␣小卒, mou4ming4␣siu2zeot1, nameless␣small.pawn, “nameless pawn”>, <不明␣飛行-物, bat1ming4␣fei1hang4-mat6, unknown␣flying-object, “UFO”>.

4.2.2 Proper Nouns

Surnames and given names, regardless of order, should be separated by the pipe. If there are any prefixes or suffixes that attached to the name, the prefix/suffix boundaries should be marked with a dash: <王|菲, wong4|fei1>, <陳-仔, can2-zai2>, <阿-黎|明, aa3-lai4|ming4>.

If a place name is specified at multiple levels, they should be separated by a pipe. Not doing so implies that they represent two or more named entities juxtaposed together: <香港|九龍, Hoeng1gong2|Gau2lung4, HongKong| Kowloon, “Kowloon (of) Hong Kong” (Kowloon is part of Hong Kong)>, <香港␣九龍␣新界, Hoeng1gong2␣Gau2lung4␣San1gaai3, HongKong␣Kowloon␣NewTerritories, “Hong Kong (Island), Kowloon & New Territories” (three places)>.

Cantonese does not use capitalization to mark proper nouns. To facilitate named-entity recognition, if a noun phrase is the name of a person, place,

[3] To claim that a disyllabic nominal is not a syntactic noun, one needs to show that the comprising etymons can either (i) both enter a *Num-CL N* structure and can be analyzed as two juxtaposed NPs in a coordination structure, or (ii) be analyzed as a clear case of Adj + N, and the adjective is a modifier.

organization, etc., existing word boundaries (denoted by a space) will be replaced by the pipe, so that syntactic word boundaries can be preserved, e.g., <香 港 | 大 學 , Hoeng1gong2|daai6hok6, HongKong|university, "The University of Hong Kong">, < 香 港 大 學 | 地 鐵 - 站 , Hoeng1gong2daai6hok6|dei6tit3-zaam6, HongKongUniversity|subway-station, "HKU MTR Station">, <傷殘|運動-會, soeng1caan4|wan6dung6-wui2, disabled|sports-meet, "The Paralympics">, <勁 歌 金 曲 | 頒 獎 | 典 禮 , Ging6go1gam1kuk1|baan1zoeng2|din2lai5, JadeSolidGold (a TV program)|prize-presentation|ceremony, "The Jade Solid Gold prize-presentation ceremony">.

A proper noun may have a complex structure. When a proper noun is embedded in another proper noun, the internal structure of the innermost embedded element is suppressed, as shown in examples above. The institution 香港大學 and the TV program name 勁歌金曲, when used in a standalone manner, will be segmented with a pipe. The full structure of these names would require a tree-like structure with details of morpheme- and word-level segmentations at all hierarchical levels. For simplicity, the pipe is used only in the top-level analysis.

4.2.3 Personal Pronouns

Personal pronouns, which can be used as separate words, are considered separate units in our scheme. The plural forms with the suffix 哋 *dei6* are considered one word: <我哋, ngo5dei6, 1SG.plural, "we">, <你哋, nei5dei6, 2SG.plural, "you">, <佢哋, keoi5dei6, 3SG.plural, "they">, <人哋, jan4dei6, people.plural, "others, other people">.

4.2.4 Boundary Issue

The morphological boundary can be inconsistent with the semantic bracketing of a noun, a situation commonly known as the bracketing paradox. For instance, the English phrase "physical therapist" refers to a person (-ist) who conducts physical therapy. The -ist suffix has semantic scope over both words ([physical therapy]+ist), and the phrase should not be interpreted as a therapist who is physical, as opposed to virtual (physical [therapy+ist]). The exact same case is also found in Cantonese (<物理 治療-師, *mat6lei5 zi6liu4-si1, physical therapy-ist*>). We acknowledge the fact that morphological structure may be a mismatch with higher-level analysis, and we always stick to the morphological level in word segmentation.

