Taiwan Journal of Linguistics
Vol. 21.2, 45-99, 2023
DOI: 10.6519/TJL.202307_21(2).0002

LEXICAL AND SUB-LEXICAL FREQUENCY EFFECTS IN CANTONESE*

Jane S.Y. Li[1,2], Heikal Badrulhisham[1], John Alderete[1]
[1]*Simon Fraser University*
[2]*Johns Hopkins University*

ABSTRACT

This report gives the first detailed account of word and sub-lexical frequency in three large corpora of Cantonese. Word frequencies across the corpora have a similar structure overall, but pairwise comparisons between corpora showed low lexical overlap and weak correlations in the frequencies of individual words. By contrast, sound structure frequencies, including segment, syllable, and tone, 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 methods used to segment words from running text, encode words orthographically and phonologically, and extract token and type frequencies from large data sets, thereby providing further access to the data. Finally, we validate the word frequency data by using it as a predictor of speech error and word recognition data in Cantonese. All of these generalizations are summarized in public data sets.

Keywords: Cantonese, word frequencies, phonological frequencies, corpus linguistics

* We gratefully acknowledge Queenie Chan, Xizi Deng, James Kirby, Melvin Yap and the audience of New Empirical Contributions to Cantonese Linguistics and Language Processing (Simon Fraser University, The University of British Columbia, and The Chinese University of Hong Kong). This research was supported in part by an Insight grant from the Social Science and Humanities Research Council of Canada (435-2020-0193).

1. INTRODUCTION

It is difficult to overstate the importance of frequency in the investigation of language behavior. Frequency norms are indispensable to psycholinguistics. Classes like high frequency and low frequency items are routinely used as factors in experimental designs, and as a continuous measure, frequency is used in a variety of ways to predict language behavior (Ambridge et al. 2015; Gordon 1983; Oldfield and Wingfield 1965; Stemberger and MacWhinney 1986; Vitevitch 1997). Frequency effects are no less important to corpus linguistics, natural language processing, and second language education, given the relevance of the underlying psycholinguistic processes to these disciplines (Bird et al. 2009; Ellis 2002; Gries 2016). Frequency has also become increasingly more important to linguistic analysis, as linguists investigate the role of frequency in a host of linguistic behaviors (Bod et al. 2003; Bybee 2001; Cohen Priva et al. 2021; Frisch 1996; Shaw and Kawahara 2018).

Success in these domains therefore depends on a solid understanding of frequency effects. Our ability to examine the impact of frequency on language behavior in major Indo-European languages like English and Dutch is strong because of the existence of rich data sets for 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. For example, Anand et al. (2011) examined 550 corpora available from the Linguistic Data Consortium, one of the largest repositories of linguistic data, and found that just five languages accounted for 85% of all data sources. Clearly, a bias exists in this documentation towards majority languages with large populations and socio-economic power, leaving far fewer resources for other under-studied languages. Our primary aim here is to address this problem by giving a comprehensive account of frequency effects in Cantonese, an under-studied Chinese language of Hong Kong, southern China, and the Cantonese-speaking diaspora.

Our account builds on past work and addresses some of its shortcomings. As demonstrated in our review below, there are many corpora of Cantonese and Chinese-language corpora that include Cantonese data, but most of them are special-purpose corpora tied to

individual projects. Three large corpora have been created for general use, however, and we analyze and compare the frequency distributions from them. They are the Hong Kong Cantonese Adult Corpus ("HKCAC" henceforth, Leung and Law, 2001), the Hong Kong Cantonese Corpus ("HKCanCor"; Luke and Wong, 2015), and the IARPA Babel Cantonese Language Pack ("IARPA"; Andrus et al., 2016).

While they are important empirical contributions in their own right, these corpora are in general under-analyzed and in their current form they do not meet the needs of linguistic or psycholinguistic studies. In part due to problems with word segmentation in Chinese languages, no corpus has been used to document word frequency, leaving this critically important factor unanalyzed. Likewise, sub-lexical structure (e.g., segments, syllables, and tone) has not yet been investigated thoroughly. While segment and tone frequency have been examined in some detail in HKCAC (Leung et al. 2004), there are problems with this preliminary account that motivate the current research. First, Leung et al.'s account gives frequencies for surface representations, that is, the word forms that follow the application of neutralizing phonological rules. This means that we do not yet know the frequencies of lexical representations that are the focus of most psycholinguistic research. Second, segment type frequencies (i.e., frequencies in the lexicon) are calculated in Leung et al. (2004) from character types rather than the more standard technique using word types, which again limits the use of frequencies documented in this work. In addition to these shortcomings, no two corpora have been directly compared, and so we do not know how general the frequency effects are in the language as a whole, or which corpus is more suitable for particular studies. Finally, no work has attempted to validate the frequency norms by using them to predict language behavior.

