Word and sound frequency in Cantonese: Comparisons across three corpora

Jane S.Y. Li, Simon Fraser University
Heikal Badrulhisham, Simon Fraser University
John Alderete, Simon Fraser University

Abstract.

This report gives detailed accounts of word frequency and sound structure frequency in three large corpora of Cantonese. Word frequencies across the corpora have similar structure in frequency rankings, but pairwise comparisons between corpora showed low lexical overlap and low correlation in frequencies for individual words. By contrast, sound structure frequencies are well-correlated, but nonetheless exhibit important differences due to the type/token distinction, orthographic encoding, word position, and speech genre. These differences inform psycholinguistic studies of Cantonese that include frequency as an experimental condition. In addition, we document the different methods used to segment words from running text, to encode words orthographically and phonologically, and to extract token and type frequencies from large data sets, thereby providing further access to the data. All of these generalizations are summarized in open data sets.

Keyworks: frequency norms, word frequency, consonants, vowels, syllables, tone, type vs. token, word segmentation, Cantonese, psycholinguistics

1. Introduction

It is difficult to overstate the importance of frequency in psycholinguistics. A range of core capacities in language comprehension are affected by frequency, including the effect of word frequency on lexical access (Forster & Davis, 1984; Gordon, 1983) and sound structure frequency on spoken word recognition (Vitevitch & Luce, 1999; Vitevitch et al., 1999). Language production processes are also clearly influenced by frequency, as shown by the impact of word and sound frequency on response latencies in picture naming (Bates et al., 2003; Levelt & Wheeldon, 1994; Oldfield & Wingfield, 1965) and the structure of speech errors (Dell, 1990; Levitt & Healy, 1985; Stemberger & MacWhinney, 1986). Since language acquisition involves both comprehension and production, it is not a surprise that frequency effects also pervade language acquisition studies (Ambridge et al., 2015). In all of these domains, experimental setups routinely distinguish high frequency versus low frequency items (e.g., Vitevitch (1997), Goldrick et al. (2011)), and frequency is also extensively used as a continuous variable that is correlated with behavioral measures (e.g., Ernestus and Baayen (2003), Gahl (2008)).

The successful analysis of language processing in any language therefore depends on a solid understanding of frequency effects. Our ability to examine the impact of frequency on language processing in major Indo-European languages like English and Dutch is strong because of the existence of rich data and analysis of these languages (e.g., Baayen et al. (1996); Kessler and Treiman (1997); Roland et al. (2007)). However, other lesser-studied languages are in comparatively weaker positions (Jaeger & Norcliffe, 2009). Our primary aim here is to build on existing accounts to provide a range of practical information about word frequency and sound

frequency in Cantonese, an under-studied Chinese language of Hong Kong, southern China, and the Cantonese-speaking diaspora.

Cantonese sound structure has been investigated in some detail in Leung et al. (2004). This work reported on the frequency of onsets, rimes, and tone in the Hong Kong Cantonese Adult Corpus ("HKCAC" henceforth) (Leung & Law, 2001), and gives the first detailed account of these phonological structures in spontaneous speech. An important aspect of this study is that it is based on surface phonological structures. These structures are phonological forms of words that result from one of many phonological processes, including assimilation, reduction, and other casual speech phonology. Surface representations are distinct from lexical representations, which are more abstract lexical structures that do not necessarily include the results of phonological processes. As the authors note, surface phonological representations provide a wealth of information concerning language development, the relationship between surface and lexical representations, and applied domains such as automatic speech recognition.

Despite its importance, the Leung et al. (2004) account of frequency effects has two design features that limit its applications to the psycholinguistic study of Cantonese. First, the focus on surface representations means that the reported frequency norms are not accurate counts of the sound structures of lexical representations. While surface representations are important to mapping acoustic structures to lexical representations, it can be argued that lexical representations are more central to accounts of language processing. Contemporary models of speech production generally posit lexical representations rather than surface representations in the inter-connected networks of word-forms that underlie language production processes (Dell et al., 2014; Levelt et al., 1999). Standard models of spoken word recognition also posit abstract lexical representations and encode processes of activating and selecting these representations as the basic processes underlying lexical access (Luce & Pisoni, 1998; Marslen-Wilson, 1984). Recent research has supported the role for these more abstract representations by showing how the distance between surface and lexical representations affects visual processing (Farris-Trimble & Tessier, 2019). Moreover, while recent exemplar models of word recognition store lexical knowledge as multiple traces of heard tokens rather than a single canonical form (Goldinger, 1998; Keith Johnson, 1997), these models nonetheless require mechanisms for selecting specific tokens in lexical access, which again abstracts away from specific instances of a single word. Therefore, in order to engage in these research paradigms, language processing in Cantonese also needs frequency norms from deeper lexical representations.

A second problem stems from the way type frequencies (i.e., frequencies in the lexicon rather than a corpus) was calculated in Leung et al. (2004). The authors calculated type frequencies from Chinese characters rather than the more standard technique of using words as the basis for typing. Words, which are not co-extensive with characters, are the conventional linguistic unit in calculating type frequencies because they support greater cross-linguistic comparison and allow for observations that are not possible with characters (Atkins et al., 1992). For example, it is not possible in the Leung et al. (2004) account to give type frequencies of tone in different positions in a word, since tone is associated with syllables, and characters are almost always a single syllable in Chinese languages. Type frequency is tremendously important to understanding psycholinguistic processes (Hay et al., 2004; Levitt & Healy, 1985), but both the nature of the representations (surface rather than lexical), and the non-standard way of calculating them in Leung et al. (2004), render the reported type frequencies less suitable for analyses of language processing.

We endeavour here to address these problems by further investigating sound frequencies in HKCAC. In addition, we document word and sound frequencies and two other data sets to broaden the empirical coverage. In particular, we report on patterns in the Hong Kong Cantonese Corpus ("HKCanCor" henceforth; Luke and Wong 2015) and the IARPA Babel Cantonese Language Pack (henceforth "IARPA"; Andrus et al. 2016) distributed by the Linguistic Data Consortium. We describe the methods used to extract word frequencies from these corpora, and report word frequencies for the first time from these sources. We also report on token and type frequencies of sound structures in all three corpora, and document similarities and differences across corpora. These investigations address the limitations discussed above by documenting lexical representations and type frequencies of sound structures based on word frequency. In addition, they provide the opportunity to examine in more detail the Cantonese syllabary, consonant and vowel structure, tone relative to word position, and the impact of speech genre on word and sound frequency. Our results provide a standard stock of frequency norms tailored to psycholinguistic analysis (https://github.com/jane-lisy/cantfreq). They also show that there are many important differences among these corpora that psycholinguists need to attend to before selecting an appropriate data collection for baseline data.

The rest of this article is structured as follows. We introduce our methods in section 2 by first explaining the linguistic structures we investigate, and reviewing and selecting Cantonese language corpora suitable for our study. We then describe the methods of segmenting words from running text in these corpora and extracting word and sound frequencies. Section 3 reports on word frequencies across corpora, summarizing frequencies in the shared lexical structure, and highlighting important differences. Section 4 gives a detailed account of a range of sound structures, including syllabic and sub-syllabic structure, consonant phonemes, vowel phonemes, tone, and phonotactics. The last section discusses some of the recurring themes of the two sections, summarizes some of the linguistic insights that can be gleaned from the results, and gives a set of recommendations about how to use the three corpora in psycholinguistic studies.

2. Methods

2.1 Phonological structures

Sound structure in Cantonese can be described at three different levels. At the segmental level, Cantonese speech is a stream of consonant and vowels. At the syllabic level, these phonological segments are organized into, syllables, or natural groupings of consonants and vowels that commonly recur in the language. In addition, Cantonese speech has the suprasegmental level, tone, or the characteristic pitch shapes that are associated with syllables, but are functionally independent of them.

Traditionally, these distinct levels are anchored in the syllable, which is structured as follows in Cantonese: (C) X_1 (X_2). The initial (C) is an optional onset slot that can be filled with one of 19 phonemes (i.e., contrastive sound units) or left empty. Broken down by manner class, the onset can be filled by stop sounds (*p* p^h *t* t^h *k* k^h k^w k^{wh}), fricatives (*f s h*), affricates (*ts* ts^h), nasals (*m n ŋ*), or approximants (*l w j*).

The (C) X_1 (X_2) syllable template, minus the onset, is traditionally called the rime. The X slots in the rime can be filled by either consonants or vowels. Open syllables can be formed by filling X_1 with one of seven monophthongal vowels (*i e y œ a: o u*) and leaving X_2 empty, or by combining a vowel in X_1 with a high vowel in X_2 to form one of the eleven diphthongs (*ei œi ɐi*

aːi oi ui, iu eu ɐu aːu ou).[1] Closed syllables, on the other hand, can be formed by combining a vowel in X_1 with a nasal or unreleased voiceless stop in X_2, as in *-am* or *-it*. There are a number of gaps in the combination of X_1 and X_2 in rimes, as shown below in Table 1. For example, the short central low vowel *ɐ*, the short counterpart to long *aː*, is restricted to the first position of a diphthong and closed syllables. In addition, some rimes are marginal (given in parentheses), either because they are rare or limited to specialized constructions, like onomatopoeic speech, as with the rime *-em*.

Table 1. Attested rimes in Cantonese.

	i	**e**	**y**	**œ**	**ɐ**	**aː**	**o**	**u**
V	i	e	y	œ		aː	o	u
V+i		ei		œi	ɐi	aːi	oi	ui
V+u	iu	(eu)			ɐu	aːu	ou	
V+m/p	im	(em)			ɐm	aːm		
V+n/t	in	(en)	yn	œn	ɐn	aːn	on	un
V+ŋ/k	iŋ	eŋ		œŋ	ɐŋ	aːŋ	oŋ	uŋ

Finally, in a small number of morphemes, syllables can be composed of a syllabic nasal *m̩* and *ŋ̩*, which fills the X_1 position, as in negative marker [m̩21] ‘not’. Syllables with syllabic nasals do not have onsets or codas.

Cantonese is a tone language, meaning that tone can signal a difference in meaning in otherwise identical words. Modern Cantonese has six tones, shown in Table 2. Tones in these examples are transcribed with Chao tone digits (suffixed to syllables), which approximate the surface pitch shapes (Chao, 1930). The six tones can be cross-classified by tone height (high, mid, low) and contour (level, rising, falling).

Table 2. The six tones of Cantonese (Matthews & Yip 2011: 27).