4.3 Classifiers and Determiners

Like many East and Southeast Asian languages, Cantonese is a numeral classifier language, where quantities as Numeral + Noun in English (e.g., *three tables*) are expressed by the pattern Numeral + Classifier + Noun: <三-本 書, saam1-bun2 syu1, three-CLF book, "three books">, <兩-個 人, loeng5-go3 jan4, two-CLF person, "two persons">.

For word segmentation, our scheme connects the numeral and classifier with a dash, whereas the classifier and noun are separated by a space: Numeral-Classifier Noun.

[Numeral-Classifier] signals a stronger morphological connection between the numeral and classifier. Observe that a numeral alone without a classifier does not license quantity, as evidenced by nominal ellipsis:

(1) 我 有 三-本 書， 佢 有
ngo5 jau5 saam3-bun2 syu1, keoi5 jau5
1SG have three-CLF book, 3SG have
{ *兩, OK 兩-本 }
{ loeng5, loeng5-bun2 }
{ two, two-CLF }
"I have three books, and he has two (books)."

Unlike Mandarin Chinese, a numeral is not the only element that can occur before a classifier in Cantonese. Other options are pronouns and full-fledged nominal expressions, for which a space is used between the nominal expression (pronoun included) and the classifier:

(2) 我 架 車
ngo5 gaa3 ce1
1SG CLF car
"my car"

(3) 樓下 士多 老闆 架 車
lau4haa6 si6do1 lou5baan2 gaa3 ce1
downstairs store owner CLF car
"the car that belongs to the owner of the store downstairs"

A classifier can be reduplicated to denote "each X". For this productive pattern, the scheme uses a dash between the two copies, e.g., <年-年, nin4-nin4, year-year, "every year">.

Determiners, such as 哩 *ni1* "this/these", 嗰 *go2* "that/those", 另 *ling6* "another", and 某 *mau5* "certain", come before the classifier, with an optional intervening numeral. The boundary between pronominals and classifiers is marked by a dash, e.g., <嗰-本␣書, go2-bun2␣syu1, that-CLF␣book, "that book">, <哩-兩-個␣人, ni1-loeng5-go3 jan4, this-two-CLF␣person, "these two people">.

Quantities for entirety (成 *seng4*, 全 *cyun4*) and halves (半 *bun3*) are treated similarly, which are word-segmented with a dash from the following classifier, e.g.: < 成 - 個 ␣ 人 , seng4-go3␣jan4, whole-CLF␣person, "the whole person">, <半-隻␣雞, bun3-zek3␣gai1, half-CLF␣chicken, "half a chicken">.

4.4 Numerals

No dashes or spaces are used for segmenting numerals, including those that involve 十 *sap6* "ten", 百 *baak3* "hundred", 千 *cin1* "thousand", etc. This treatment is similar to German, where numerals such as 55 and 200 are single words (German *fünfundfünfzig* versus English *fifty-five,* German *zweihundert* versus English *two hundred*): <二十, ji6sap6, twenty, "twenty">, < 一 百 三 十 五 , jat1baak3saam1sap6ng5, one.hundred.thirty.five, "one hundred and thirty-five">. Ranges with 至 *zi3* for "from X to Y" are treated with 至 as a word separated by spaces from the surrounding numerals: <五␣至␣

十-分鐘, ng5␣zi3␣sap6-fan1zung1, five␣to␣ten-minute, “five to ten minutes”>.

4.4.1 Approximation and Uncertainty

For consistency, 幾 *gei2* for expressing approximation is also not segmented, i.e., <幾十, gei2sap6, few.tens, “A few dozens”>, <一百零幾, jat1baak3ling4gei2, one.hundred.zero.few, “a hundred or so”>. However, 唔知幾多 *m4zi1gei2do1* “don’t know how much” for an uncertain quantity requires dashes before and after it: <四百-唔知幾多-十, sei3baak3-m4zi1gei2do1-sap6, four.hundred-don’t.know.how.much-ten, “four hundred and a few dozens”>, <唔知幾多-萬, m4zi1gei2do1-maan6, don’t.know.how.much-ten.thousand, “a few tens of thousands”>.

