Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3069–3077
Marseille, 11–16 May 2020

Cifu: a frequency lexicon of Hong Kong Cantonese

Regine Lai, Grégoire Winterstein
The Chinese University of Hong Kong, Université du Québec à Montréal
Department of Linguistics and Modern Languages, Département de Linguistique
ryklai@cuhk.edu.hk, winterstein.gregoire@uqam.ca

Abstract

This paper introduces Cifu, a lexical database for Hong Kong Cantonese (HKC) that offers phonological and orthographic information, frequency measures, and lexical neighborhood information for lexical items in HKC. The resource can be used for NLP applications and the design and analysis of psycholinguistic experiments on HKC. We elaborate on the characteristics and challenges specific to HKC that were relevant in the design of Cifu. This includes lexical, orthographic and phonological aspects of HKC, word segmentation issues, the place of HKC in written media, and the availability of data. We discuss the measure of Neighborhood Density (ND), highlighting how the analytic nature of Cantonese and its writing system affect that measure. We justify using six different variations of ND, based on the possibility of inserting or deleting phonemes when searching for neighbors and on the choice of data for retrieving frequencies. Statistics about the four genres (written, adult spoken, children spoken and child-directed) within the dataset are discussed. We find that the lexical diversity of the child-directed speech genre is particularly low, compared to a size-matched written corpus. The correlations of word frequencies of different genres are all high, but in general decrease as word length increases.

Keywords: Hong Kong Cantonese, lexicon, lexical frequency, lexical neighborhood, neighborhood density

1. Introduction

This paper describes the construction of Cifu: a lexical database for Hong Kong Cantonese.[1] The resource lists lexical elements of Hong Kong Cantonese accompanied by a measure of their frequency in various genres, as well as morphological and orthographic information. The resource is modeled after comparable ones for other languages such as LEXIQUE3 for French (New and Pallier, 2019), or CELEX2 for English, Dutch and German (Baayen et al., 1995). In addition, the resource lists information about the "lexical neighborhood" of the items in the lexicon, i.e. other items in Cifu that are phonologically similar to the source item via some measuring distance (e.g. Levenshtein Distance).

Such a resource is useful for both NLP applications (e.g. as the input of word segmenters, PoS taggers…) and for designing the materials of psycholinguistic experiments, typically for balancing the frequency of lexical items, or in order to find candidate items for priming experiments. Evidence from psycholinguistics methodology such as priming have shown that factors such as word frequency and phonological similarity to other words in the lexicon have a sizeable impact on language processing such as in word recognition and production, as well as in language acquisition (e.g. (Dell and Gordon, 2003; Storkel et al., 2006; Vitevitch and Luce, 1999)). These effects are found in populations in a wide range of age groups, ranging from children (e.g. (German and Newman, 2004; Munson, 2001; Munson et al., 2005; Storkel and Lee, 2011)) to older adults (e.g. (Newman and German, 2005; Sommers and Danielson, 1999)), populations with speech and language disorders, hearing impairment (Dirks et al., 2001) and aphasia (Gordon and Dell, 2001; Gordon, 2002; Vitevitch and Castro, 2015). Although the effects of frequency and phonological similarity are robust, cross-linguistic differences have been observed in languages that are typologically close like English, French and Spanish (Dufour and Frauenfelder, 2010; Vitevitch and Rodríguez, 2005). As pointed out in (Vitevitch and Luce, 2016) , the root cause of these cross-linguistic differences can only be discovered when the effect is studied in a wider range of languages which differ from English, French and Spanish in terms of phoneme inventory, typical length of words and morphological structure. Cantonese being a morphologically isolating language and a language with lexical tones should be a good test case. However, relatively few studies have tested the effect phonological similarity using Cantonese. This is probably due to the lack of lexical information on the languages such as the ones that are constructed in the current lexical database. Cifu is the first Cantonese database which provides public access to such crucial lexical information for conducting psycholinguistics research.

There are however a number of issues pertaining to the specific case of Cantonese that need to be addressed when constructing such a resource. The major one has to do with the peculiar status of the written medium to which Hong Kong Cantonese speakers are used to, and which routinely mixes elements from Cantonese with Mandarin and is not standardized in the same way as (for example) Mandarin Chinese.

In Sec. 2 , we introduce the main features of Hong Kong Cantonese (HKC) that help determine the lexical elements listed in the database and the characteristics associated with these elements. In Sec. 3 , we discuss the spoken and written forms of HKC, and consider the available resources for establishing our frequency database. Sec. 4 describes the concept of Neighborhood Density. Sec. 5 describes the resource and its building process. In this section, we mention comparable resources for HKC and which were instrumental in building Cifu. We conclude in Sec. 6

2. General features of Cantonese

Cantonese is a sinitic language spoken in Hong Kong, Macau, the Guandong region of southern PRC and through-

[1] Cifu: 詞庫 (ci4fu3: /ts^{h}i˩fu˧/) is a compound word in Cantonese. The first character 詞 means 'words' and the second character 庫 refers to 'storage', and as a compound it refers to 'lexicon'.

out the Chinese diaspora. It is the second most spoken sinitic language, after Mandarin Chinese, and is the one that is the most spoken outside of the PRC (Wiedenhof, 2015). Here, we focus on the variety spoken in Hong Kong, where about 89% of the population use it as a first language.[2] We first discuss what constitutes a lexical item in Cantonese, then discuss aspects of the phonology of Cantonese that will be relevant when discussing neighborhood density.