High level	55	憂 jɐu55 ‘worry’
High rising	25	油 jɐu25 ‘paint’
Mid level	33	幼 jɐu33 ‘thin’
Low falling	21	油 jɐu21 ‘oil’
Low rising	23	有 jɐu23 ‘have’
low level	22	又 jɐu22 ‘again’

The three level tones have “allotones” in so-called checked syllables ending in unreleased *p t k* that are shorter in duration than their counterparts in non-checked syllables. Some speakers also have a high falling [53] tone that is either in free variation with [55] (common in older speakers from Hong Kong) or contrastive with it (as in Guangzhou Cantonese), though this tone

[1] Our transcription of vowels is phonemic and intended to avoid the potential confusion created by including the following allophonic details. Monophthongs in open CV syllables are longer in duration and sometimes written with “ː”. The high vowels are generally long, but have /u i/ lax counterparts [ʊ ɪ] in syllables closed with a velar. The mid vowels /e o œ/ are generally realized as long [ɛː ɔː œː], except in V1 of a diphthong, as in [ei ou øi]; [øi] is sometimes written [øy], reflecting another practice of sometimes writing V2 as a consonantal glide. /o/ is also [ø] before alveolar coda consonants.

is rare among younger speakers. Acoustic studies of Kong Hong Cantonese have also revealed a change in progress in which some speakers do not discriminate between the rising tones [25]/[23], the level tones [33]/[22], and [21]/[22], in production and perception tasks (Bauer et al., 2003; Mok et al., 2013). These three tonal phenomena are not represented in the corpora we examine, so we do not investigate them further.

There are many different phonetic systems for transcribing Cantonese sound structure, with no clear consensus. This lack of consensus is also found in the corpora we investigate, though to be fair, their coding principles are designed for textual searches, not ease of reading or linguistic insight. As with the illustrations above, we will use the IPA (International Phonetic Alphabet) and Chao tone digits throughout for consistency (though, as explained in footnote 1, we abstract over certain vowel allophones to avoid confusion). Appendix B gives the correspondences between IPA and two commonly used phonetic systems, Yale romanization and Jyutping (the latter is the phonetic system developed by the government of Hong Kong and the Linguistic Society of Hong Kong). The appendix also gives the corresponding sounds and tones for the three main corpora we investigate here, namely HKCAC, HKCanCor, and the IARPA corpus. See Bauer & Benedict (1997: 471) for correspondences with several other phonetic systems, including the specific transcription system used in this authoritative work.

2.2. Corpora

We reviewed 10 major Cantonese language corpora created in the past 40 years (see Table 3) to identify data collections suitable for our research focus. As our goal is to investigate language usage in adult spontaneous speech, we excluded five speech corpora built from child language acquisition research or other projects involving pre-planned speech (the first five projects in the table). Pre-planned speech is different from spontaneous speech because it involves more reading and less free expression of ideas, and so it invokes different psycholinguistic processes. Of the remaining five corpora, Xu and Lee (1998) and the PolyU Corpus of Spoken Chinese are comparative corpora that compare Cantonese with other Chinese languages, like Shanghainese or Mandarin. While these corpora do contain some spontaneous language data, a large percentage of the data sets are constrained to specific criteria required to make comparisons across the languages, and so are not well-suited to our needs.

The three remaining corpora are large data collections of adult natural speech. The Hong Kong Cantonese Adult Corpus has over 170,000 syllables of transcribed speech, and is the primary data source for Leung et al. (2004), the first rigorous account of sound structure frequencies in Cantonese.[2] It is gathered from natural conversations and phone-in radio programs on a variety of topics, and so it is mostly composed of natural unscripted speech. It is accessed through the FileMakerPro database program, which represents the corpus both as a list of syllables and as a list of sentences. A unique property of this corpus is that it gives detailed phonetic transcriptions that have surface phonological structure, that is phonological structures after the application of phonological processes.

The Hong Kong Cantonese Corpus is a similar corpus built from spontaneous speech in radio call-in shows and other phone conversations (230,000 characters from approximately 30 hours of speech, including 52 in-person conversations and 42 radio conversations, most of which were two- or three-party conversations). The corpus is segmented at both the sentential and word level, and each word is a structured representation tagged for its orthographic form, phonological

[2] We are grateful to the authors of this data collection for making the data available to us.

form (in Jyutping), and part of speech. Thus, while the corpus has phonological representations that can be investigated, it is unlike HKCAC in that it does not show surface phonology. A useful aspect of this data collection is that it can be accessed through a Python package developed for it, PyCantonese (J. L. Lee, 2015), which supports quick and easy searches of structured linguistic data.

The IARPA Babel Cantonese Language Pack is the largest data collection, based on over 200 hours of scripted and spontaneous telephone conversations in Cantonese spoken in China (in particular, Guangdong and Guangxi). Consistent with our focus on spontaneous speech, we only investigated the unscripted speech in this data set. Though the IARPA language pack itself was built for the development of speech recognition technology, it is comparable to HKCAC and HKCanCor because it is composed of spontaneous conversations with many different adult speakers. However, the Cantonese of IARPA is from different dialect groups (central Guangdong, northern Guangdong, northern and southern Pearl River Delta, Guangxi, and western Guangdong) than those of HKCAC and HKCanCor, which focus primarily on Hong Kong Cantonese. The authors note that there are differences in lexical choice and pronunciation among the dialect groups. The extent of these differences can be assessed below, at least in part, by comparing the frequency data of IARPA relative to HKCAC and HKCanCor. The corpus itself is a set of .txt files in which words are segmented by spaces and indexed with a time stamp, and larger sentence breaks can be inferred from pauses. Each recording has two separate files, one with an orthographic transcription and another with a phonetic transcription that was transcribed with phonemic representations.

Table 3. Cantonese language corpora.

Project (Authors)	**Description**
A Linguistic Corpus of Mid-20th Century Hong Kong Cantonese (Chin & Tweed, 2019)	A corpus of approximately 60 Cantonese movie dialogues in the mid-20th century, intended for a diachronic analysis of Cantonese. Size: ~800,000 characters.
CHILDES (V. Yip & Matthews, 2007)	A longitudinal database of eight Cantonese-English bilingual children. Intended to investigate bilingual acquisition in infants.
The Hong Kong Cantonese Child Language Corpus (T. H.-T. Lee & Wong, 1998)	A diachronic corpus of eight children (age 1-3) documented over the span of a year.
HKU-70 Corpus (Fletcher et al., 2000)	70 transcribed audio files of ~20 minute interviews with pre-schoolers. Intended to investigate syntactic and lexical forms of children.
Hong Kong spoken Cantonese database (So, 1992)	A database of native speakers of Hong Kong Cantonese pronouncing syllables of Cantonese. Size: ~1,800 syllables.
Xu and Lee (1998)	Transcriptions of Cantonese, Shanghainese, and Mandarin plays, television shows, news broadcasts, and unstructured interviews.
PolyU Corpus of Spoken Chinese (Hong Kong Polytechnic, 2015)	28 transcribed audio recordings of conversations, debates, and phone-in radio shows in Cantonese and Mandarin.
Hong Kong Cantonese Adult Language Corpus (Leung & Law, 2001)	Audio transcriptions of spontaneous radio phone-in programs. Size: ~170,000 Chinese characters.
Hong Kong Cantonese Corpus (Luke & Wong, 2015)	A collection of transcribed spontaneous conversations and radio phone-in programs. The corpus has been segmented by part-of-speech. Size: ~230,000 Chinese characters.
IARPA Babel Cantonese Language Pack (Andrus et al., 2016)	A collection of spontaneous and scripted telephone conversations of Cantonese speakers in Guangdong and Guangxi, China. Intended for speech recognition training. Size: ~215 hours of audio.

There are some differences among these corpora, including regional differences, some of

the conversational formats, and the level of phonological and phonetic detail, that we attend to in our searches below. However, these differences are overshadowed by the similarities among them in the use of adult speech, the unscripted spontaneous nature of the speech, and their relatively large sizes. Perhaps more important are the differences in encoding language in the corpora: what constitutes a word, how sounds map to the IPA, and how the representation of filler words and reduced forms are not completely consistent. In the next section we outline our methods for reducing the impact of these representational differences by attempting to standardize word segmentations and the representation of sound sequences.

2.3. Word segmentation

What counts as a word is a complex problem in Chinese languages (Packard, 2000), and the specific methods used for extracting words from running text and speech, can have an impact on the counts we examine here. These methods affect word frequencies because different segmentations result in different counts for specific words, and they also affect the frequencies of sound structure because segmentation can affect both its identity, as well as the position of the sound within a word. We review the segmentation methods used to segment two corpora, HKCanCor and IARPA, and our methods for segmenting HKCAC, so that any discrepancies among the corpora can be investigated as a consequence of these methods.

The texts in HKCanCor are structured data that include part of speech tags, so they are already segmented. The creators have in part followed the annotation scheme used in a standard written corpus of Mandarin, namely the People's Daily POS Tagged Chinese Corpus (see Zhan et al. (2006)). HKCanCor makes use of the tagset of 26 part of speech categories, but the creators of this corpus expanded that tagset to 46 categories to account for new structures in this corpus. The details of how words are segmented and tagged are not fully documented in the source material, but can be summarized as follows. A small subset of the corpus was first manually segmented and POS-tagged and then used as the training corpus to process the remainder of the corpus following the procedures given in Fu et al. (2005).

Word boundaries are encoded as white space between words in IARPA, as in written English, but unlike standard Chinese language orthography. The specific methods are not explained in the documentation, but they are described as following a set of proprietary principles developed by the software company Appen. In addition, the authors note that their segmentation follows standard linguistic practice of recognizing compounds, affixed words, and tense-aspect markers as single words, though their treatment of reduplicated words is non-standard (see below). The segmentation was generated automatically, but phonetically-rich material (e.g., deviations from standard pronunciation) was manually verified. Because of the lack of transparency, we did a step sample of the lexicons of both HKCanCor and IARPA, checking for segmentation errors every 20th word and for the existence of the corresponding words in the other data set, and found a high degree of consistency. The only difference we found was that IARPA assumes that reduplicated words separated by a character are words, but HKCanCor does not. For example, the very first word of the IARPA corpus is 聽唔聽到 'can hear or cannot hear' [t^{h}eŋ55m21t^{h}eŋ55tou25], which IARPA treats as a single word, but HKCanCor treats as three: 聽, 唔, 聽到. Our searches of the IARPA data set indicates that approximately 0.4% (*N*=3,680) of the larger data set have this reduplicated structure. Thus, while the difference in segmenting words will lead to increased tokens and reduced type counts in HKCanCor, the effects will be relatively minor.

HKCAC is not segmented into words (though one section of the database segments by syllable/character). For our investigation, the corpus was segmented using two separate methods and then compared for consistency. First, the corpus was parsed to detect unique words not found in other sample lexicons (i.e., the lexicon built from HKCanCor and an electronic version of Huang dictionary (Huang 1970)). The unique words uncovered at this stage were then incorporated into an expanded reference lexicon that combined new and old words. This parsing process is inspired by the stick-by-longest-matching segmentation strategy documented in Fung and Bigi (2015). However, unlike Fung and Bigi's fully-automated procedure, we manually parsed sentences when the automated segmentation was unable to match all characters in the sentence into a set of words. The second parse was done automatically through jieba, a Python package built for automatic parsing in Chinese languages (Sun, 2020). We imported the lexicon from the first parse, and sampled sentences of various lengths to ensure that the parse was performed accurately. The code for the two phases of segmentation is available on the project's GitHub page.