Approximations expressed by consecutive numerals are treated as if they were a single word: <十二三, sap6ji6saam1, ten.two.three, “twelve or thirteen”>, <一兩-個, jat1loeng5-go3, one.two-CLF, “one or two”>, <兩三千, loeng5saam1cin1, two.three.thousand, “two or three thousand”>. The main reason for not segmenting the numeral sequence like these examples is that the numerical consecutiveness is a strong constraint, e.g., *四六-月 *sei3luk6-jyut6* (four.six-month) is not allowed for the intended meaning “April or June”.

4.4.2 Dates and Times

Time units (e.g., year, month, day) are like classifiers in that they directly combine with numerals. They are therefore word-segmented with a dash between the numeral and themselves: <二零二零-年, ji6ling4ji6ling4-nin4, 2020-year, “the year 2020”>, <六-點␣二十-分, luk6-dim2␣ji6sap6-fan1, six-o’clock␣twenty-minute, “twenty past six”>. These time units can also combine with a directional expression instead of a numeral, e.g., last week, next year. The same word segmentation treatment applies: <舊-年, gau6-nin2, last-year, “last year”>, <上-年, soeng6-nin2, last-year, “last year”>, <下-年, haa6-nin2, next-year, “next year”>, <遲␣三-個 月, ci4␣saam1-go3 jyut6, late␣three-CLF month, “in three months”>.

Weeks are expressed with the numeral following the word for “week” (禮拜 *lai5baai3* and 星期 *sing1kei4*). They form a unit separated by a dash between “week” and the numeral: <禮拜-四, lai5baai3-sei3, week-four, “Thursday”>, <星期-五, sing1kei4-ng5, week-five, “Friday”>.

A following monosyllabic qualifier of a date or time is segmented with a dash: <三-月-中, saam1-jyut6-zung1, three-month-middle, “mid March”>, <九-點-半, gau2-dim2-bun3, nine-o’clock-half, “half past nine”>.

4.4.3 Ordinals, Fractions and Decimals

Ordinals with 第 *dai6* are treated as determiners (section 4.3): <第-一, dai6-jat1, order-one, “first”>, <第-二-隻, dai6-ji6-zek3, order-two-CLF, “second one (something compatible with the classifier 隻, e.g., an animal)”>.

Fractions and decimals in the form of “X 分之 Y” and “X 點 Y”, respectively, are word-segmented with the obligatory parts as shown marked with dashes in-between the numerals: <四-分之-三, sei3-fan6zi1-saam1, four-portion.of-three, “three quarters”>, <三-點-一四, saam1-dim2-jat1sei3, three-point-one.four, “3.14”>, <百-分之-五十, baak3-fan6zi1-ng5sap6, hundred-portion.of-fifty, “50%”>.

4.5 Verbal Elements

4.5.1 Verbal Suffixes

Cantonese has a vast inventory of verbal suffixes, encoding aspects, modality and event structure, such as completion, ability, telicity, and inchoativity. These suffixes typically follow the lexical verbs immediately, and often precede objects or post-verbal modifiers. As such, they are often treated as an integral part of the verbs. This is closely related to whether these suffixes should be segmented as separate morphemes. Aspect markers like perfective 咗 *zo2* are functional categories that are always suffixed to the first syllable in the verb: <買-咗␣一-本␣字典, maai5-zo2␣jat1-bun2␣zi6din2, buy-Perf␣one-CLF␣dictionary, “bought a dictionary”>, <結-咗␣婚, git3-zo2␣fan1, mar-Perf␣ry, “got married”>. Morphemes that receive the same segmentation treatment include experiential 過 *gwo3*, durative 住 *zyu6*, completion 完 *jyun4*, and exhaustive 晒 *saai3*.

The examples below show contrasts between the verbs of sight and search in Cantonese. In English, the difference is encoded with different lexical items, such as *look* vs. *see* and *search* vs. *find*: <睇, tai2, look, “look”>, <睇-到, tai2-dou2, look-reach, “see”>, <搵, wan2, search, “search”>, <搵-到, wan2-dou2, search-reach, “find”>. The suffix 到 *dou2* encodes the telos, i.e., the natural endpoint of the event. It is an independent morpheme, but it is also treated as part of the verbal cluster. We use the two-tiered segmentation, where the morpheme *dou2* is connected with a dash to the verb.