We address all of these problems by giving the first comprehensive account of lexical and sub-lexical structures in these three corpora and validating the data with cross-corpora comparison and behavioral data. While our primary aim is an empirical one, motivated by the need for better frequency norms for Cantonese, we also make some theoretical contributions in the applicability of frequency to syllabaries of Cantonese, phonotactics and the analysis of speech errors.

The rest of this article is structured as follows. We introduce our methods in Section 2 by first explaining the linguistic structures we investigate, reviewing Cantonese language corpora, and selecting three corpora that are 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 important frequency distributions and highlighting important differences between the corpora. 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.

These results provide a standard stock of frequency norms for both word and sound frequencies and are available as open data sets (https://github.com/jane-lisy/cantfreq). In Section 5, we validate the data by using established frequency distributions to predict word recognition data from Tse et al. (2017) and also the incidence of speech errors in a new corpus of Cantonese speech errors (Alderete and Chan 2018). 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 on using 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 consonants 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 tones at the suprasegmental level. Tones are 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). (These and other important terms are given in Box 1 for easy reference.) 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*).

Box 1. Important terms and concepts

Cantonese syllable template: (C) X_1 (X_2), where (C) is an optional *onset*, X_2 can be filled with a consonant to form an optional syllable *coda*, and X_1 can be filled with an obligatory vowel or nasal to create the *nucleus*, or the nucleus can be a diphthong filling X_1X_2.

Sub-lexical structure: any linguistic structure below the word level (i.e., morphological, phonological, or phonetic)

Token frequency: counts of a structure in a text

Type frequency: counts of a structure in a lexicon

Lexical representation: the representation of a wordform before any phonological processes have applied

Surface representation: the representation of a wordform after all phonological processes have applied.

Probability of x (e.g., of a word or sound structure): the count of *x* in some text *t* divided by the total number of like items in *t*.

Phonotactic constraint: a constraint on the legal combinations of sounds in a word.

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 aː 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 11 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

[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 ":". 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.

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) because they are either 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**	**a**ː	**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, as 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 is rare among younger speakers. Acoustic studies of Hong Kong 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). While the three corpora have audio recordings associated with them, supporting acoustic methods for studying these mergers, our investigation relies on corpus methods for searching electronic written records, and so we leave the mergers for future research.

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 2, 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 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 and 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 (MacDonald 2016). 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. 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.

While the sub-lexical frequencies documented in HKCAC have had an impact on psycholinguistics, they have two shortcomings that limit their generalizability. First, their focus on surface representations means that reported frequency norms are not accurate counts of the sound structures of lexical representations. Differences between surface form and lexical form may arise from connected and casual speech, such as lenition of aspirated stops (e.g., /k^{h}œi23 wa:22/ □ [hœi23 wa:22] ‘said (3.sg)’) and long/short vowel variation (as in, [tsek55 hɐk55 ~ ha:k55] ‘immediately’). 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), and standard models of spoken word recognition also posit abstract lexical

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

representations and encode processes of activating and selecting these representations as the basic processes underlying lexical access (Luce and Pisoni 1998; Marslen-Wilson 1984). 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) were 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. While we recognize that determining what is a word in Cantonese is a non-trivial task (see e.g., Wong (2006)), and that this may have factored into the authors' decision to use characters, it is words (which are not co-extensive with characters) that are the conventional linguistic unit in calculating type frequencies. This is 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 and 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. For these reasons, we have reanalyzed the HKCAC data and also investigated frequency effects in two other large corpora.

The Hong Kong Cantonese Corpus (HKCanCor) is similar to HKCAC in that it is 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 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 represent surface phonology. A useful aspect of this data

collection is that it can be accessed through a Python package developed for it, PyCantonese (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 has 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.