A final step for all corpora involved excluding English code-switched words (though not English loans or loans from other languages). Punctuations and interjections were also omitted, consistent with our focus on Cantonese lexical items.

2.4. Methods for extracting data from corpora

We explain below the data processing used to extract lexicons from corpora, and then generate counts of phonological structures in the corpora and lexica. The three corpora are structured as lists of words, or an ordered list of all the tokens produced. Each token in the list of words is connected to its phonological form in a structured representation, which documents the specific phonological variant of the token. Since we are only interested in Cantonese utterances, we removed words that are orthographically encoded as English words (e.g., code-switched English words like *trial* – [tʰraɪəl] in HKCAC). However, some Cantonese words, including adapted loans, can lack a standard orthographic representation with a Chinese character (e.g., [liu55lɐŋ55] 'blathering' in HKCanCor), and these words were retained. In sum, each list of words contains three principal word classes: regular (dictionary found) Cantonese utterances, Cantonese colloquial words (which may or may not have corresponding characters), and adapted English loanwords.

The following procedures were employed to generate the token and type frequencies at the word level. The list of words for each corpus was used to generate all token frequencies, such as word size counts and tallies of word tokens in a corpus. Word types, for the purpose of calculating the type frequency of sound structures, were tallied algorithmically by applying the Python function `collections.Counter()` to the list of words, which returns the unique set of words with their respective counts in the corpus. Part of speech tags are important to distinguishing word types, but the available data leads to a problem as to how to use them. HKCanCor has POS tags as part of their structured representations and can be used to distinguish word types, but the other two corpora do not. Because all corpora use POS tagging in some way to distinguish word types, we assume that it is best to be faithful to the raw data in HKCanCor, and that this tagging is broadly consistent with the POS labels used for word segmentation in HKCAC and IARPA. Though outside the scope of this project, future studies may employ the same POS tagging methods for all corpora to ensure complete consistency.

Words in the corpus have variant phonological realizations, which are part of the structured representation of a word. The fact of phonological variation required us to make

certain assumptions about word typing. In the process of word type generation, we assume that words that are equivalent orthographically but distinct phonologically are treated as a single type. For example, the word 但係 'but', which is standardly pronounced as [ta:n22hɐi22] but sometimes reduced as [ta:22əi22], is counted as one word in our tabulation. The reasoning behind this assumption is that we seek to document lexical representations, or the phonological forms before phonologically processes that are triggered by neighboring words have applied. While [ta:22əi22] is a distinct form from [ta:n22hɐi22], we are interested in the lexical representation of the word 但係, which does not distinguish between the two possible pronunciations.

The requirement that there is a single type for a word leads to the problem of determining which variant represents the lexical representation. Like many Chinese languages, Cantonese does not have a large set of synchronic phonological processes that lead to alternations in surface forms (Pulleyblank, 1997), so the problem of variation is minor compared to other languages. However, it does have a well-defined set of casual speech rules that create surface phonological variation (Bauer, 2013; Bauer & Benedict, 1997), and one does occasionally encounter words with more than one phonological variant. How should we select the phonological variant representing the lexical representation for the purposes of calculating frequencies of the sound structure inside these representations? A conceptually simple way would be to construct lexical representations by working with surface phonological representations, and undoing the assumed phonological processes. This approach is problematic, however, because Cantonese has many words with surface phonology that have supplanted the representations that would be the input to this phonology. For example, Cantonese has a rule deleting [ŋ] in onset position, as in /ŋɐu21/ → [ɐu21] 牛 'cow', but the form [ŋɐu21] is very rare in colloquial speech, suggesting that it has been reanalyzed as [ɐu21] in the present-day lexicon.

Instead of a blanket inversion of phonological rules, we propose to construct lexical representations by selecting the most frequent variant. This approach both addresses the problem of lexical re-analysis sketched above, and is consistent with behavioral research showing the importance of frequency in selecting canonical lexical representations (Connine et al., 2008; Pitt et al., 2011). Concretely, the variant representations are tallied algorithmically in the list of words (where the structured representations encode phonological variants), and the most frequent variant is selected as the lexical representation. We can illustrate with a coda neutralization process that reduces a [k] in the syllable coda to either [t] or the glottal stop [ʔ], or deletes the segment altogether. For the word 百'hundred', [pa:k33] was chosen as the lexical representation because it is by far the most common form (n=68) compared to [pa:t33] (n=9) and [pa:33] (n=1). We have investigated this particular pattern of variation in our corpora, and found that extensive variation is rather uncommon, and that a frequency-based selection process almost always prefers the [k] variant. Combined with the psycholinguistic support discussed above, this confirms that frequency is a successful way of identifying lexical representations.

This focus on lexical representations also enabled us to establish a common stock of sound structures in the three corpora. Recall that HKCAC is transcribed phonetically, which includes details of phonetic variants. While this level of detail is useful to the analysis given in Leung et al. (2004), we cannot use the phonetic transcriptions from HKCAC to compare it with other corpora because the others lack this detail. In addition, lexical representations are the primary data for type frequency, so a focus on these structures gives a more accurate characterization of frequency in the lexicon.

Counts of syllabic and sub-syllabic structures were created as follows. The phonological representations of all words were parsed and structured as [onset, nucleus, coda, tone] per syllable (see supplemental code). We apply the same process to the unique set of words, which allow sub-lexical and sub-syllabic information to be extracted by both type and token. In our analyses in section 4, counts of sound structures, such as onsets, rimes, or entire syllables, are done combinatorically, to give an accurate count of the common stock of sound structure for all three of the corpora that is supported by contemporary research (Bauer & Benedict, 1997; Leung et al., 2004). For example, the syllable [kei55] will be a part of the query: onset = "k", 'rime = "ei", tone = "55". These searches will not uncover the frequency of reduced forms or the output of certain phonological rules, as in Leung et al. (2004), which can be referenced for such information.

Our methods for generating type frequencies described also enabled us to address a problem in the calculation of type frequency in HKCAC. Leung et al. (2004: 502) used characters as basic units for establishing types rather than the method used here based on words. As an illustration of the difference, in the set of words 傷心 'sad' and 心臟 'heart', our word-based typing has two types, one for each word, whereas Leung et al. (2004) has three types, one for each of the characters inside these words, 傷, 心, and 臟. These different types correspondingly lead to different counts in type frequencies. While character information is important for psycholinguistic inquiry, the correspondences between Cantonese character, morpheme, and words are unpredictable, which makes it harder to use the data set as the basis of other studies. The use of characters for typing also complicates the comparison of other languages that do not use characters, and as well as Chinese languages that use word typing (Atkins et al., 1992; Bird et al., 2009). Additionally, our analyses with words as units have provided us with more nuanced insight, such as the differences in tone distributions across word positions (see section 4.5), which would be unattainable in a character-based analysis. In summary, our segmentation of the HKCAC corpus therefore enables us to give word-based type frequencies, which is both more conventional and more useful.

3. Word frequency across corpora

This section describes the distribution of words in the three corpora we investigate, as well as in the lexicons derived from the corpora. It documents word frequency in some detail, an important measure in many psycholinguistic investigations.

First, we explore the relationship between corpus size and the size of the lexicon derived from a corpus (as opposed to a comprehensive lexicon of the language in general). As shown below, IARPA is considerably larger than the other two corpora (even after our exclusion of scripted text). It is roughly eight times larger than HKCAC and seven times larger than HKCanCor. The lexicon of unique word types in IARPA is also larger, roughly two and a half times larger than the other two lexicons. Corpora also differ in the ratio of corpus size to lexicon size: words on average occur more frequently in IARPA than the other two. Lexical diversity, the inverse of this ratio (Johansson, 2009), is correspondingly smaller for IARPA.

Table 4. Corpora and lexica sizes.

	Words in corpus	Words in lexicon	Lexical diversity	Corpus/lexicon
HKCAC	99,344	7,726	0.07777	12.86
HKCanCor	119,977	7,449	0.06209	16.11
IARPA	830,745	19,575	0.02356	42.43

Despite these differences, all three corpora have similar distributions of high frequency items relative to low frequency items. Figure 1 displays word frequency in the three corpora, with the frequency ranking from high to low (left to right) on the *x*-axis and token frequency (log scale) on the *y*-axis. All three corpora appear to have a Zipfian distribution (Zipf, 1949), whereby a relatively small number of lexical items have very high frequencies, and their relative frequencies drop very quickly and start to level off, as shown by the fit of the curve on the red line marking the 97th percentile (i.e., all words to the left of the line are in the top 3% in frequency rank.

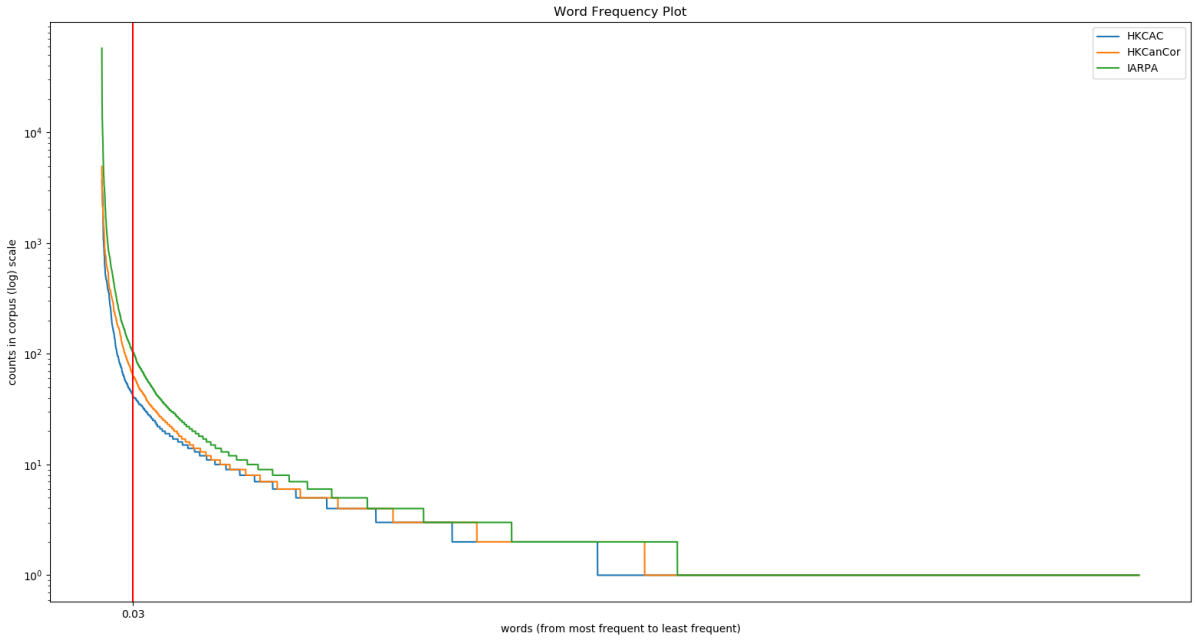

Figure 1. Word frequency for three corpora.