2.1. Lexical units and segmentation

Like all sinitic languages, Cantonese is an analytic language with little to no inflectional morphology. It is often argued that almost every syllable in the language is a morpheme (Matthews and Yip, 2011). However, not every isolated syllable can function as a word in the sense of an autonomous unit of syntax. The question of how to determine whether a sequence of syllables is a word remains an open problem in Cantonese, as with other sinitic languages (see e.g. (Magistry, 2013) for extensive discussion). The following examples illustrate recurring issues.

- The expression 甲由–*gaat6zaat2*[3] ('cockroach') is a word, but neither of its two syllable/character compounds exists on its own, nor is used in any other word, i.e. aren't morphemes.
- The expression 左右–*zo2jau2* ('approximately') is a word made of two morphemes, meaning respectively 'left' and 'right', but which combine in a non-compositional way.
- Cantonese has a series of affixes (that indicate, a.o., aspectual or modal information) that attach to verbs (e.g. 食咗–*sik6zo2* eat-ASP 'ate', 行緊–*hang4gan2* walk-ASP 'walking'. etc.) These units cannot exist on their own, though in some works they are treated as individual words (Luke and Wong, 2015).
- Loan words regularly involve multiple syllables with no morphemic status since they aim at reproducing the sonority of the term being borrowed (e.g. 士多啤梨–*si6do1be1lei2* 'strawberry', 馬拉松–*maa5laai1cung4* 'marathon', 維他命–*wai4taa1ming6* 'vitamin' etc.).

The usual way to write Cantonese (and any other Chinese language) does not use any spacing. The issues pointed above are therefore directly relevant for the task of segmentation that necessarily comes before counting items frequencies. Some of the corpora used in building `Cifu` were manually segmented (see infra), some were segmented automatically with existing tools, and others were not segmented at all.[4] This entails that segmentation schemes are not consistent across corpora, and distinguish units with different levels of granularity.

[2] cf. `https://www.censtatd.gov.hk/`, consulted Aug. 12, 2019

[3] The Jyutping romanization system for Cantonese is used in this paper, as it is endorsed by the Linguistic Society of Hong Kong (`https://www.lshk.org/jyutping`)

[4] There exists detailed segmentation schemes for Mandarin Chinese that could be used for Cantonese (Huang et al., 1996), but they have not yet been used in the Cantonese corpus community.

2.2. Syllable structure

HKC syllables follow a rather stable template characterized by:

- An optional consonant onset among the following: *p, t, k, b, d, g, z, c, m, n, ng, l, w, j, h, f, s, gw, kw*. An example of word that allows a zero onset is 鴨–*aap3* 'duck', and one with an onset is 塔–*taap3* 'tower'. Onset consonants have few restrictions in Cantonese. All nineteen consonants in the inventory can occur in the onset position, but clusters are not allowed. *gw* and *kw* are considered as singletons that are labialized.
- A vowel: there are eight monophthongs *i, yu, u, e, oe, o, a, aa* and eleven diphthongs *ui, ei, eoi, oi, ai, aai, eu, iu, ou, au, aau*. Diphthongs only occur in open syllables, whereas monophthongs can occur either in open or closed syllables. Two syllablic consonants, *m* and *ng*, can form the nucleus of a syllable in Cantonese, but neither onsets nor codas are allowed to cooccur with these syllabic consonants within a syllable. A syllable can be limited to a single vowel, e.g. 呀–*aa3* (a sentence final particle) and 唔–*m4* (negation morpheme) which consists of a syllabic consonant.
- An optional coda among a subset of the consonants. Possible consonants are the plain unapirated stops *p, t, k* and nasals *m, n, ng*. Examples are: 試–*si3* 'test' for an element without coda, and 扇–*sin3* 'fan' with coda.
- A tone: there are six tones in Cantonese, three of which are level tones (Tone 1: High level, Tone 3: Mid level, Tone 6: Low level), two are rising (Tone 2: High rising, Tone 5: Low rising) and one is a falling tone (Tone 4: Falling). Only level tones are allowed in closed syllables (in underived forms). Tones in open syllables are unrestricted. In some descriptive studies (e.g. (Hashimoto, 1972)), three additional tones, the entering tones, are included in the Cantonese tonal inventory. These three entering tones only appear in closed syllables, they are in complementary distribution with the three level tones, and have the same prosodic contour and pitch level as the level tones, and thus we assume a six-tones inventory instead of nine.

As mentioned above, `Cifu` follows the Jyutping romanized scheme. Jyutping can straightforwardly be mapped back to an IPA phonemic transcription. See Table 1 for the conversion of Jyutping to IPA.

3. Transcribing Cantonese

3.1. Spoken vs. written Cantonese

Cantonese, like other Chinese languages, is commonly written with Chinese characters. However, HKC has not been standardized in writing in a way comparable to Mandarin. Instead, in Hong Kong, children are taught to write in standard written Chinese, which itself is based on Mandarin. Standard written Chinese differs from HKC in considerable ways, similar to the differences between Mandarin and HKC, which includes lexical differences. For example, the third person pronoun in HKC is 佢–*keoi5*, whereas standard

Cons Jyutping	Cons IPA	Vowel Jyutping	Vowel IPA	Tone Jyutping	Chao's Tone Letters
p	p^h	*i*	i	*1*	55
t	t^h	*yu*	y	*2*	25
k	k^h	*u*	u	*3*	33
b	p	*e*	ɛ	*4*	21
d	t	*oe*	œ	*5*	23
g	k	*eo*	ɵ	*6*	22
c	ts^h	*o*	ɔ		
z	ts	*a*	ɐ		
m	m	*aa*	a		
n	n				
ng	ŋ				
l	l				
w	w				
j	j				
h	h				
f	f				
s	s				
kw	k^{wh}				
gw	k^w				

Table 1: Conversion of Jyutping to IPA

written Chinese uses the character 他–*taa1*, which corresponds to Mandarin tā.