4.5.2 “Have” and “Not Have”

As in many other languages, 有 *jau5* “have” is versatile in Cantonese. The most common uses are the verbal use as “have” for possession and the nominal use as a marker of the existential construction. With an explicit subject of the possession, 有 *jau5* in these sentences is seen as verbal “have”: <即係␣佢␣有␣個␣model␣answer,
zik1hai6␣keoi5␣jau5␣go3␣model␣answer,
that.is␣3SG␣have␣CLF␣model␣answer,
“That means they have a model answer.”>.

In sentences where the noun introduced by 有 *jau5* is the subject of the sentence, these tokens are often judged as existential markers. However, for word segmentation, the existential marker 有 *jau5* is no different from the verbal 有 *jau5* for possession, as the existential marker is like the verb but without an overt subject: <有␣個␣男仔␣frown␣我␣走␣喇,
jau5␣go3␣naam4zai2␣me1␣ngo5␣zau2␣laa1,
EXIST␣CLF␣boy␣carry␣1SG␣leave␣SFP,
“There was a boy who carried me on his back.”>.

The perfective “have” is considered a separate word: <上-年␣書展␣我␣都␣有␣去, soeng6-nin2␣syu1zin2␣ngo5␣dou1␣jau5␣heoi3,
last.year␣book.fair␣1SG␣also␣HAVE␣go, “I visited

the book fair last year, too.”>. However, some other cases show that 有 *jau5* can be sublexical and combine productively with some other elements. For example, 有 *jau5* often forms a cluster with 得 *dak1* to indicate possibility of the event denoted by the following verb. The cluster is therefore connected with a dash to indicate that they form a unit. This also applies to its negative counterpart 冇 *mou5* “not have”, where 有-得 *mou5-dak1* denotes unavailability or impossibility: <先至␣ 有-得␣ 睇␣ 嚟, sin1zi3␣ jau5-dak1␣ tai2␣ gaa3, only.then␣ HAVE-dak␣ watch␣ SFP, “Only then (it) is available to watch.”>, <有-得␣ 睇␣ Sailor_Moon␣ 啊, mou5-dak1␣ tai2␣ Sailor_Moon␣ aa3, not.have-dak␣ watch␣ Sailor_Moon␣ SFP, “(The child) cannot watch the Sailor Moon cartoon.”>

4.6 Adjectives

Similar to English, the most common type of collocation with adjectives is degree modification in Cantonese. Degree modification markers, such as 好 *hou2* “very”, 幾 *gei2* “rather”, 非常 *fei1soeng4* “extraordinarily”, are treated as separate words and segmented by spaces: <好␣悶, hou2␣mun6, very␣ boring/bored, “very boring/bored”>, <非常␣好, fei1soeng4␣ hou2, extraordinarily␣ good, “great”>.

Some adjectives are compositionally formed by an adjective and a verb: <易-搵, ji6-wan2, easy-search, “easy to find”>, <易-入口, ji6-jap6hau2, easy-put_in_mouth, “palatable”>.

In some cases, the same adjective-verb combination may provide a different meaning that is opaque and non-compositional. These adjectives are segmented as their own units, and there is no space or dash between them: <好食, hou2sik6, good.eat, “delicious” (not: good / healthy to consume)>, <難食, naan4sik6, difficult.eat, “unpalatable” (not: difficult to eat)>.

In practice, these cases can often be disambiguated by annotators in context. Similar to verbal predicates, adjectives co-occur with suffixes denoting comparison and change of states. For example, the suffix 啲 *di1* is clustered with the adjective by a dash. The term 遲-啲 *ci4-di1* “later” is a frequent combination that can also act as an adverbial, although we have analyzed it in a similar way to other adjective-suffix combinations for consistency. Other suffixes like aspect marker 咗 *zo2* (denoting change of states) and comparative 過 *gwo3*, are analyzed the same way. <買␣多-啲␣書, maai5␣ do1-di1␣ syu1, buy␣ much-Comp␣ book, “buy more books”>, <凍-咗, dung3-zo2, cold-Perf, “(become) colder”>.