Another way to investigate similarities across corpora is to examine specific lexical items at the high end of the frequency spectrum. In Table 5, we list and order by rank the top 12 words in each corpus. The Zipfian-like distributions suggested above would seem to predict similar frequency rankings across corpora for specific lexical items in this list. For example, since the word with the highest frequency is predicted to be so much higher than all others, it should in principle be highest in all corpora. However, we do not find that in our comparisons, and indeed, the top two items in HKCAC, as well as the third and fourth ranked items, are nearly tied in frequency, contra Zipf's Law. There is some common ground, however, in that seven of the top 12 are shared across all corpora (color-coded below): these include the three personal pronouns (1st person 我, 2nd person 你, and 3rd person 佢), the negator 唔, the sentence final particle 啊, the

modifier 咁, and the predication marker 咁; in fact, all of the 18 morphemes listed in Table 5 are function words.[3]

Table 5. Top 12 lexical items by frequency in each corpus.

HKCAC			HKCanCor			IARPA		
character	IPA	counts	character	IPA	counts	character	IPA	counts
呢	li55	3664	係	hɐi22	4936	啊	a:55	57683
係	hɐi22	3634	啊	a:55	3540	係	hɐi22	33609
唔	m21	3052	我	ŋo23	2668	你	nei23	26036
嘅	ke33	3028	你	nei23	2535	我	ŋo23	24395
啊	a:55	2905	佢	kʰœi23	2224	佢	kʰœi23	17820
咁	kɐm25	2785	都	tou55	2149	唔	m21	17050
我	ŋo23	2589	呢	li55	2134	咁	kɐm25	16946
你	nei23	2449	咁	kɐm25	2102	個	ko33	13397
佢	kʰœi23	2171	唔	m21	1944	喇	la:33	13161
即	tsek55	1718	㗎	ka:33	1779	都	tou55	12973
就	tsɐu22	1459	就	tsɐu22	1759	冇	mou23	11615
喇	la:33	1140	即係	tsek55hɐi22	1632	有	jɐu23	11161

Comparisons based on raw frequencies are difficult, because the corpora differ in size. By converting frequencies to probabilities, we can standardize the data and make comparisons at various positions in frequency rank (Gries, 2015). Thus, $P(x)$ denotes the probability of a lexical item x in a corpus, and it is calculated as the count of x divided by the total number of tokens in the corpus. For instance, the most common word in HKCAC, 呢 (a sentence final particle) occurred 3,664 times in the corpus with 99,344 items, so P(呢) = 3664/99344 = 0.0369. As shown in Table 6, the probabilities of the first and second ranked items are rather different across corpora, because IARPA's two top-ranked items have rather high probabilities relative to the other two corpora. However, these differences become less exaggerated as we move down the list to the 12th ranked item and the median. Table 6 gives two additional counts capturing facts at opposite ends of the frequency spectrum: the hapax legomena (% H. L.), or the percentage of the corpus made up of words that only occur once in the corpus, and %Top 1-6, the percentage of the corpus made up of the six most frequent items. IARPA differs from HKCAC and HKCanCor in that it has the largest percentage for % Top 1-6, and, correspondingly, the lowest % H. L., presumably due to the sparser lexical diversity of IARPA (see above).

[3] The meaning of the function words in Table 5 are as follows: [li55] 呢, sentence final particle (question, rhetorical); [hɐi22] 係, predication, 'yes'; [m21] 唔, negation; [ke33] 嘅, possessive/adjective linker; [a:55] 啊, sentence final particle (declarative); [kɐm25] 咁, '-ly'; [ŋo23] 我, 1st singular pronoun; [nei23] 你, 2nd singular pronoun; [kʰœi23] 佢, 3rd singular pronoun; [tsek55] 即, 'then, namely'; [tsɐu22] 就, 'then; [la:33] 喇, sentence final particle (exclamation, after a list of examples); [tou55] 都, 'also'; [ka:33] 㗎, sentence final particle (modal); [tsek55hɐi22] 即係, 'then it is, which is to say'; [ko33] 個, generic classifier; [mou23] 冇, 'don't have'; [jɐu23] 有, 'exist, have'.

Table 6. Probabilities at different frequency ranks and percentage occurrence of hapax legomena and the top six items.

	P(1st)	*P*(2nd)	*P*(12th)	*P*(median)	% H.L.	% Top 1-6
HKCAC	0.0369	0.0366	0.0115	0.000010	4.00	19.19
HKCanCor	0.0411	0.0295	0.0136	0.000010	2.96	15.05
IARPA	0.0694	0.0405	0.0134	0.000002	1.05	21.26

Another way to make comparisons between two corpora is to investigate a set of words that both corpora have in their lexicons, and ask how well correlated the shared items are in frequency (Kilgarriff, 2001). To this end, we constructed three lists of shared items across the three possible corpora comparisons, and examined the correlations (Pearson's *r*) within each list between the frequency of a particular item in one corpus and its frequency in another. In order to match items, we had to first convert the IARPA entries to traditional Chinese characters, which we did with the Python package HanziConv (Yue, 2016), so that they could be matched with entries in HKCAC and HKCanCor, which use traditional Chinese characters. The counts of matched lexical items are shown Table 7 and correlations in relative word frequency (i.e., probabilities) between these shared items are given in Table 8.[4] The first observation is that the three lexicons do not overlap very much. For example, HKCAC and HKCanCor have lexicons with 7,726 and 6,259 words (for stacked words) respectively, but only 2,387 shared items between them, or an overlap of roughly 30.90% (of HKCAC) and 38.14% (of HKCanCor). The second observation is that, among shared items, corpora are well-correlated, but much less so for HKCAC and IARPA. It is difficult to assess precisely why IARPA has a lower correlation with HKCAC, and not with HKCanCor, since both document the speech of Hong Kong Cantonese in a similar register. However, the differences are important enough to suggest that language researchers need to attend to this difference when using these data sets to analyze language processes.

Table 7. Shared items in lexicons of three corpora (% of total lexical items in corpus in row, column).

	HKCAC	HKCanCor
HKCAC		
HKCanCor	2,387 (38.14, 30.90)	
IARPA	3,064 (15.65, 39.66)	3,376 (17.24, 53.94)

Table 8. Correlation coefficients for relative word frequency for shared lexical items.

	HKCAC	HKCanCor
HKCAC		
HKCanCor	0.7440	
IARPA	0.6244	0.8520

Psycholinguistic research is also interested in groups of lexical items, and frequently makes the distinction between 'high frequency' and 'low frequency' items in experimental stimuli. To assess the common ground in these groupings, we examined the items that occurred in the shared lists, and binned them into 'high', 'mid', and 'low' frequency groups based on their

[4] While we do use POS tags for creating the HKCanCor lexicon (see section 2.4), we depart from that practice here, and leave words "stacked" together (i.e., undifferentiated by part of speech), because otherwise we cannot compare HKCanCor with other corpora. The percentages reflect the stacked words in HKCanCor.

frequency rank in each corpus (i.e., the top third is 'high', middle third 'mid', and bottom third 'low'). This meant that, though two items are shared in the lexicons, they could be in any of the three frequency groups because they were assigned to a group independently based on their rank in each corpus. Table 9 gives the percentage of shared items that match in frequency groups for each comparison. The results show that the best matches in all comparisons were for high frequency items, and that the best overall matches are between HKCanCor and another corpus, which is consistent with the correlations reported above. We also note that some of the mid frequency categories may not be significantly above chance levels (33.33%), and so assignment of these labels to lexical items should be taken with a grain of salt.

Table 9. Percentage of shared items matching in 'high', 'mid' and 'low' frequency groups.

Comparison	High	Mid	Low
HKCAC, HKCanCor	66.33	47.68	63.73
HKCAC, IARPA	61.64	41.13	55.53
HKCanCor, IARPA	68.53	43.87	56.09

The above results can be used to inform psycholinguistic research that uses large data sets to answer questions about how word frequency impacts language processing. If the breadth of a lexicon is important, then the lexicon based on IARPA is by far the largest. The frequency distributions of the lexicons of IARPA and HKCanCor seem to be well-correlated, and if these two large data sets are more representative, then either of them is probably a good choice in terms of assigning word frequency values to individual items. Perhaps the most important finding is that, though all corpora are relatively large, they are all unique and characterized by different frequency distributions, especially in the regions between high and low frequency groups. Subsequent research may therefore benefit from conducting separate analyses with more than one data set, to investigate a question with two or more independent tests.

4. Sound structure frequencies

4.1. General overview

We report below on the token and type frequencies of sound structures, investigating all of the sound categories introduced in section 2.1. We start at the syllabic level and work our way down to sub-syllabic structures. As with word frequencies, we are interested in looking across corpora to see how well the corpora are correlated. In addition, we investigate differences between token and type frequency, as well as new linguistic structures that have not been explored in past accounts.

4.2 Syllabic and subsyllabic structure

We begin with tallies of the size of words in terms of syllables. As shown in Table 10, for token frequencies, monosyllabic words are by far the most numerous, and frequencies fall steeply in successively larger polysyllabic words in all corpora. This pattern of descending frequency is likely due to the relatively high frequency of monosyllabic grammatical morphemes, like the personal pronouns (see section 3), because this trend is not repeated in lexical frequencies. As shown in Table 11, disyllabic words are the most frequent words in the lexicon, followed by monosyllabic words, before returning to the downward trend. The rise in disyllabic words, relative to smaller monosyllabic words, is likely due to the importance of compounding as a word-formation device in Cantonese, which produces polysyllabic words by combining two or

more monosyllabic morphemes, though two-stem compounds are the most frequent (Matthews & Yip, 2011).

Table 10. Word size in syllables, token frequencies.

	1	2	3	4	5	n ≥ 6	Totals
HKCAC	62,004	34,658	2,277	369	24	12	99,344
HKCanCor	85,759	31,853	1,921	377	51	16	119,977
IARPA	639,605	170,046	18,111	2,791	167	25	830,745

Table 11. Word size in syllables, type frequencies.

	1	2	3	4	5	n ≥ 6	Totals
HKCAC	1,178	5,382	901	235	21	9	7,726
HKCanCor	2,023	4,334	745	294	41	12	7,449
IARPA	2,380	11,544	4,498	1,042	90	21	19,575

Another way to compare and contrast corpora is to examine the range of attested syllables, and compare them against the set of logically possible syllables predicted from subsyllabic structures. As discussed in section 2, syllables can be broken down into an onset and a rime. Cantonese has 20 distinct onsets (19 overt onsets plus the empty onset) and 56 rimes, predicting with free combination 1,120 distinct syllables. To this number, we can add two syllables created by the syllabic nasals *m* and *ŋ*, yielding 1,122. This count is a total for atonal syllables (syllables without tone). We do not expect to observe this number of syllables in any corpus because prior research has shown that, because of phonotactic restrictions and the history of the language, Cantonese employs far fewer syllables in words. Thus, Bauer and Benedict (1997) propose a syllabary of 750 attested syllables, drawing on past research and their own work probing possible syllables with native speakers.