Therefore, speakers of HKC are exposed to distinct languages when learning to speak and write. Nevertheless, there is a natural tendency, for example on social media, but also in journalism or advertising (Snow, 2004), to write in a manner that more closely resembles the way people speak and to which speakers of HKC are accustomed. Therefore, there exists a written form of HKC, which is placed on a continuum between spoken HKC and standard written Chinese. Written HKC typically uses the Cantonese versions of functional items since it tends to follow Cantonese syntax (pronouns, particles…), but there is considerable variation in the use of lexically full items, e.g. 'tomorrow': 聽日–*ting1jat6* (spoken) vs. 明天–*ming4tin1* (written); 'forget': 唔記得–*m4gei3dak1* (spoken) vs. 忘記–*mong4gei3* (written). See (Matthews and Yip, 2011) and (Snow, 2004) for a more systematic investigation of written HKC and its relation to both spoken HKC and standard written Chinese.

3.2. Cantonese characters

The characters used to transcribe HKC belong to the set of traditional characters (as in Taiwan, among other places) as opposed to the simplified ones favored in PRC, Singapore and Malaysia. As mentioned above, although written Cantonese is not explicitly taught in schools, it is used in some media outlets, social media, and communications through text messaging apps on cell phones. The choice of character to represent a given morpheme in Cantonese is not standardized and there are variations. Regional differences between some of the character choices are found between Guangzhou and Hong Kong Cantonese speakers. For example, the genitive marker *ge3* is typically written as 嘅 by HKC speakers, whereas Guangzhou speakers usually opt for 噶. Differences among speakers in Hong Kong are also observed. The most common and systematic difference is the (non-)use of the 'mouth' radical 口 for words that are only allowed in colloquial context. The genitive marker is again an example of such case. The character can be represented either by the character with the 'mouth' radical, as in 嘅, or the one without, as in 既. These variations are tolerated because the phonetic information is conveyed by the other component of the character, and the exclusion of the radical does not impede the understanding of the text (e.g. because syntax makes clear which item is denoted by the character). Beyond these cases, there are no standardized forms in many other less systematic cases, including very frequent ones. For example there are two frequent characters used to write the adverb *zung6* 'still' (仲 or 重). In the perspective of `Cifu`, we need separate information for each character, even if they're referring to the same linguistic unit, since (among other things) in some priming studies, the frequency of a glyph matters in addition to the frequency of the linguistic unit, and we thus need to evaluate the frequency of a character/linguistic unit compound.

In addition, some HKC words entirely lack characters. Examples include *kwaak1* 'to outline' (as in *kwaak1 go3 hyun1* 'outline a circle', or *soe4* 'to slide'[5] (as in *soe4waat6tai1* 'slide down a slide' (Matthews and Yip, 2011)), and are customarily written either by choosing a Mandarin equivalent or by using roman letters.

Finally, other units do have a character, such as the morpheme *di1* 啲 and the loanword for 'baby' *bi4bi1*, but tend to be written in alternate ways, often using roman script (in that case capital D, and BB, respectively), e.g. on social media.

3.3. Orthographic information: characters structure

A character can be seen as the composition of several basic components, each of which is a sequence of strokes of a particular type. The simplest character has one stroke (一–*jat1* 'one'), going up to 64 strokes (though the characters in question fell out of use long ago, and have no unicode point). The number of strokes in a given character is one of the factors that correlates with the complexity of recognizing/identifying the character, along with phonological factors and the individual frequencies of the sub-characters that compose the character (see (Feldman and Siok, 1999) for a review of such factors).

There are ambiguities when it comes to writing a given character. One example is the character 免 which counts 7 or 8 strokes and can be decomposed either as a two or three layered character.

The parts that compose a character often give indications on the meaning and pronunciation of the character, though this is not done in a systematic and predictable way. The most common type of characters is made up of a semantic and a phonetic component. The semantic component is usually the radical of a character, and radicals are how characters

[5]The character 瀡 has recently been gaining traction as a way to write 'soe4', but remains far from being used commonly.

organized in a dictionary. There is however no systematic way to identify the radical of a character based on its decomposition (e.g. the radical does not occupy a specific position in the character, though there are general trends).

4. Neighborhood information

Neighborhood density is a way to quantify a word's phonological similarity to other words in the lexicon. A word that is said to be in a dense neighborhood means it has many similar sounding words. On the contrary, if a word is considered to be in a sparse neighborhood, it has few words that sound similar to it. Consistent findings obtained from word perception experiments have shown that phonological similarity of a word affects the speed and accuracy of its processing. More specifically, words that have more phonological neighbors are processed slower than those with less (for a review, see (Vitevitch and Luce, 2016)).