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 focus on 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 filler words and reduced forms are represented 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.

Table 3. Cantonese language corpora

Project (Authors)	**Description**
A Linguistic Corpus of Mid-20th Century Hong Kong Cantonese (Chin and 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 (Yip and Matthews 2007)	A longitudinal database of eight Cantonese-English bilingual children. Intended for investigating bilingual acquisition in infants.
The Hong Kong Cantonese Child Language Corpus (Lee and 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 for investigating 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 and Law 2001)	Audio transcriptions of spontaneous radio phone-in programs. Size: ~170,000 Chinese characters.
Hong Kong Cantonese Corpus (Luke and 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.

2.3 Pre-processing: Preparing the Data for Analysis

As the three corpora are different in their data structures and collection methods, pre-processing is necessary to ensure that their frequency norms are comparable at both the word and the sound levels. This subsection explains the specific data cleaning procedures taken to ensure that the frequency counts of each corpus were represented accurately. At the word level, we segmented the three corpora and part-of-speech tagged the resulting words and then inductively defined the criteria of a word. For linguistic reasons and because it is necessary for word typing, we unified phonological representations of each word by selecting the most frequent variant, as explained in more detail below.

What counts as a word is a complex problem in Chinese linguistics (Packard 2000). Although both HKCanCor and IARPA are segmented corpora, they differ in their criteria for words, leading to potential differences in type and token counts. Given that the IARPA segmentation strategies were proprietary, we performed a step sample of the lexicons of both HKCanCor and IARPA, checking for segmentation differences every 20th word. 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 phrase 聽唔聽到 'can hear or cannot hear' t^{h}eŋ55m21t^{h}eŋ55dou25 is treated as a single word in IARPA, whereas HKCanCor treats it as three: 聽, 唔, 聽到. Our searches of the IARPA data set indicate 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.

The HKCAC is not segmented, so we applied a two-step process for segmenting words from full sentences. In the first step, we obtained a close estimate of lexical items in HKCAC by applying a recursive stick-by-longest-matching algorithm (Fung and Bigi 2015) to each HKCAC sentence. The algorithm matches, from left to right, the longest lexical item from the sample lexicon built from existing lexicons, namely the Huang dictionary (Huang 1970) and HKCanCor. It then repeats the same process to the sub-sentence without the matching lexical item until either the sentence is fully parsed or when it encounters a word unattested in the

sample lexicon. In the latter case, we manually segment the unattested word, and if needed, the previous word. These new words are added to an induced lexicon that aids in the second parse. In the second step, segmentation is done automatically through the Jieba Python package (Sun 2020), where we re-segmented HKCAC by loading the same sample lexicon and the induced lexicon from the first parse as custom dictionaries.

All three corpora were then tagged with the Universal Dependencies part of speech tagset via the PyCantonese package (Lee 2015). Words with the same orthography but different part of speech tags are considered two distinct types. Finally, we excluded all English code-switched words (though not English loans or loans from other languages), punctuations, and interjections, as our goal is to analyze the Cantonese lexicon.

In contexts where one word has multiple phonetic variants, we selected the most frequent variant as the phonological representation of the word. For example, the word 百 'hundred' has the phonological representation [pak33] because it is the most common (n = 68), compared to the other variants [pat33] (n = 9) and [pa33] (n = 1). This selection process was necessary for three reasons. First, type frequency tabulations at the sound level require a single representation. Second, the selection process eliminates phonological alternations (especially those triggered by neighboring words), which is consistent with our goal to document lexical representations. While Cantonese does not have a large set of synchronic phonological processes that produce alternations in surface forms (Pulleyblank 1997), it does have a clear set of casual speech rules that create surface phonological variation (Bauer 2013; Bauer and Benedict 1997). A well-defined protocol that minimizes the selection of casual speech variants is therefore necessary to probe the lexical representation of each word. Third, HKCAC contains more phonetic detail and in turn, more phonetic variants because of its transcription methods. By selecting the most frequent variant, it allows for comparisons between HKCAC and the two other corpora.