As shown in Table 12, all corpora undershoot this logical total by a wide margin, but there are also some important differences among them in terms of their attested syllables. The values under Attested Syllables give the counts of all attested syllable types, regardless of their frequency. Under Adjusted Token and Adjusted Type, which are derived from token and type frequencies respectively, we exclude marginal syllables that have less than three examples because these syllables are not really viable in the language. This table also relates each count to the total possible (1,122), giving the percentage occurrence of that total in parentheses.

Table 12. Syllabary size by attested syllables (sum of nonzero frequencies), token and type frequencies (greater than 3).

	Attested Syllables	Adjusted Token (n>3)	Adjusted Type (n>3)
HKCAC	605 (53.92)	495 (44.12)	446 (39.75)
HKCanCor	589 (52.50)	506 (45.10)	453 (40.37)
IARPA	596 (53.12)	558 (49.73)	519 (46.26)

One generalization that can be derived from these facts is that all of the corpora undershoot the 750 item syllabary of Bauer and Benedict (1997) by a wide margin. The attested syllables of HKCAC come closest, but undershoot it by 145 syllable types. The most comprehensive syllabary based on lexicons (i.e., derived from adjusted types) is that of IARPA, which is missing 231 syllable types. In addition, all corpora have a large number of marginal syllables because adjusted totals drop drastically from attested syllables. The average drop from

attested syllables to syllables based on adjusted types is 20.80%, though the drop in the IARPA corpus is far less (12.92%), likely due to its size. To summarize, all corpora undershoot both the logically possible (1,122) and conjectured (750) syllabaries, though larger corpora like IARPA are more representative when marginal syllables are excluded.[5]

The set of attested syllables and their frequencies can be broken down by the way syllables are encoded. In particular, Bauer and Benedict's syllabary distinguishes regular syllables that have a standard character-based representation, colloquial syllables that lack standard characters, the syllables of adapted loanwords (chiefly English loans), and a large number of impossible syllables, i.e., syllables that are logically possible combinations of Cantonese onsets and rimes but are not attested. The attested syllables from above are broken down into the categories in Table 13. With this breakdown, we can see that size of HKCAC's syllabary based on attested syllables is due largely to a larger number of impossible syllables; IARPA's attested syllables are much higher when regular character-based syllables are considered.

Table 13. Attested syllables and syllable frequencies by encoding type (upper bound for attested in parentheses).

		Regular (584)	Colloquial (126)	Loan (40)	Impossible (372)	Total (1,122)
Attested	HKCAC	525	36	8	36	605
	HKCanCor	534	43	2	10	589
	IARPA	555	37	2	2	596
Token	HKCAC	122,095	16,520	195	282	139,092
	HKCanCor	140,644	16,415	11	26	157,096
	IARPA	955,885	90,217	8	73	1,046,183
Type	HKCAC	15,112	345	18	89	15,564
	HKCanCor	14,127	239	5	17	14,388
	IARPA	42,679	1,014	3	14	43,710

The above patterns investigate differences in whether a syllable is attested or not, but ignores the frequency distributions of these syllables. In general, it would be valuable to compare the syllable frequencies across corpora, again to gauge similarities and contrasts across corpora and syllable frequency is often needed to balance experimental items. As shown in Table 14, syllable frequencies across corpora are highly correlated, though these correlations are slightly smaller for token frequencies. Correlations between syllable token and type frequency within a corpus are much lower (Table 15), presumably because of the loss of many syllables in high frequency words.

[5] We note that the attested syllables in HKCAC also undershoot the 753 syllables reported in Leung et al. (2004). This discrepancy is due to the fact that we conducted different searches: we have restricted our search here to combinations of licit onsets and rimes, whereas this work documented many casual speech phenomena that includes both new segments (e.g., *ɐʔ*) and new combinations due to reduction, assimilations, and casual speech phonology.

Table 14. Correlations of syllable frequencies across corpora.

Comparison	Token	Type
HKCAC, HKCanCor	0.8130	0.9132
HKCAC, IARPA	0.8795	0.8808
HKCanCor, IARPA	0.8548	0.9025

Table 15. Correlations between syllabaries from type and token frequencies.

	Token, Type
HKCAC	0.6576
HKCanCor	0.5233
IARPA	0.5835

Finally, we further probe syllable frequencies by investigating syllable shapes across corpora. Table 16 and Table 17 give the token and type frequencies of the five basic shapes of syllables, distinguishing open syllables with monophthongs (CV) and diphthongs (CVV), syllables closed with a nasal (CVN) or a stop (CVS), as well as syllables with a syllabic nasal (N). As shown by the percentage frequencies in both tables, the relative frequencies of all shapes are very similar across corpora. However, there are clear differences when comparing token and type in the same corpus. Open CV and CVV syllables are the most prevalent syllable shape in token frequencies, followed by CVN, CVS, and then N. In type frequencies, on the other hand, the frequencies of open syllables drop considerably, especially for CV syllables. This drop is compensated by an increase in closed syllables, whereby CVN becomes the most frequent shape in all lexicons. Syllabic nasals are by far the least frequent in both token and type frequency.

Table 16. Syllable shape token frequencies across corpora.

	CV	CVV	CVN	CVS	N
HKCAC	45,010 (32.19)	45,159 (32.30)	32,953 (23.57)	12,677 (9.07)	4,028 (2.88)
HKCanCor	49,764 (31.68)	52,736 (33.57)	33,566 (21.37)	17,019 (10.83)	4,011 (2.55)
IARPA	355,433 (33.97)	363,552 (34.75)	217,021 (20.74)	81,484 (7.79)	28,693 (2.74)

Table 17. Syllable shape type frequencies across corpora.

	CV	CVV	CVN	CVS	N
HKCAC	2,927 (18.56)	4,560 (28.91)	5,791 (36.72)	2,395 (15.19)	98 (0.62)
HKCanCor	2,446 (17.00)	4,215 (29.30)	5,265 (36.59)	2,339 (16.26)	123 (0.85)
IARPA	6,826 (15.62)	13,968 (31.96)	15,971 (36.54)	6,233 (14.26)	712 (1.63)

4.3 Consonants

We now turn to the distributions of consonants across the three corpora. As noted in section 2.1, some consonants (stops and nasals) can appear in both onset and coda position, and two nasals, namely *ŋ* and *m*, can function as syllable nuclei. Therefore, our counts below distinguish consonants by their syllabic role, but sounds that occur in more than one slot can be summed if a more general tally is desired (see the data supplement). Table 18 gives the token and type frequencies of all consonants. Several of the more salient phonemes have similar distributions across corpora. For example, *k* has the highest token frequency in all corpora, and is ranked high in all type frequencies as well. Likewise, *t, ts,* and *j* have high frequency across all

columns. Interestingly, *h* has high frequency in all token counts, but not type counts, and *s* has the opposite pattern in all corpora. These are two cases that clearly distinguish token and type frequency.

Table 18. Consonant frequencies by corpus and type/token.

		HKCAC		HKCanCor		IARPA	
		Token	Type	Token	Type	Token	Type
Onset	ø	13,560	572	12,032	353	107,881	990
	p	3,091	642	3,690	680	26,936	1,890
	p^h	956	210	1,005	259	5,725	744
	t	11,459	1,217	14,254	1,045	93,996	3,569
	t^h	2,911	506	3,204	545	25,575	1,700
	k	21,762	1,771	21,794	1,310	138,075	4,094
	k^h	3,871	279	3,769	263	27,868	829
	k^w	427	128	1,734	216	2,949	404
	k^{wh}	71	30	63	30	358	84
	f	3,089	685	2,620	550	19,786	2,040
	s	8,852	1,928	9,579	1,875	62,558	5,275
	h	12,541	898	18,607	873	111,302	3,140
	ts	13,676	1,729	13,845	1,463	79,200	4,262
	ts^h	3,327	785	4,047	956	30,649	2,841
	m	5,917	782	6,433	696	92,540	2,783
	n	296	95	7,841	269	4,179	193
	ŋ	1322	176	3,515	152	237	115
	w	4,329	452	5,119	427	32,575	1,387
	l	14,480	1,045	8,210	933	91,535	3,289
	j	13155	1,634	15,735	1,493	92,259	4,081
Coda	p	899	332	1,091	348	7,236	848
	t	5,598	949	6,099	870	33,846	2,157
	k	5,813	1,103	9,829	1,121	40,402	3,228
	m	6,209	594	6,931	589	35,862	1,681
	n	12,433	2,441	13,613	2,261	93,710	6,778
	ŋ	14,220	2,703	13,022	2,415	87,449	7,512
Nucleus	m	4,008	85	3,850	76	28,663	702
	ŋ	19	13	161	47	30	10

These anecdotal observations are backed up by the correlations shown in Table 19 and Table 20. Thus, all comparisons between corpora show that consonant frequencies are highly correlated, especially with type frequencies, though HKCanCor has slightly lower correlations with the other corpora in token frequency. Correlations between token and type frequencies within a corpus drop considerably, similar to the drop found in syllable frequencies.

Table 19. Correlations between corpora in consonant frequency.

Comparison	Token	Type
HKCAC, HKCanCor	0.91357	0.98787
HKCAC, IARPA	0.95098	0.98335
HKCanCor, IARPA	0.90000	0.98663

Table 20. Correlations in token and type frequency in consonants.

	Token, Type
HKCAC	0.77887
HKCanCor	0.68211
IARPA	0.70782

4.4. Vowels

We now turn to the distribution of vowels across the three corpora. Recall from section 2.1 that monophthongs can either appear in open syllables or the first part of a VC rime. We therefore sum their frequencies in both contexts, given that they both occupy the X_1 position of the rime. Table 21 provides type and token counts for all vowels across the three corpora. The monophthongs *i, a:, o,* and *ɐ* consistently have the highest-ranking token and type frequencies. The diphthongs *ɐi, ou, ei,* and *ɐu* also have high frequencies relative to other diphthongs, though the high rank of *ɐi* drops considerably in type frequency, likely due to the impact of [hɐi22] 係, which is either the first or second most common word in these corpora.

Table 21. Vowel frequencies, by corpus and type/token.

		HKCAC		HKCanCor		IARPA	
		Token	Type	Token	Type	Token	Type
Monophthongs	i	17,515	2782	20,317	2,674	127,116	6,536
	e	11,301	427	9,641	357	48,411	1,195
	y	3,159	618	3,327	608	17,342	1,882
	œ	4,036	763	3,973	660	22,462	1,939
	u	5,227	1220	5,031	1,121	32,028	3,524
	o	17,824	1494	18,451	1,056	13,8401	3,716
	a:	15,226	1646	22,013	1,763	17,6962	5,914
	ɐ	15,650	1979	17,596	1,811	99,926	4,990
Diphthongs	ei	8,380	689	8,478	652	62,303	1,674
	œi	4,788	386	4,861	343	39,844	1,009
	ui	1,069	175	1,090	92	3,130	293
	oi	1,597	280	1,285	250	8,803	821
	ɐi	10,361	598	14,765	694	81,225	1,961
	ai	2,069	384	2,853	354	19,739	1,346
	iu	1,756	325	1,839	269	14,548	925
	eu	5	3	11	7	0	0
	ou	7,921	806	9,880	749	72,997	2,540
	ɐu	6,581	720	6,801	654	46,936	2,139
	au	600	171	873	151	5,317	594

Correlation data derived from the aggregated vowel counts (Table 22 and Table 23) support these observations. All correlations between corpora are very strong, and the least correlated pair, HKCAC and IARPA, compare with their correlations in word frequency. While correlations between token and type frequencies within each corpus are weaker, the vowel patterns are strongly correlated, more so than with consonants.