4.1. Definition

A common way to quantify phonological similarity is using Levenshtein distance. A neighbor of a target word is defined as a word that is one phoneme different from any of the phonemes of the target word. The one phoneme difference can be realized through substitution, insertion or deletion (Greenberg and Jenkins, 1964; Landauer and Streeter, 1973). For example, the English target word 'beam' /bim/ has phonological neighbors such as 'seem' /sim/ (substitution), 'bee' /bi/ (deletion), 'bream' /brim/ (insertion). The phonological neighborhood of a word consists of the set of neighbors that are produced using Levenshstein Distance, and the neighborhood frequency of a word is the mean of its neighbors' frequencies of occurrences (Vitevitch and Luce, 2016). The neighborhood density of an item is then defined as the average of the individual frequencies of the terms in its phonological neighborhood. Since tones are phonemic in Cantonese, they are also considered in the calculation of neighborhood density. For example, the tonal neighbors of a Cantonese word *si1* 'poem' are *si2* 'history', *si3* 'exam', *si4* 'time', *si5* 'city', *si6*'yes'. Since a syllable contains one and only one tone, no tone deletion or insertion can be applied. As shown in the examples above, if the phonemic inventory remains constant, i.e. the numbers of consonants and vowels are the same for two languages, the possibility of a word to have neighbors is higher for a word in a tonal language than in a non-tonal language.

4.2. Uses

The effect of neighborhood density calculated in the manner described above has been shown in various tasks, such as, word learning, visual word recognition, short-term and long-term memory tasks (Luce and Large, 2001; Storkel, 2004; Yates et al., 2004; Roodenrys et al., 2002; Sommers and Lewis, 1999). Among the sparse literature on the effect of neighborhood density in Cantonese, Kirby and Yu (2007) found that neighborhood density had a significant effect on grammaticality judgements of Cantonese monosyllabic words and nonwords. Although literature on English have confirmed that the effect of neighborhood density of longer words (bisyllabic) is comparable to that of monosyllabic words (Cluff and Luce, 1990; Vitevitch et al., 2008), the effect in longer words is potentially different in Cantonese as longer words in Cantonese are less likely formed via affixation, and this sets Cantonese apart from English, French and other other Indo-European languages that are well studied in this respect. Whether or not this difference in morphology influences how language is processed requires further investigations, and the relevant lexical information on longer words that is necessary for such investigations are available in our database.

5. The `Cifu` Resource

The `Cifu` resource was created with all the previous considerations in mind. Here, we first describe the choice of corpora to analyze, then the steps of the data processing in these corpora to extract words, their frequencies, and their neighborhood densities in order to create `Cifu`.

5.1. Data

We opted to distinguish four different genres for which we calculate word frequencies. These genres are: *Written* discourse, *Spoken* discourse by both adults and children, and *Child directed* speech. The last one is distinguished on account of its idiosyncratic properties compared to standard adult speech (e.g. a lower lexical variety), and because it is the direct input children receive in addition to their ambient languages. Carlson et al. (2014) have shown that the lexical properties of child-directed speech are a better predictor of children's vocabulary growth than those of adult-directed speech. We thus have made both sets of frequencies available to those who are interested in research in the area of language acquisition. We relied on existing corpora for all the spoken genres. The sources used for the three spoken genres are summarized in Table 2.

Genre	Corpora
Adult Spoken	HKCanCor (Luke and Wong, 2015), HKCAC (Leung and Law, 2001), CantoMap (Lai and Winterstein, 2019)
Child Spoken / Child-directed	HKU-70 (Fletcher et al., 2000), Lee/Wong/Leung Corpus (Lee et al., 1994)

Table 2: Corpora used for the spoken genres

The use of movie subtitles, routinely used to approximate spoken data (New and Pallier, 2019), is not an option here since the vast majority of movie subtitles are in written Chinese, straying far from HKC (in addition to the general gap between genuine spoken data and scripted movie dialogues).

For the written genre, we scraped 3,841 chapters of amateur novels from the website `https://www.shikoto.com/`. The website is a platform for digital literature. We chose that source on account of the variety of the themes of the stories (science fiction, love stories, motivational, horror, comedy etc.) and the fact that they are fairly representative of the written text to which HKC speakers are exposed on a daily basis. Though some of the materials in the corpus resembles written Chinese, it is not prevalent throughout

the data, striking a desirable balance between prescriptive norms of writing and styles more closely resembling spoken HKC. The target audience of the digital literature on this platform is young people (roughly below 30 years of age), it thus has the potential of being biased in age. However, this is the age group that is the most fluent in and most likely to use written Cantonese and therefore it reflects the current usage of written Cantonese in the society. The use of written Cantonese is less consistent in other sources from the internet. Many of the websites (e.g. news outlet) often choose to use the standard variety (Chinese Mandarin) for most of their content and occasionally with sections (such as gossip columns) that are written in Cantonese. Due to the inconsistency of Cantonese usage on the web, it is unrealistic for us to use the internet as the source of our database.
The size of each corpus, measured in tokens, is indicated in Table 4. Authors like Brysbaert and New (2009) suggest that a size of about 15 millions tokens guarantees a robust estimation of the term frequencies (i.e. which correlates well with psycholinguistic measures). However, because HKC is not particularly well resourced and that corpus data (especially spoken) is not always freely available, we made do with what is available at the time, and have much less than that target size.

5.2. Processing

5.2.1. Segmentation

The first step of processing was to segment all the data in the corpora in a consistent way. We tested three segmenting tools: `Thulac` (Li and Sun, 2009) , `Jieba` (Junyi, 2019) and `SPPAS` (Bigi, 2015) , all of which handle Chinese characters. Out of the three, `SPPAS` has a module dedicated to the segmentation of Cantonese, but we ended up selecting `Jieba` for its better overall qualitative performance, once we paired it with a custom dictionary. That custom dictionary was created on the basis of the wordlist used by the `MOR` module of the `CLAN` software to tag parts of speech files (MacWhinney, 1995).[6] That word list proved superior to other existing wordlists available for Cantonese, for example the `yedict` list we used later to access the romanization of terms (see infra). Note that, though it uses a word list to bootstrap segmentation, the segmenter is able to correctly segment forms outside of its known vocabulary.
As mentioned earlier, some Cantonese corpora are already segmented, sometimes manually, sometimes automatically (e.g. CantoMap, HKCanCor, CHILDES). However, since the segmentation procedures differ across corpora (or were not documented), we decided to re-segment all our data in a consistent way to ensure that all the data was treated in a uniform way.