For adjective reduplication, a dash is used to link the reduplicated adjectives and the diminutive marker 哋 *dei2*: <怪-怪-哋, gwaai3-gwaai2-dei2, strange-strange-DIM, “a little strange”>, <凍-凍-哋, dung3-dung2-dei2, cold-cold-DIM, “a little cold”>.

At a more abstract level, this is consistent with the treatment of nominal reduplication denoting plural objects, in which the reduplicated elements are connected with a dash. Following the same principle, reduplications in V-one-V and Adj-one-Adj formats are also segmented with dashes, allowing the identification of the lexical verbs and acknowledging that they form a unit: <逛-一-逛, gwaang3-jat1-gwaang3, roam-one-roam, “to roam around briefly”>, <食␣到␣飽-一-飽, sik6␣dou3␣baau2-jat1-baau2, eat␣until␣full-one-full, “eat until very full”>.

4.7 Reduplication

Cantonese has a great variety of reduplication. Across the different reduplication forms, they may not display the same pattern with regard to their relation with the base form, whether all components contribute to the meaning of the whole reduplication, or which component in the reduplication bears the meaning. Therefore, the reduplication forms in Cantonese are segmented differently.

4.7.1 The AXX Template

This pattern is usually found in adjectives, especially with color terms. A disyllabic suffix can be added to the root adjective in the form of adjective-XX, where the syllable X appears twice. Occasionally the two instances of X may have different tones or initials. The reduplicated part cannot be used independently, and is therefore not separated with any marker. A dash is inserted after the root adjective to show the morphemic boundary: <紅-卜卜, hung4-bok1bok1, red-IDEO, “bright red”>, <懵-盛盛, mung2-sing6sing6, muddled-IDEO, “muddled and confused”>.

4.7.2 The XXY Template

This 2+1 (XXY) template has several different structures, many of them unproductive. One pattern is in the form of a disyllabic onomatopoeic expression followed by a verb, the morphological structure of the first part is obscure and should not be further segmented: <條條-掏, tiu4tiu2-fing6, IDEO-hang, “dangling”>, <紮紮-跳, zaat3zaat3-tiu3, IDEO-jump, “vivacious”>.

The suffixes 貢 *gung3* and 震 *zan3* are treated like other verbal particles following reduplicated verbs (i.e., 下 *haa5*, 哋 *dei2*). The dash is used between the reduplicated elements: <串-串-貢, cyun3-cyun3-gung3, cocky-cocky-IDEO, “cocky and provocative”>, <搞-搞-震, gaau2-gaau2-zan3, mess-mess-IDEO, “messing around”>. If the XXY template is created by reduplicating the first syllable of a disyllabic word, add a dash after the first syllable: <貿-貿然, mau6-mau6jin4, REDUP-impetuous, “impetuous”>, <戚-戚然, cik1-cik1jin4, REDUP-sad, “sad”>. If the internal structure of an XXY template is unknown, then the word is kept as is, with no dashes inserted: <死死地氣, sei2sei2dei6hei3, “reluctantly”>.

4.7.3 The AABB Template

Nouns in the form of AABB are segmented in the format of A-A B-B (notice the space between the two reduplications, unlike tetrasyllabic generics in section 4.2.1 above). The rationale behind this is that they are equivalent to phrases like “knives and forks”, “bread and butter” in English. The reduplication is a way to denote genericity. Here, reduplication in Cantonese is treated as a morphological device, and the reduplicated form can be used with another reduplicated form (e.g., hand-hand is used with foot-

foot): <手-手␣ 腳-腳, sau2-sau2␣ goek3-goek3, hand-REDUP␣ foot-REDUP, “hands & feet”>, <刀-刀␣ 叉-叉, dou1-dou␣ caa1-caa1, knife-REDUP␣ fork-REDUP, “knives & forks”>. This type of reduplication also applies to verbs and adjectives: <行-行␣ 企-企, haang4-haang4 ␣ kei5-kei5, walk-REDUP␣ stand-REDUP, “aimless & idle”>, <跑-跑␣ 跳-跳, paau2-paau2␣ tiu3-tiu3, run-REDUP␣ jump-REDUP, “bouncy & active”>.