Table 22. Correlations between corpora in vowel frequency.

	Token	Type
HKCAC, HKCanCor	0.97365	0.98800
HKCAC, IARPA	0.93474	0.97083
HKCanCor, IARPA	0.97586	0.97942

Table 23. Correlations between type and token frequency in vowels.

	Token, Type
HKCAC	0.84340
HKCanCor	0.84174
IARPA	0.86294

4.5 Tone

Finally, we report on the distribution of suprasegmental tone in the three corpora. To begin, we note that tone is affected by syllable shape because contour tones (T2, T4, T5) are restricted in checked syllables ending in *p t k* (=CVS). This is illustrated below in Table 24 with token frequencies from HKCanCor, where we see that T5 is unattested, T4 is marginal, and T2 is underrepresented in CVS syllables (the expected frequency of T2 in CVS based on column totals is 2,837). This systematic gap is true of all corpora.

Table 24. Tone frequencies in HKCanCor (token) by syllable shape.

	High level T1, 55	High rising T2, 25	Mid level T3, 33	Low falling T4, 21	Low rising T5, 23	Low level T6, 22	Totals
CV(V)	19,032	15,942	21,521	10,300	17,132	22,584	106,511
CVN	9,305	10,110	2,580	6,270	787	4,514	33,566
CVS	8,270	132	3,161	5	0	5,451	17,019
Totals	36,607	26,184	27,262	16,575	17,919	32,549	

The following two tables give context-free frequencies of the six tones. They all appear to follow the same trend, whereby T1 and T6 have slightly higher than expected frequencies (based on a one-in-six chance rate of 16.66%), the low contour tones (T4 and T5) have slightly lower frequencies, and the remaining tones, T2 and T3, are very close to chance levels. This trend seems to be exaggerated in type frequencies, where all corpora but IARPA have even higher frequency for T1, and all corpora have marked drops in the frequency of T5, while T4 gets a boost.

Table 25. Tone token frequencies by corpora.

	High level T1, 55	High rising T2, 25	Mid level T3, 33	Low falling T4, 21	Low rising T5, 23	Low level T6, 22
HKCAC	28,775 (20.92)	22,614 (16.44)	26,543 (19.30)	16,886 (12.28)	14,833 (10.79)	27,880 (20.27)
HKCanCor	36,607 (23.30)	26,184 (16.67)	27,262 (17.35)	16,575 (10.55)	17,919 (11.41)	32,549 (20.72)
IARPA	309,659 (29.60)	139,562 (13.34)	169,386 (16.19)	99,930 (9.55)	136,688 (13.07)	190,958 (18.25)

Table 26. Tone type frequencies by corpora.

	High level T1, 55	High rising T2, 25	Mid level T3, 33	Low falling T4, 21	Low rising T5, 23	Low level T6, 22
HKCAC	3,818 (24.43)	2,335 (14.94)	2,930 (18.75)	2,320 (14.85)	974 (6.23)	3,250 (20.80)
HKCanCor	3,683 (25.60)	2,293 (15.94)	2,282 (15.86)	2,455 (17.06)	783 (5.44)	2,892 (20.10)
IARPA	11,850 (27.11)	6,815 (15.59)	6,862 (15.70)	7,921 (18.12)	2,503 (5.73)	7,759 (17.75)

Frequency distributions for tone, however, are affected by context, and this needs to be factored into calculations of the impact of frequency in language processing. Table 27 and Table 28 give the counts relative to the first or second syllable in disyllabic words. We assume that there will be similar trends in polysyllabic words greater than two syllables, but we focus on disyllabic words because they are far more numerous and generalizing from them is more straightforward. The relative frequencies in words with three syllables and greater can be explored with parallel queries by manipulating word size in the data supplement. By contrasting the percentage occurrence in σ_1 versus σ_2, we see that T1 and T6 swap ranks: T1 is the most common tone in initial syllables, but it is demoted to the second or third rank because T6 is promoted to the highest rank in the second syllable. This trend is observed in both token and type frequencies, but more muted in the latter. These trends are interesting because there does not appear to be clear patterns in Cantonese phonology that would preclude a tone from having the same frequency in both positions. Indeed, one potential pattern, *pinjam* (變音) "changed tone", has the tendency to change all tones to either the high level (55) or high rising (25) tone in the final syllable of many disyllabic word (Chen, 2000; M. Yip, 1980), the opposite of what is found here.

Table 27. Tone frequencies (token) by word position in disyllabic words across corpora.

		High level T1, 55	High rising T2, 25	Mid level T3, 33	Low falling T4, 21	Low rising T5, 23	Low level T6, 22
HKCAC	σ_1	9,688 (28.10)	6,193 (17.96)	4,269 (12.38)	5,306 (15.39)	3,238 (9.39)	5,787 (16.78)
	σ_2	6,517 (19.48)	5,521 (16.50)	7,414 (22.16)	3,914 (11.70)	962 (2.88)	9,123 (27.27)
HKCanCor	σ_1	10,167 (31.92)	6,179 (19.40)	3,285 (10.31)	5,458 (17.13)	2,407 (7.56)	4,357 (13.68)
	σ_2	6,274 (19.70)	4,865 (15.27)	4,107 (12.89)	3,645 (11.44)	2,238 (7.03)	10,724 (33.67)
IARPA	σ_1	54,002 (31.76)	36,926 (21.72)	16,229 (9.54)	23,958 (14.09)	13,157 (7.74)	25,774 (15.16)
	σ_2	40,722 (23.95)	24,423 (14.36)	22,804 (13.41)	23,429 (13.78)	12,584 (7.40)	46,084 (27.10)

Table 28. Tone frequencies (type) by word position in disyllabic words across corpora.

		High level T1, 55	High rising T2, 25	Mid level T3, 33	Low falling T4, 21	Low rising T5, 23	Low level T6, 22
HKCAC	σ_1	1440 (27.07)	833 (15.66)	946 (17.78)	763 (14.34)	396 (7.44)	942 (17.71)
	σ_2	1187 (22.10)	815 (15.18)	1100 (20.48)	765 (14.25)	261 (4.86)	1242 (23.13)
HKCanCor	σ_1	1,237 (28.54)	663 (15.30)	684 (15.78)	736 (16.98)	255 (5.88)	759 (17.51)
	σ_2	919 (21.20)	743 (17.14)	723 (16.68)	748 (17.26)	207 (4.78)	994 (22.93)
IARPA	σ_1	3,298 (28.57)	1,663 (14.41)	1,927 (16.69)	1,951 (16.90)	671 (5.81)	2,034 (17.62)
	σ_2	2,963 (25.67)	1,941 (16.81)	1,880 (16.29)	2,051 (17.77)	643 (5.57)	2,066 (17.90)

4.6 Phonotactics

We can also compare and contrast data sets on how well they respect phonotactic constraints, or the constraints on legal combinations of sounds. Many production and perception processes are affected by phonotactic constraints (Dell et al., 1993; Goldrick, 2004; Hay et al., 2004). Further, phonotactics are in many ways the heart of phonological analysis (Hayes & White, 2013; Prince

& Tesar, 2004), so an assessment of the three data sets relative to these constraints is of interest to both linguists and psycholinguists.

Cantonese phonotactics can be characterized as a set of negative constraints against combinations of syllabic positions, like a ban on a particular nucleus + coda combination. In Table 29, seven constraints from the linguistics literature are shown schematically (Cheng, 1991; M. Yip, 1997), with the banned phoneme sequences given on the last line.[6] The frequencies reported here show that most constraints are respected by all corpora, and in many cases, they are categorically respected in the sense that there are no observed violations. The sporadic violations of the other constraints seem to be mainly limited to sound symbolic words and loans, as in [pɐm55] in 'ping pong' from IARPA (violates constraint a Table 29), and [tep55] 'sound of chewing' from HKCanCor (violates constraint e Table 29).

Table 29. Frequencies of phonotactic violations by constraint and corpus.

	HKCAC		HKCanCor		IARPA	
	Token	Type	Token	Type	Token	Type
a. *Ons…Coda lab … lab *2 x /p pʰ m f kʷ kʷʰ/	1	1	0	0	16	5
b. *Nuc Coda [+round] lab *up um op om yp ym	13	3	0	0	0	0
c. *Ons Nuc lab [-back, +round] */p pʰ m f kʷ kʷʰ/+/y œ/	3	2	0	0	0	0
d. *Ons Nuc Ons cor [+bk, +rd] cor */t tʰ s n l/ + /o, u/ + cor	11	7	1	1	0	0
e. *Nuc Coda e lab/cor *em en ep et	4	4	8	3	0	0
f. *Ons Nuc cor u */t tʰ s n l/ + u	2	2	14	12	0	0
g. *Nuc Coda [+high] dorsal *ik iŋ yk yŋ uk uŋ	5	2	0	0	0	0

5. Discussion

5.1 Recurring themes

The three corpora reviewed here are similar in kind in that they are large collections of spontaneous speech in adults. Despite this common ground, however, we have documented important differences in the frequency distributions of words and sounds. The differences are greater in word frequencies. Perhaps surprisingly, there is little lexical overlap among the corpora, and among the shared lexical items, word frequency is weakly correlated when broken

[6] The schematic constraints use the following distinctive features to define classes of sounds: lab(ial) for bilabial and labial-dental sounds, [+round] or [+rd] for vowels with lip rounding, [-back] for front vowels, such as *y* and *œ*, [+back] or [+bk] for back vowels like *o*, cor(onal) for coronal sounds using the front and tip of the tongue.

down by frequency class. While sound frequencies are better correlated, with correlations rarely dipping below .9, important differences are documented here as well in attested syllables and the breadth of atonal syllables outside of traditional syllabaries. In sum, there are important differences across corpora that must be attended to when selecting a corpus and interpreting data relative to the frequencies reported in that corpus.

The distinction between token and type frequencies is also necessary as we found important differences between the two in just about every dimension of sound structure. It affects word size, syllabaries, consonant and vowel occurrence, and tone because the sound structures represented multiple times in high frequency items are reduced in the lexicon. For example, correlations between corpora in syllable frequencies range between .81 and .91, but correlations between token and type frequencies within a corpus are between .52 and .65. Our findings show the magnitude of differences, and correspondingly, how consequential this decision can be. Finally, we have also found frequency distributions to be affected by other factors, including encoding type, syllable shape, and word position, which must also be taken into consideration.