5.2.2. Creating the list / counting frequencies

Once all the data were segmented, we made a basic count of occurrences across genre of terms appearing in the data. We filtered terms by excluding:

- terms that have no Chinese characters in them, e.g. English words in code switching situations etc.
- hapaxes that were not found in the `yedict` dictionary, i.e.,any term that only appeared once across all genres and that was not found in `yedict` was excluded.

The final list comprises 51,798 distinct entries across the four genres.
Measures of frequencies per million tokens were calculated. These were calculated on the basis of all tokens in the corpus, excluding punctuation, but including non-Chinese words and hapaxes that were removed from `Cifu`.

5.2.3. Jyutping transcription and definitions

The Jyutping transcriptions of the entries in `Cifu` were automatically obtained from the `yedict` lexical database[7], itself an adaptation of the CEDICT Mandarin-English dictionary[8] for Cantonese. `yedict` offers both a Cantonese romanization of its terms and their definition in English. That list is very large (over 147,000 entries), with many entries corresponding to entire phrases (which partly explains why it is unfit as the basis for the word segmenter, cf. supra). Yet, some entries in `Cifu` were not found in `yedict`. For these unknown terms, we reconstructed their Jyutping romanization by using a `MaxMatch` algorithm searching for the biggest known sub-parts of the term in `yedict`. In total, there are 29,995 such reconstructed forms (44.4% of the resource), indicated with a `*` symbol in `Cifu`.
Besides the romanization, we also extracted the English definition of the terms and added it to `Cifu` entries. In total, 21,860 entries in `Cifu` have such a definition.

5.2.4. Character information

To retrieve information about the characters in the entries of `Cifu`, we used the IDS data available at `https://github.com/cjkvi/cjkvi-ids` (part of the Kanji Database Project `http://kanji-database.sourceforge.net/`). Specifically, we relied on the data stemming from the "Chinese Document Processing lab" at Academia Sinica.
For each entry in `Cifu`, we indicate:

- The number of strokes in each character in the entry. When there are several possibilities, they are all indicated, separated by commas.
- The composition of the character in terms of the relative placement of the sub-character components (e.g. ⿰ to indicate two components side by side), followed by the components occupying each position. This representation is recursive if the components occupying a certain position can be further decomposed. Some of the sub-character components are characters in their own right, but not all of them are. In some instances, a position is occupied by a descriptive entity that is not a character. These entities are identified with codes starting with `&`. Finally, the composition information includes optional indications in square brackets relative to the scheme used to decompose it (details can be found on the relevant website).

[6] The wordlist is available on `https://talkbank.org/morgrams/`

[7] `https://writecantonese8.wordpress.com/2012/02/04/cantonese-cedict-project/`

[8] `https://cc-cedict.org/`

For both kinds of information, the informations about the different characters in the entry are separated by colons.

5.2.5. Neighborhood densities

The neighborhood density (ND) of an entry in `Cifu` is the average of the frequencies of the neighbors of a word, where neighbors are characters defined as elements that have a Levenshtein distance of 1 with the target word (cf. Sec. 4). We considered several variations on the calculation of ND, based on two independent factors.

First, we distinguished ND by whether inserting or deleting phonemes was allowed when searching for neighbors. Given that both tone and vowel are obligatory elements of a Cantonese syllable, this meant considering (or not) the possibility of adding/removing the onset and coda of every syllable in the source word. For example, in case insertion and deletion are not considered, the neighbors for the entry of 鴨–*aap3* ('duck') will only be searched by considering alternate vowels, tones and non-empty codas. If insertion and deletion are allowed, then the addition of all possible onsets will also be considered, as well as the possibility of an empty coda. In theory though, a neighbor of a word can be obtained by inserting a vowel, e.g. a V could be inserted next to CVC word to create a bisyllabic word of any of these forms: V.CVC/ VC.VC/CV.CV. Similarly, deleting a vowel could also generate smaller neighbors, e.g. CVCV → CVC. In order to restrict the neighbors of a word to elements of the same length in terms of the number of syllables it contains, we have decided that in the cases above and where a word contains an onset and a coda, e.g. /\–*baat3* ('eight'), no insertion was applied. Table 3 summarizes the processes that were applied to different syllable types in neighborhood calculation.

Syllable Type	Substitution	Insertion	Deletion
CVC	✓	×	✓
CV	✓	✓	✓
VC	✓	✓	✓
V	✓	✓	×

Table 3: The processes applied to different syllable types

Second, we calculated separate ND by using either only the frequencies in the Written genre, the Spoken Adult genre or the average frequency of the two. This was done because depending on the type of experiment done (e.g. relying on written or audio stimuli) the relevant frequencies might differ.