However, if AB is a disyllabic verb or adjective that cannot be further divided into separate morphemes, i.e., the meaning of AB is non-compositional (e.g., 林審 *lam4sam2*), then the reduplication is treated as if it is a suppletive form of the original form, i.e., no dashes will be inserted: < 林 林 審 審 , lam4lam4sam2sam2, odds.REDUP, “odds and ends”>, < 靚 靚 仔 仔 , leng3leng3zai2zai2, handsome.REDUP, “nice-looking”>.

4.8 A-not-A Question

Alternative questions in Cantonese often take the A-not-A format, where negator 唔 *m4* occurs between the two occurrences of the predicate. With the verb 係 *hai6* “be”, the alternative question becomes 係-唔-係 *hai6-m4-hai6* “be-not-be”. In our segmentation, the entire cluster is segmented by dashes on both sides of the negator. The same rule applies to all elements that can occur in A-not-A, such as modals, verbs and adjectives: (Modals) <會-唔-會, wui5-m4-wui5, will-NEG-will, “Will or will not?”>, <可-唔-可以, ho2-m4-ho2ji5, can-NEG-can, “Can or cannot?”>; (Verbs) <搵-唔-搵-到, wan2-m4-wan2-dou2, search-NEG-search-reach, “Did you find it?”>, <鍾-唔-鍾意, zung1-m4-zung1ji3, like-NEG-like, “Do you like it?”>.

For complex verbal clusters, as exemplified just above, the segmentation of 搵-到 *wan2-dou2* “find” follows the more general rule that it should be split by a dash. In contrast, 鍾意 *zung1ji3* “like” forms its own morpheme and is not split by a dash. That is, the segmentation of the verb phrase is an independent decision from A-not-A. Adjectives are treated in the same way as modals and verbs: <好-唔-好, hou2-m4-hou2, good-NEG-good, “Is it good?”>, <後-唔-後生, hau6-m4-hau6saang1, young-NEG-young, “Is s/he young or not?”>.

Note that for disyllabic predicates, it is typical that only the first syllable occurs before the negator, while some speakers also accept and produce the A-not-A with both syllables occurring before the negator.

4.9 Sentence-Final Particles (SFPs)

Sentence-final particles (SFPs) refer to grammaticalized particles found at the end of an utterance that perform a wide range of grammatical, attitudinal or discourse functions. Cantonese is known to have a large inventory of SFPs, like many other Southeast Asian languages across multiple language families, and some of these particles can be further divided into sub-syllabic morphemes. For the sake of simplicity, each SFP syllable will be treated as one morpheme, and consecutive SFPs will be separated by a dash: <我␣ 睇␣ 嚟-咋-喎, ngo5␣ tai2␣ gaa3-zaa3-wo3, 1SG␣ look␣ SFP1(realis)-SFP2(only)-SFP3(informative),“I want to say that I am in fact only looking around”>.

Some disyllabic SFPs cannot be further decomposed. In these cases, the morpheme boundaries do not coincide with the syllabic boundaries, and are therefore treated as one unit: <吖嗎, aa1maa3>, <嗄嗎, gaa1maa3 (ge3 + aa1maa3)>, <咋嗎, zaa1maa3 (ze1 + aa1maa3)>, < 喇 嗎 , laa1maa3 (laa3 + aa1maa3)>.

Some SFPs at the very end of an utterance behave more similarly to interjections, e.g., 吓 *haa2*, which will be separated from other particles with a space.