5.2 Applications to experimental designs

How do these facts apply to experimental designs and decisions about experimental stimuli? This question is more important for word frequency than for sound structures frequency, because sound frequency is in general better correlated across the corpora than word frequency. The lexicon based on IARPA is by far the largest with close to 20,000 entries, so if breadth of the lexicon is the primary criteria, it is the best option. HKCanCor is also a good option if the design requires words to be sub-grouped by part of speech class, as the structured representations are POS-tagged and have support for Python programming. HKCanCor is also well-correlated in word frequency with both IARPA and HKCAC, so it seems to have word frequencies typical of the larger population. Given the lack of lexical overlap, researchers may encounter words that they wish to include in their study, but are not listed in a given corpus. If this arises, then the frequencies reported here can be used to create probabilities based on frequencies reported in another corpus, which can help fill in some gaps.

The frequency of sound structure is less affected by the corpus, so selecting one over the other is likely a matter of the specific kind of information. IARPA has a more representative syllabary when marginal syllables are excluded and larger baselines in general. However, the structured representations of HKCanCor make it easy to cross-classify the data by part of speech categories, and word segmentation is likely to be more reliable than IARPA. If surface representations are required, then HKCAC is the only option, and the facts of Leung et al. (2004) should be consulted. If the distributions of particular structures seem to differ in different corpora, researchers can also sum the frequencies in the tables reported here from all corpora, and derive average values that are less affected by corpus selection. The data supplements to this article, word frequencies and sound frequencies, give the raw frequencies of all the structures reported here in a single data table, and can be easily manipulated to achieve these results (see Appendix A).

5.3 Future work

Though we have compared the three corpora on the basis of how well they obey certain phonotactic constraints, our investigation in section 4.6 is preliminary in the sense that it focuses on established constraints from the literature that are essentially categorical. Research on phonotactics in a variety of languages, however, has shown that phonotactic restrictions are gradient in nature and this research recognizes constraints against structures that are attested but

under-represented in the lexicon (Frisch et al., 2000; Treiman et al., 2000). Gradient phonotactics has in fact been investigated in Cantonese by Kirby and Yu (2007), and found to support a departure from classical generative phonology that only distinguishes between attested and prohibited structures. In particular, this study probed native speaker intuitions about the well-formedness of syllables in three classes: attested syllables, unattested syllables that violate phonotactic constraints (systematic gaps), and unattested syllables that do not violate phonotactic constraints (accidental gaps). They found that regression models with neighborhood density (i.e., the degree of confusability of a word with other words) and phonotactic probability as predictors accounted for a moderate amount of the variation, though phonotactic probability was found to be weaker than other studies of English, and perhaps even unnecessary in explaining the data.

We accept the larger point about gradient phonotactics in Cantonese, but our findings suggest that the claimed diminished role of phonotactic probabilities in explaining word-likeness data can be fruitfully re-examined. Our findings show important differences between type and token frequencies. Kirby and Yu used a combination of type and token frequencies in calculating phonotactic probability, which could have reduced some of the impact of this measure. They also used the type frequencies from Leung et al. (2004), but, as explained above, these frequencies are problematic. Though which frequency measure to use is still somewhat controversial, type frequency has emerged as a standard measure for correlations with grammatical well-formedness (Hay et al., 2004). Given this problem, we think that a follow-up study correlated with the type frequencies reported here will be more conclusive about the role of phonotactic probability.

Another understudied aspect of the corpora is the linguistic behavior of bilinguals. The use of English by Cantonese speakers has risen considerably in the past 25 years, so much so that in 2016, approximately 53% of Hong Kong residents actively use English (Liu, 2017). The prevalence of English can be observed in the texts, as many of the native speakers are bilingual in English and Cantonese and switch freely between the two languages. Though English is redacted from the IARPA corpus, it is represented in both the HKCAC and HKCanCor corpora. English words account for about 0.8% of the words in HKCAC and 1.9% in HKCanCor. We have focused on documenting the frequencies of Cantonese language structures, but it is a fact that many of the speakers are producing Cantonese words while also sometimes switching to English. This fact, and the linguistic annotations in these corpora that distinguish individual speakers, support a variety of research questions. Which linguistic contexts lead to switches between the two languages, and are there individual differences? What characterizes the Cantonese words supplanted by English ones, and are there prosodic or other markers that can help predict switches? While these questions can be investigated in HKCAC and HKCanCor, it should be noted that the corpora were not designed with many of these questions in mind. A more recent corpus, SpiCE (Khia Johnson et al., 2020), was in fact designed to address questions like these. This corpus includes 19 hours of high-quality recordings of bilingual speech in English and Cantonese, detailed transcriptions (force-aligned phonetic transcripts), and robust search functions, and is ideally suited to address these and other questions.

Acknowledgements.

This work was supported in part by Social Sciences and Humanities Resource Council grant 435-2014-0452. There are no conflicts of interest.

Appendix A. Data supplements.

All of the data and scripts discussed in this article are available at: github.com/jane-lisy/cantfreq. Two consolidated data files are especially useful. The file <wordfrequencies_master> provides all the information on word frequencies that we investigated in section 3. This document has 26,506 rows for all the words in the three corpora, and 15 columns for word attributes, including frequencies and probabilities from the three corpora, traditional and simplified orthographic representations, a phonological representation in Jyutping, and word size. The file <soundfrequencies_master> likewise assembles all the information about sound frequencies reported in section 4. It has 1,232 rows representing all of the sound structures in Cantonese and it distinguishes segments, rimes, syllables, and tones. There are 17 columns for reporting frequencies and probabilities, as well as attributes that cross-classify the sounds by syllable role, syllable shape, encoding type, and structure type for selecting the appropriate baselines, which are declared in special rows.

The data tables and Python scripts for each corpus are also available on the GitHub page for corpus-specific exploration. These corpus-specific data tables are associated with the Python notebooks that generated them, which are fully commented and enable users to replicate the results reported here.

Appendix B. Phonetic symbols used in different systems and corpora.

Phonetic Description	IPA	Yale	Jyutping	HKCAC	HKCanCor	IARPA	Example (phonetic)
Obstruents							
bilabial unaspirated stop	p	b	b	p	b	b	爸 ba:55 ‘father’
bilabial aspirated stop	p^h	p	p	pH	p	p	爬 pa:21 ‘crawl’
dental unaspirated stop	t	d	d	t	d	d	大 da:i22 ‘large, great’
dental aspirated stop	t^h	t	t	tH	t	t	頭 tau21 ‘head’
velar unaspirated stop	k	g	g	k	g	g	家 ga:55 ‘family, home’
velar aspirated stop	k^h	k	k	kH	k	k	球 kau21 ‘ball’
labial-velar unaspirated stop	k^w	gw	gw	kw	gw	gw	軍 gwan55 ‘army, troops’
labial-velar aspirated stop	k^{wh}	kw	kw	kwH	kw	kw	裙 kwan21 ‘skirt’
labial-dental fricative	f	f	f	f	f	f	肥 fei21 ‘fat’
dental fricative	s	s	s	s	s	s	時 si21 ‘time’
glottal fricative	h	h	h	h	h	h	下 ha:22 ‘below, to descend’
dental unaspirated affricate	ts	j	z	ts	z	j	姐 dze25 ‘older sister’
dental aspirated affricate	ts^h	ch	c	tsH	c	ch	車 tse55 ‘car’
Sonorants							
bilabial nasal	m	m	m	m	m	m	媽 ma:55 ‘mother’
dental nasal	n	n (~l)	n	n	n	n	年 nin21 ‘year’
velar nasal	ŋ	ng	ng	N	ng	ng	牙 ŋa:21 ‘teeth’
bilabial glide	w	w	w	w	w	w	畫 wa:25 ‘painting’
dental lateral approximant	l	l	l	l	l	l	籃 la:m21 ‘basket’
palatal glide	j	y	j	j	j	y	兒 ji21 ‘son, infant’
Simple vowels							
high front unrounded	i	i	i	i	i	i	撕 si55 ‘to tear’
high front rounded	y	yu	yu	y	yu	yu	瘀 jy35 ‘bruise’
high back rounded	u	u	u	u	u	u	湖 wu21 ‘lake’
mid front unrounded	e [ɛ]	e	e	E	e	e	笛 dek22 ‘flute’
mid front rounded	œ	eu	oe	J	oe	eu	樣 jœŋ22 ‘kind, sort’
mid back rounded	o [ɔ]	o	o	O	o	o	菠 bo55 ‘spinach’
low central short	ɐ	a	a	A	a	a/aa	龜 gwai55 ‘turtle’
low central long	a:	aa	aa	a	aa	a	爸 ba:55 ‘father’
Diphthongs							
high front unrounded +u	iu	iu	iu	iu	iu	iu	笑 siu33 ‘laugh’
high back rounded +i	ui	ui	ui	ui	ui	ui	會 wui25 ‘meeting’
mid front unrounded+i	ei	ei	ei	ei	ei	ei	四 sei33 ‘four’
mid front unrounded +u	eu [ɛu]	ew	eu	Eu	eu	ew	掉 deu22 ‘throw’
mid front rounded +i	œi [ɵy]	eui	eoi	0y	eoi	eui	水 sœi25 ‘water’
mid back +i	oi [ɔy]	oi	oi	Oi	oi	oi	菜 tsoi33 ‘vegetable’
mid back +u	ou	ou	ou	ou	ou	ou	好 hou25 ‘good’
low central +i	ɐi	ai	ai	Ai	ai	ai	西 sai55 ‘west’
low central +u	ɐu	au	au	Au	au	au	夠 gau33 ‘enough’
low central long +i	a:i	aai	aai	ai	aai	aai	嘥 sa:i55 ‘waste’
low central long +u	a:u	aau	aau	au	aau	aau	教 ga:u33 ‘teach’

Tones							
high rising tone	a25	á	a2	2	2	2	使 si25 ‘to cause, make’
high level tone	a55	ā	a1	1	1	1	詩 si55 ‘poem’
(high falling tone)	a53	à	a1	1	1	1	(絲) si53 ‘silk’
mid level tone	a33	a	a3	3	3	3	試 si33 ‘to try’
low rising tone	a23	áh	a5	5	5	5	市 si23 ‘city, market’
low level tone	a22	ah	a6	6	6	6	事 si22 ‘matter, affair’
low falling tone	a21	àh	a4	4	4	4	時 si21 ‘time’

References

Ambridge, B., Kidd, E., Rowland, C. F., & Theakson, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language, 42*, 239-272.

Andrus, T., Dubinski, E., Fiscus, J., Gillies, B., Harper, M., Hazen, T. J., . . . Wong, J. (2016). *IARPA Babel Cantonese Language Pack IARPA-babel101b-v0.4c LDC2016S02. Web Download*.

Atkins, S., Clear, J., & Ostler, N. (1992). Corpus design criteria. *Literary and Linguistic Computing, 7*, 1-16.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1996). *CELEX2*. Philadelphia: Linguistic Data Consortium.

Bates, E., D'amico, S., Jacobsen, T., Székely, A., Andonova, E., Devescovi, A., . . . Tzeng, O. (2003). Timed picture naming in seven languages. *Psychonomic Bulletin & Review, 10*, 344-380.