To search for the neighbors of an entry in `Cifu`, we decomposed its Jyutping transcription as a list of syllables, themselves decomposed in their four components: onset, vowel, coda and tone (cf. Sec. 2 2). The decomposition was done automatically using the `pycantonese` library (Lee et al., 2015) and ensuring that dipthongs were treated as a single vowel. The only `Cifu` entries we considered for ND (both as source and neighbor) were such that their Jyutping transcription was not reconstructed (cf. Sec. 5 2 3), or, if reconstructed, contained no ambiguity. For all the retained entries, we substituted in turn each component of each syllable with all alternative options for that component. For example, in the sequence *zung1*, the tone was replaced by all five other options, resulting in the forms *zung2/zung3/zung4/zung5/zung6*. All resulting forms were searched within `Cifu` and aggregated for the calculation of ND. When considering an alternate word, we made sure that it had at least one pronunciation that differed from the pronunciation of the source word. This is because both source words and potential neighbors might have multiple pronunciations. For example, the entry of 阿 has two possible pronunciations: *aa3* and *o1*. Among its potential neighbors is 呀 which has three pronunciations: *aa3,aa1* and *aa6*. Though one of these matches the pronunciation of the source word 阿, its other pronunciations do not, and therefore 呀 counts as a neighbor of 阿 (and vice-versa in this case, though not in the general case).

The ND for a `Cifu` entry is then calculated as the average of the individual frequencies of the items in its neighborhood set (i.e. with no repetition of the items). For example, the entry for 好 has two possible pronunciations: *hou2* and *hou3*. This entails that a sequence like *hou4*, which is a neighbor of both *hou2* and *hou3*, will only be used once when computing the neighborhood of 好, i.e. each of the character that have *hou4* as their pronunciation (i.e. 嗥豪蠔嚎濠毫) will appear only once in the neighborhood of 好.

As mentioned above, three different ND were calculated by varying the source of the frequencies of the neighbors: from the Written genre, from the Spoken Adult genre, or an average of those two. This thus gives 6 different ND measures in total for each entry in `Cifu`.

5.3. Basic statistics

Table 4 shows the aggregated sizes of the corpora used to compute frequencies in each genre, along with information about the number of word types and character types (i.e. individual characters that are used to form words) for each genre.

Genre	**# Tokens**	**# Chinese tokens**	**# Word types**	**# Char. types**
Written	766,461	748,282	49,626	4,430
Adult Spoken	331,623	310,037	11,970	2,764
Child Spoken	173,741	234,611	2,489	1,546
Child-directed	705,791	586,977	3,531	1,925

Table 4: Corpus size by genre

As expected, the Written genre offers the most diverse lexicon, and the spoken adult part sharply contrasts with the child directed speech in terms of lexical diversity. As can be seen, though it has a comparable size to the Written part, the Child-directed section has a much lower number of word types (about 14 times less), and also has more than three times less types than the Adult Spoken part which is only

about half the size of the Child-directed section. A likely explanation for this is that the main purpose of the child-directed speech that was obtained from the Cantonese corpora from CHILDES was to elicit children's production. It thus mainly consists of confirmation questions and repetitions, and this might have driven the number of tokens high, yet the number of types remain low.

A Pearson correlation test using the *r* statistic was calculated to test the relationship between the word frequencies in the four genres, see Table 5 . Children's speech is the most correlated with Child-directed Speech ($r = .890$, $p < .001$) further confirming that Child-directed involves a lot of reprises, corrections etc., and that data is also highly correlated with adult speech ($r = .741$, $p < .001$). The word frequencies of the written and spoken genres are highly correlated ($r = .785$, $p < .001$). The correlations between the word frequencies between child-directed speech and children's speech remain high in words of length from 1 to 4 syllables. However, the correlations between adult spoken and children spoken decrease as words become longer (bi-syllables: $r = .162$, $p < .001$; trisyllables: $r = .890$, $p = .010$; quadrasyllabic: $r = -.009$, $p > .05$).

Genre	All data	1-syll.	2-syll.	3-syll.	4-syll.
Adult Spok. vs. Writ.	.785***	.788***	.652***	.496***	.364***
Child-dir. vs. Child Spok.	.890***	.890***	.770***	.919***	.809***
Adult Spok. vs. Child Spok.	.740***	.755***	.162***	.208**	-.009
Adult Spok. vs. Child-dir.	.741***	.753***	.233***	.044***	.809***

Table 5: Correlations of Word Frequencies between Genres

5.4. Entry example

Table 6 shows all the information available in an entry of `Cifu` along with the type of each piece of information.

5.5. Availability

`Cifu` is available on a `github` repository at the following URL: `https://github.com/gwinterstein/Cifu`. `Cifu` is released under the GNU General Public License v3.0. Future updates of the resource will be hosted in the same place, and contributions to the project by interested parties are welcome.

[7]The frequency information appears for the four genres in `Cifu`: Written, Spoken Adult, Spoken Child and Child directed.

[8]The neighborhood densities appear in 6 versions, as described in Sec. 5.2.5.

Feature	**Type**	**Example**
Word	UTF-8 string using the HKSAR character set	不過
Jytuping	ASCII string	bat1gwo3
Frequency (occ.)[7]	integer	5599
Frequency (per million tokens)	float	2318.744
# Strokes	list of integers (colon and comma separated)	4:11,12,13
Structure (composition of the characters)	UTF-8 string	⿰ 一 &CDP-8665;: ⿺ 辶 咼
Definition	ASCII string	only / merely / no more than / but / however / anyway (to get back to a previous topic)
ND[8]	integer	3.8904

Table 6: The values of an entry in `Cifu`

6. Conclusion and outlooks

`Cifu` is the first resource that offers frequency and neighborhood information for terms in HKC. As such it distinguishes itself from other existing lexicon of HKC (such as the word list of CHILDES or the `yedict` dictionary), and should be useful to researchers who need such information, for example in designing psycholinguistics experiments.