4.10 Idioms

Idioms are fixed expressions that often have opaque meanings. In many cases, the idioms cannot be modified or inserted by any elements while preserving the idiomatic meaning, such as 孔 融 讓 梨 *hung2jung4joeng6lei4* “Kongrong giving away the pear” or 一石二鳥 *jat1sek6ji6niu5* “(kill) two birds (with) one stone”. In these examples, the idioms are seen as their own lexical units and therefore no dash or space is inserted. Other examples also include expressions with more than four characters, such as 女大不中留 *neoi2daai6bat1zung1lau4* “daughters should not be kept unmarried”, 條條大路通羅馬 *tiu4tiu4daai6lou6tung1lo4maa5* “all roads lead to Rome”, and 身 在 福 中 不 知 福 *san1zoi6fuk1zung1bat1zi1fuk1* "blessed without knowing". Even though the internal structure of these expressions is visible to morphology, as long as they cannot be modified by intervening elements, they are considered inseparable or “unsegmentable” units.

On the other hand, there are also idioms where some parts can be modified. For example, the idiom 露出馬腳 is segmented as < 露 - 出 ␣ 馬 腳 , lou6-ceot1␣ maa5goek3, expose␣ horse.hoof, “to give (oneself) away”>, because the verb can be modified by aspect or possessive markers. While the meaning of the phrase is non-compositional, the individual morphemes are separable. Therefore, the segmentation treats these multiword expressions as regular sentences, even though their meaning may be opaque: < 露 - 出 - 咗 ␣ 佢 哋 ␣ 隻 ␣ 馬 腳 , lou6-ceot1-zo2␣ keoi5dei6␣ zek3␣ maa5goek3, expose-Perf␣ 3PL␣ CLF␣ horse.hoof, “They gave themselves away.”>.

4.11 Infixes

Infixes in Cantonese are often expletives. For word segmentation, the way an infix is delimited depends on the morphological environment in which infixation occurs. When the target position of infixation has no internal morphological boundary, the infix is not signaled, as if it were part of the original word: <麻鬼煩, maa4gwai2faan4, maa4faan4 “troublesome” + infix gwai2 “ghost”, “freaking troublesome”>.

If the infix occurs at a morphological boundary (e.g., marked by dashes in our scheme), then the infixal word segmentation respects that information by preserving the same word segmentation marking before and after the infix: <日-鳩-日, jat6-gau1-jat6, day-INFIX-day, “every goddamn day” >.

5. Potential Use and Significance

5.1 Facilitating Search

Multi-tiered segmentation provides flexibility for how fine-grained a corpus search is. For example, a user interested in the coarser units may ignore all the dashes or even the pipes, whereas another user concerned with grammatical properties may search by the dash-delimited tokens.

5.2 Natural Language Processing

For natural language processing (NLP), the proposed multi-tiered segmentation for Cantonese readily demarcates larger, coarser units (signaled by pipes) that correspond to units of interest in named-entity recognition. Moreover, our companion re-segmented HKCanCor data will serve as a reference dataset for the proposed multi-tiered segmentation, facilitating the creation of more NLP resources.

5.3 Teaching and Learning

For Cantonese language teaching and learning, the multi-tiered segmentation scheme provides a way to generate multiple word lists for curriculum design. For instance, a word list can be generated by removing all dashes, and a morpheme list can be obtained by splitting at all the dashes. This also makes word-size measures comparable across materials, which is an important step in the development of graded materials and standardized tests for Cantonese proficiency.

6. Conclusion

In this paper, we have outlined a series of word segmentation principles for Cantonese, a Chinese variety with a non-delimited writing system. Given that what counts as a word varies depending on both linguistic and practical factors, the principles are designed as a multi-tiered segmentation scheme to accommodate different levels of granularity for wordhood. As reference data, we have also released a 150,000-character HKCanCor dataset re-segmented by the proposed multi-tiered word segmentation scheme. While the multi-faceted nature of word segmentation makes it challenging to solve word segmentation, our work represents a step forward in both surfacing and tackling some of the nuances with additional segmentation information and guidelines for segmentation decisions.

7. Bibliographical References

Bauer, R. S. (2018). Cantonese as written language in Hong Kong. *Global Chinese, 4*(1), 103-142.