Bauer, R. S. (2013). Phonetic features of colloquial Cantonese. In G. Peng & F. Shi (Eds.), *Eastward flows the Great River: Festschrift in honor of Professor William S-Y. Wang on his 80th birthday* (pp. 30-42). Hong Kong: The City University of Hong Kong.

Bauer, R. S., & Benedict, P. K. (1997). *Modern Cantonese phonology* (Vol. 102). Berlin: Mouton de Gruyter.

Bauer, R. S., Cheung, K.-h., & Cheung, P.-m. (2003). Variation and merger of the rising tones in Hong Kong Cantonese. *Language Variation and Change, 15*, 211-225.

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python*. Sebastopol, CA: O'Reilly Media, Inc.

Chao, Y. R. (1930). A system of tone letters. *Le Maître Phonétique, 45*, 24-27.

Chen, M. (2000). *Tone sandhi: Patterns across Chinese dialects*. Cambridge: Cambridge University Press.

Cheng, L. L.-S. (1991). Feature geometry of vowels and co-occurrence restrictions in Cantonese. In *Proceedings of the 9th West Coast Conference on Formal Linguistics 9* (pp. 107-124).

Chin, C. O., & Tweed, A. M. (2019). *The corpus of mid-20th century Hong Kong Cantonese (second phase) and its applications.* Paper presented at the Workshop on Cantonese (WOC): Cantonese Study: An Empirical Approach, Hong Kong, The Hong Kong Polytechnic University.

Connine, C. M., Ranbom, L. J., & Patterson, D. J. (2008). Processing variant forms in spoken word recognition: The role of variant frequency. *Perception and Psychophysics, 70*, 403-411.

Dell, G. S. (1990). Effects of frequency and vocabulary type on phonological speech errors. *Language and Cognitive Processes, 5*, 313-349.

Dell, G. S., Juliano, C., & Govindjee, A. (1993). Structure and content in language production: A theory of frame constraints in phonological speech errors. *Cognitive Science, 17*, 149-195.

Dell, G. S., Nozari, N., & Oppenheim, G. M. (2014). Word production: Behavioral and computational considerations. In M. Goldrick, V. Ferreira, & M. Miozzo (Eds.), *The Oxford handbook of language production* (pp. 88-104). Oxford: Oxford University Press.

Ernestus, M., & Baayen, R. H. (2003). Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language, 79*, 5-38.

Farris-Trimble, A., & Tessier, A.-M. (2019). The effect of allophonic processes on word recognition: Eye-tracking evidence from Canadian raising. *Language, 95*, e136-e160.

Fletcher, P., Leung, C. S. S., Stokes, S., & Weizman, Z. (2000). *Cantonese pre-school language development. A guide. (Report of the project "Milestones in the learning of spoken Cantonese by pre-school children"*. Hong Kong: Language Fund.

Forster, K. I., & Davis, C. (1984). Repetition priming and frequency attenuation in lexical access. *Journal of Experimental Psychology: Learning, Memory and Cognition, 10*, 680-698.

Frisch, S. A., Large, N. R., & Pisoni, D. S. (2000). Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. *Journal of Memory and Language, 42*, 481-496.

Fu, G., Luke, K.-K., & Wong, P. P.-W. (2005). Description of the HKU Chinese Word Segmentation System for Sighan Bakeoff 2005. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*.

Fung, R., & Bigi, B. (2015). *Automatic word segmentation for spoken Cantonese*. Paper presented at the 2015 International Conference Oriental COCOSDA.

Gahl, S. (2008). Time and Thyme are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language, 84*, 474-496.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*, 251-279.

Goldrick, M. (2004). Phonological features and phonotactic constraints in speech production. *Journal of Memory and Language, 51*, 586-603.

Goldrick, M., Baker, H. R., Murphy, A., & Baese-Berk, M. (2011). Interaction and representational integration: Evidence from speech errors. *Cognition, 121*, 58-72.

Gordon, B. (1983). Lexical access and lexical decision: Mechanisms of frequency sensitivity. *Journal of Verbal Learning and Verbal Behavior, 22*, 24-44.

Gries, S. T. (2015). Quantitative designs and statistical techniques. In D. Biber & R. Reppen (Eds.), *The Cambridge Handbook of English Corpus Linguistics* (pp. 50-71). Cambridge: Cambridge University Press.

Hay, J., Pierrehumbert, J., & Beckman, M. (2004). Speech perception, well-formedness and the statistics of the lexicon. In J. Local, R. Ogden, & R. Temple (Eds.), *Phonetic Interpretation: Papers in Laboratory Phonology VI* (pp. 58-74). Cambridge: Cambridge University Press.

Hayes, B., & White, J. (2013). Phonological naturalness and phonotactic learning. *Linguistic Inquiry, 44*, 45-75.

Hong Kong Polytechnic, U. (2015). *PolyU Corpus of Spoken Chinese* Retrieved from: http://asianlang.engl.polyu.edu.hk/

Jaeger, T. F., & Norcliffe, E. J. (2009). The cross-linguistic study of sentence production. *Language and Linguistic Compass, 3*, 866-887.

Johansson, V. (2009). Lexical diversity and lexical density in speech and writing: a developmental perspective. *Lund Working Papers in Linguistics, 53*, 61-79.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 145-166). San Diego: Academic Press.

Johnson, K., Babel, M., Fong, I., & Yiu, N. (2020). SpiCE: A new open-access corpus of conversational bilingual speech in Cantonese and English. In *Proceedings of the 12th conference on language resources and evaluation* (pp. 4082-4088).

Kessler, B., & Treiman, R. (1997). Syllable structure and the distribution of phonemes in English syllables. *Journal of Memory and Language, 37*, 295-311.

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics, 6*, 97-133.

Kirby, J., & Yu, A. (2007). Lexical and phonotactic effects on wordlikeness judgements in Cantonese. In *Proceedings of the 16th International Congress of the Phonetic Sciences (ICPhS XVI)* (pp. 1161–1164). Saarbrücken, Germany.

Lee, J. L. (2015). PyCantonese Python package. Chicago: University of Chicago.

Lee, T. H.-T., & Wong, C. (1998). CANCORP: The Hong Kong Cantonese Child Language Corpus. *Cahiers de Linguistique Asie Orientale, 27*, 211-228.

Leung, M. T., & Law, S.-P. (2001). HKCAC: The Hong Kong Cantonese Adult Language Corpus. *International Journal of Corpus Linguistics, 6*, 305-326.

Leung, M. T., Law, S.-P., & Fung, S.-Y. (2004). Type and token frequencies of phonological units in Hong Kong Cantonese. *Behavior Research Methods, Instruments, and Computer, 36*, 500-505.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22*, 1-75.

Levelt, W. J. M., & Wheeldon, L. (1994). Do speakers have access to a mental syllabary. *Cognition, 50*, 239-269.

Levitt, A., & Healy, A. (1985). The roles of phoneme frequency, similarity, and availability in the experimental elicitation of speech errors. *Journal of Memory and Language, 24*, 717-733.

Liu, J. (2017, June 29). Cantonese v Mandarin: When Hong Kong languages get political. *BBC News*.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear Hear, 19*, 1-36.

Luke, K.-K., & Wong, M. L.-Y. (2015). The Hong Kong Cantonese Corpus: Design and uses. *Journal of Chinese Linguistics, 25*, 309-330.

Marslen-Wilson, W. (1984). Function and process in spoken word recognition: A tutorial review. In H. Bouma & D. G. Bouwhis (Eds.), *Attention and performance X: Control of language processes* (pp. 125-150). Hillsdale, NJ: Erlbaum.

Matthews, S., & Yip, V. (2011). *Cantonese: A comprehensive grammar*. London: Routledge.

Mok, P. P.-K., Zuo, D., & Wong, P. W.-Y. (2013). Production and perception of a sound change in progress: Tone merging in Hong Kong Cantonese. *Language Variation and Change, 25*, 314-370.

Oldfield, R. C., & Wingfield, A. (1965). Response latencies in naming objects. *Quarterly Journal of Experimental Psychology, 17*, 273-281.

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

Pitt, M. A., Dilley, L., & Tat, M. (2011). Exploring the role of exposure frequency in recognizing pronunciation variants. *Journal of Phonetics, 39*, 304-311.

Prince, A., & Tesar, B. (2004). Learning phonotactic distributions. In R. Kager & J. Pater (Eds.), *Fixing priorities: Constraints in phonological acquisition* (pp. 245-291). Cambridge: Cambridge University Press.

Pulleyblank, E. (1997). The Cantonese vowel system in historical perspective. In W. Jialing & N. Smith (Eds.), *Studies in Chinese phonology* (pp. 185-217). Mouton de Gruyter: Berlin.

Roland, D., Dick, F., & Elman, J. (2007). Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language, 57*, 348-379.

So, L. K. H. (1992). *Hong Kong spoken Cantonese database*.

Stemberger, J. P., & MacWhinney, B. (1986). Frequency and the lexical storage of regularly inflected forms. *Memory & Cognition, 14*, 17-26.

Sun, J. (2020). "Jieba" Chinese text segmentation, v0.42. Retrieved from https://pypi.org/project/jieba.

Treiman, R., Kessler, B., Knewasser, S., Tincoff, R., & Bowman, M. (2000). English speakers' sensitivity to phonotactic patterns. In M. B. Broe & J. B. Pierrehumber (Eds.), *Papers in laboratory phonology V: acquisition and the lexicon* (pp. 269-282). Cambridge: Cambridge University Press.

Vitevitch, M. S. (1997). The neighborhood characteristics of malapropisms. *Language and Speech, 40*, 211-228.

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

Vitevitch, M. S., Luce, P. A., Pisoni, D. B., & Auer, E. T. (1999). Phonotactics, neighborhood activation, and lexical access for spoken words. *Brain and Language, 68*, 306-311.

Xu, L. J., & Lee, T. (1998). *Parametric variation in three Chinese dialects, Cantonese, Shanghainese and Mandarin*.

Yip, M. (1980). *The tonal phonology of Chinese.* (Doctoral dissertation). MIT,

Yip, M. (1997). Consonant-vowel interaction in Cantonese. In J. Wang & N. Smith (Eds.), *Studies in Chinese phonology* (pp. 251-274). Berlin: Mouton de Gruyter.

Yip, V., & Matthews, S. (2007). *The Bilingual Child: Early Development and Language Contact*. Cambridge: Cambridge University Press.

Yue, B. (2016). Simplified and traditional Chinese character conversion, v0.3.2. Retrieved from https://github.com/berniey/hanziconv

Zhan, W., Chang, B., Duan, H., & Zhang, H. (2006). Recent developments in Chinese corpus research. In *Proceedings of the 13th NIJL International Symposium of Language Corpora: Their Compliation and Application* (pp. 3.6-7). Tokyo, Japan.

Zipf, G. K. (1949). *Human behavior and the principle of least effort* Reading, MA: Addison-Wesley.