Our preliminary results show that the length of most Cantonese words lies between 1 to 3 syllables, see Table 7. The number of homophones in HKC is high, and at the same time, many characters in HKC have various pronunciations. In addition, tones are phonemic in Cantonese, and the cumulative effect of all these factors greatly increase the neighborhood density of a lexical item in HKC. It is unclear how or whether this has an effect on the general processing of words, either visually or auditorily on its speakers, and whether such an effect leads to a greater dependency on the syntactic and pragmatic contexts during the processing of words. Another aspect that is worth investigating is the effect of the six different types of ND that we calculated in `Cifu`. They are all included in the database because it is unclear, based on the current findings in the field, whether one is more relevant than others. The effect of the various measures of ND, such as the ones obtained from orthographic and phonemic transcriptions, should be tested in future psycholinguistic investigations as they could very well be tapping into different aspects of language processing in Cantonese, or more generally, Chinese speakers.

Future work about `Cifu` will focus on three additional as-

Word Length	no. of occur.	Freq.
1	3,935	.076
2	37,910	.729
3	8,406	.162
4	1,462	.028
5	59	.001

Table 7: The Frequency of Words of Length up to 5 syllables

pects. First, we will improve and evaluate thoroughly the automatic segmentation of the data. Right now, the entries in `Cifu` are conservative to an extent to ensure that they correspond to actual words. This was done by ensuring that the terms retained in `Cifu` are either already listed in existing dictionaries, or are not hapaxes. Second, we intend to add part of speech information to the entries. This is however a big challenge since there is no established and consensual inventory of PoS for HKC (though some works already used their own inventory: (Luke and Wong, 2015)). One option would be to follow the framework of Universal Dependencies for which annotated data already exists (Wong et al., 2017). Third, we aim at using larger datasets to compute frequencies, though in the case of spoken data this entails either recording new material, or using alternate means of accessing data, again not a small endeavour.

7. Bibliographical References

Bigi, B. (2015). SPPAS - Multi-lingual Approaches to the Automatic Annotation of Speech. *The Phonetician*, I–II(111–112):54–69.

Brysbaert, M. and New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English . *Behavior Research Methods*, 41(4):977–990.

Carlson, M. T., Sonderegger, M., and Bane, M. (2014). How children explore the phonological network in child-directed speech: A survival analysis of children's first word productions. *Journal of memory and language*, 75:159–180.

Cluff, M. S. and Luce, P. A. (1990). Similarity neighborhoods of spoken two-syllable words: Retroactive effects on multiple activation. *Journal of Experimental Psychology: Human Perception and Performance*, 16(3):551.

Dell, G. S. and Gordon, J. K. (2003). Neighbors in the lexicon: Friends or foes. *Phonetics and phonology in language comprehension and production: Differences and similarities*, 6:9–37.

Dirks, D. D., Takayanagi, S., Moshfegh, A., Noffsinger, P. D., Fausti, S. A., et al. (2001). Examination of the neighborhood activation theory in normal and hearing-impaired listeners. *Ear and hearing*, 22(1):1–13.

Dufour, S. and Frauenfelder, U. H. (2010). Phonological neighbourhood effects in french spoken-word recognition. *The Quarterly Journal of Experimental Psychology*, 63(2):226–238.

Feldman, L. B. and Siok, W. W. (1999). Semantic Radicals in Phonetic Compounts: Implications for Visual Character Recognition in Chinese. In Jian Wang, et al., editors, *Reading Chinese Script, A Cognitive Analysis*. Lawrence Erlbaum Associates.

Fletcher, P., Leung, S. C.-S., Stokes, S. F., and Weizman, Z. O., (2000). *Cantonese pre-school language development: A guide.* Hong Kong University, Department of Speech and Hearing Sciences, Hong Kong.

Fung, R. and Bigi, B. (2015). Automatic Word Segmentation for Spoken Cantonese. In *The Oriental Chapter of COCOSDA*. International Committee for the Coordination and Standardization of Speech Databases and Assessment Techniques.

German, D. J. and Newman, R. S. (2004). The impact of lexical factors on children's word-finding errors. *Journal of Speech, Language, and Hearing Research*.

Gordon, J. K. and Dell, G. S. (2001). Phonological neighborhood effects: Evidence from aphasia and connectionist modeling. *Brain and Language*, 79(1):21–23.

Gordon, J. K. (2002). Phonological neighborhood effects in aphasic speech errors: Spontaneous and structured contexts. *Brain and language*, 82(2):113–145.

Greenberg, J. H. and Jenkins, J. J. (1964). Studies in the psychological correlates of the sound system of american english. *Word*, 20(2):157–177.

Hashimoto, O.-k. Y. (1972). *Phonology of Cantonese*, volume 1. Cambridge University Press.

Huang, C.-R., Chen, K.-j., and Chang, L.-L. (1996). Segmentation Standard for Chinese Natural Language Processing. In *Proceedings of the 1996 International Conference on Computational Linguistics (COLING 96)*, Copenhagen, Denmark.

Kirby, J. P. and Yu, A. C. (2007). Lexical and phonotactic effects on wordlikeness judgments in cantonese. In *Proceedings of the International Congress of the Phonetic Sciences XVI*, volume 13891392.

Landauer, T. K. and Streeter, L. A. (1973). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning and Verbal Behavior*, 12(2):119–131.