Duanmu, S. (2007). *The Phonology of Standard Chinese*. Oxford University Press.

Fung, R., & Bigi, B. (2015). Automatic word segmentation for spoken Cantonese. In 2015 International Conference Oriental COCOSDA held jointly with 2015 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE) (pp. 196-201). IEEE.

Gong, C., Li, Z., Zhang, M., & Jiang, X. (2017, September). Multi-grained Chinese word segmentation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 692-703).

Hou, X. [侯兴泉] & Wu, N. [吴南开] (2017). 信息处理用粵方言字词规范研究 [A study of the standardization of characters and words in the Yue dialect for information processing]. 广东人民出版社 [Guangdong Renmin Chubanshe].

Luke, K. K., & Lau, C. M. (2008). On loanword truncation in Cantonese. *Journal of East Asian Linguistics*, 17, 347-362.

Matthews, S., & Yip, V. (2011). *Cantonese: A Comprehensive Grammar*. 2nd edition. Routledge.

Myers, J. (2022). Wordhood and Disyllabicity in Chinese. In C. Huang, Y. Lin, I. Chen, & Y. Hsu (Eds.), *The Cambridge Handbook of Chinese Linguistics* (pp. 47-73). Cambridge University Press. https://doi:10.1017/9781108329019.005

Packard, J. (2000). *The Morphology of Chinese: A linguistic and cognitive approach*. Cambridge University Press.

Tang, S. W. [鄧思穎] (2015). 粵語語法講義 [Lectures on Cantonese Grammar], 商務印書館 [The Commercial Press].

Wu, A. (2003). Chinese word segmentation in MSR-NLP. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing* (pp. 172-175).

8. Language Resource References

Chin, A. C.-O. (2013). New resources for Cantonese language studies: A Linguistic Corpus of Mid-20th Century Hong Kong Cantonese. *Newsletter of Chinese Language 92*(1): 7-16.

Lau, C. M., Chan, G. W.-Y., Tse, R. K.-W., & Chan, L. S.-Y. (2022). Words.hk: A Comprehensive Cantonese Dictionary Dataset with Definitions, Translations and Transliterated Examples. In Proceedings of the 1st Workshop on Dataset Creation for Lower-Resourced Languages (DCLRL) @LREC2022 (pp. 53-62), Marseille, France. European Language Resources Association.

Lee, J. L., Chen, L., Lam, C., Lau, C., & Tsui, T. H. (2022). PyCantonese: Cantonese Linguistics and NLP in Python. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (pp.* 6607–6611), Marseille, France. European Language Resources Association.

Lee, T. H.T., Wong, C. H., Leung, S., Man. P., Cheung, A., Szeto, K., and Wong, C. S. P. (1994). *The Development of Grammatical Competence in Cantonese-speaking Children*, Report of RGC earmarked grant 1991-94. https://childes.talkbank.org/access/Chinese/Cantonese/LeeWongLeung.html

Luke, K. K., & Wong, M. L. Y. (2015). The Hong Kong Cantonese Corpus: Design and Uses. *Journal of Chinese Linguistics Monograph Series*, *25*, 312–333.

Shen, F., Yu, W., Min, C., Ye, Q., Xia, C., Wang, T., & Wu, Y. (2021). CyberCan: A New Dictionary for Cantonese Social Media Text Segmentation. https://doi.org/10.31235/osf.io/tyjr7

Winterstein, G., Tang, C., & Lai, R. (2020). CantoMap: a Hong Kong Cantonese maptask corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference* (pp. 2906-2913).

Wong, T. S., Gerdes, K., Leung, H., & Lee, J. S. (2017). Quantitative comparative syntax on the Cantonese-Mandarin parallel dependency treebank. In *Proceedings of the fourth international conference on Dependency Linguistics (Depling 2017)* (pp. 266-275).

Yip, V., & Matthews, S. (2007). *The Bilingual Child: Early Development and Language Contact.* Cambridge University Press. https://childes.talkbank.org/access/Biling/YipMatthews.html