Lee, T., Wong, C., Leung, S., P., M., A., C., Szeto, K., and Wong, C. (1994). The Development of Grammatical Competence in Cantonese-speaking Children. Technical report, RGC, Hong Kong.

Lee, J. L., Chen, L., and Tsui, T.-H. (2015). PyCantonese: new perspectives on Cantonese linguistics. In prep.

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

Li, Z. and Sun, M. (2009). Punctuation as Implicit Annotations for Chinese Word Segmentation. *Computational Linguistics*, 39(4):505–512.

Luce, P. A. and Large, N. R. (2001). Phonotactics, density, and entropy in spoken word recognition. *Language and Cognitive Processes*, 16(5-6):565–581.

Luke, K. K. and Wong, M. L. (2015). The Hong Kong

Cantonese Corpus: Design and Uses. *Journal of Chinese Linguistics*, 25:312–333.

MacWhinney, B. (1995). *The CHILDES Project: Tools for Analyzing Talk*. Lawrence Erlbaum Associates, Mahwah, NJ, 3rd edition.

Magistry, P. (2013). *Unsupervised word segmentation and wordhood assessment : the case for mandarin chinese*. Ph.D. thesis, Université Paris Diderot – Paris 7.

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

Munson, B., Swenson, C. L., and Manthei, S. C. (2005). Lexical and phonological organization in children. *Journal of Speech, Language, and Hearing Research*.

Munson, B. (2001). Phonological pattern frequency and speech production in adults and children. *Journal of Speech, Language, and Hearing Research*.

Newman, R. S. and German, D. J. (2005). Life span effects of lexical factors on oral naming. *Language and Speech*, 48(2):123–156.

Roodenrys, S., Hulme, C., Lethbridge, A., Hinton, M., and Nimmo, L. M. (2002). Word-frequency and phonological-neighborhood effects on verbal short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(6):1019.

Snow, D. (2004). *Cantonese as a Written Language*. Hong Kong University Press, Hong Kong.

Sommers, M. S. and Danielson, S. M. (1999). Inhibitory processes and spoken word recognition in young and older adults: the interaction of lexical competition and semantic context. *Psychology and aging*, 14(3):458.

Sommers, M. S. and Lewis, B. P. (1999). Who really lives next door: Creating false memories with phonological neighbors. *Journal of Memory and Language*, 40(1):83–108.

Storkel, H. L. and Lee, S.-Y. (2011). The independent effects of phonotactic probability and neighbourhood density on lexical acquisition by preschool children. *Language and Cognitive Processes*, 26(2):191–211.

Storkel, H. L., Armbrüster, J., and Hogan, T. P. (2006). Differentiating phonotactic probability and neighborhood density in adult word learning. *Journal of Speech, Language, and Hearing Research*.

Storkel, H. L. (2004). Do children acquire dense neighborhoods? an investigation of similarity neighborhoods in lexical acquisition. *Applied Psycholinguistics*, 25(2):201–221.

Vitevitch, M. S. and Castro, N. (2015). Using network science in the language sciences and clinic. *International journal of speech-language pathology*, 17(1):13–25.

Vitevitch, M. S. and Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language*, 40(3):374–408.

Vitevitch, M. S. and Luce, P. A. (2016). Phonological neighborhood effects in spoken word perception and production. *Annual Review of Linguistics*, 2:75–94.

Vitevitch, M. S. and Rodríguez, E. (2005). Neighborhood density effects in spoken word recognition in spanish. *Journal of Multilingual Communication Disorders*, 3(1):64–73.

Vitevitch, M. S., Stamer, M. K., and Sereno, J. A. (2008). Word length and lexical competition: Longer is the same as shorter. *Language and Speech*, 51(4):361–383.

Wiedenhof, J. (2015). *A Grammar of Mandarin*. John Benjamins, Amsterdam.

Wong, T.-s., Gerdes, K., Leung, H., and Lee, J. (2017). Quantitative Comparative Syntax on the Cantonese-Mandarin Parallel Dependency Treebank. In *Proceedings of the Fourth International Conference on Dependency Linguistics*, pages 266–275, Pisa, Italy.

Yates, M., Locker, L., and Simpson, G. B. (2004). The influence of phonological neighborhood on visual word perception. *Psychonomic Bulletin & Review*, 11(3):452–457.

8. Language Resource References

Baayen, R.H. and Piepenbrock R. and L. Gulikers. (1995). *CELEX2*. Linguistic Data Consortium, 2.0, ISLRN 204-698-863-053-1.

Bigi, Brigitte. (2015). *SPPAS*.

Fletcher, P. and Leung, S. C-S. and Stokes, S. F. and Weizman, Z. O. (2000). *HKU-70*. 1.0.

Junyi, Sun. (2019). *Jieba*.

Lai, Regine and Winterstein, Grégoire. (2019). *CantoMap*. 1.0, ISLRN 167-857-138-471-9.

Lee, T.H.T. and Wong, C.H. and Leung, S. and Man. P. and Cheung A. and Szeto, K. and Wong, C.S.P. (1994). *Lee/Wong/Leung Corpus*. 1.0.

Leung, Man-Tak and Law, Sam-Po. (2001). *HKCAC: The Hong Kong Cantonese Adult Language Corpus*. Hong Kong University, 1.0.

Zhongguo Li and Maosong Sun. (2009). *Thulac*.

Luke, Kang Kwong and Wong, May L.Y. (2015). *HKCanCor: The Hong Kong Cantonese Corpus*. 1.0.

Boris New and Christophe Pallier. (2019). *LEXIQUE 3*. CNRS, 3.0.