This use of frequency raises the question of whether the most frequent variants are typically the standard dictionary forms. To address this, we sampled words with more than one variant at random, and we found that the vast majority of selected variants are indeed the dictionary forms. There are some exceptions to this, but these are always low frequency

items with two attested forms of equal frequency. For example, the word 啞 'silent/unable to speak' has two tokens in HKCAC, one with the dictionary form [a: 25] and the other with a variant form [a: 23]. Because they were equal in frequency (both $n = 1$), the algorithm picked [a:23] by chance. We believe that finding better predictors for this task in the absence of frequency information would be a fruitful line of research, but given that there are very few affected items, we do not address this issue in this article.

These pre-processing procedures not only neutralize the main representational differences between the three chosen corpora, but also address the issues from Leung et al.'s (2004) previous analysis of Cantonese frequency. Specifically, the segmentation of HKCAC enables type calculations at the word level instead of the character/syllabic level, which is consistent with languages that do not use characters and analyses of Chinese languages that use word-based typing (Atkins et al. 1992; Bird et al. 2009). Additionally, the phonetic variant selection process captures the lexical representation of a word, which is consistent with behavioral research showing the importance of frequency in selecting canonical lexical representations (Connine et al. 2008; Pitt et al. 2011).

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 them. 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 roughly two and a half times larger than the other two lexicons. The 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

	Corpus size	Lexicon size	Lexical diversity	Corpus/lexicon
HKCAC	99,420	10,097	0.10156	9.85
HKCanCor	119,855	7,386	0.06162	16.23
IARPA	859,040	31,996	0.03725	26.85

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, where 0 = a frequency of 1) 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.

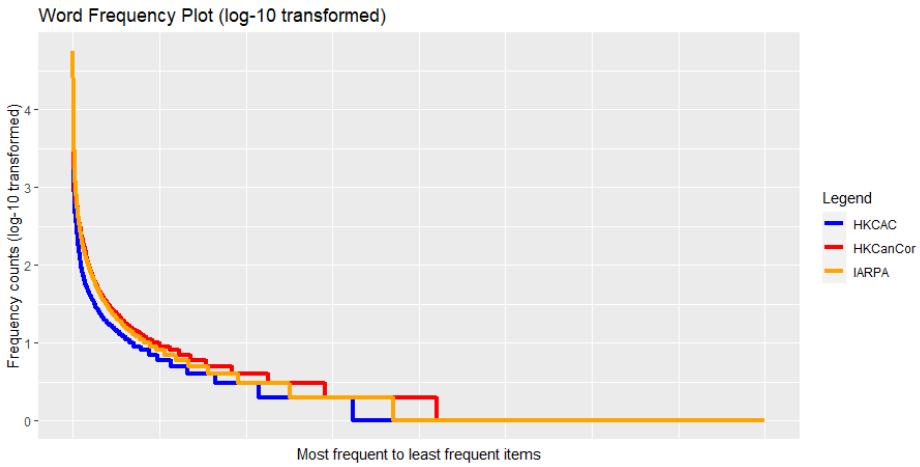

Figure 1. Word frequency for three corpora

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 this pattern 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 係. 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**		
word	IPA	counts	word	IPA	counts	word	IPA	counts
呢	le55	3664	係	hɐi22	4936	啊	a:55	57654
係	hɐi22	3634	啊	a:55	3540	你	nei23	26311
唔	m21	3052	我	ŋo23	2668	我	ŋo23	24593
嘅	ke33	3028	你	nei23	2535	係	hɐi22	24539
啊	a:55	2904	佢	k^hœi23	2224	佢	k^hœi23	17976
咁	kɐm25	2785	都	tou55	2149	唔	m21	17606
我	ŋo23	2586	呢	le55	2134	咁	kɐm25	17233
你	nei23	2445	咁	kɐm25	2102	哦	o55	13622
佢	k^hœi23	2171	唔	m21	1944	都	tou55	12999
即	tsek55	1717	㗎	ka:33	1779	冇	mou23	11761
就	tsɐu22	1458	就	tsɐu22	1759	有	jɐu23	11298
㗎	ka:33	1055	即係	tsek55hɐi22	1632	去	hœi22	10971

[3] The meaning of the function words in Table 5 are as follows: [le55] 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; [k^hœi23] 佢, 3rd

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 tokens, so *P*（呢） = 3664/99420 = 0.0369. As shown in Table 6, the probabilities of the first and second ranked items are different across corpora because IARPA's top-ranked item has a rather high probability 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 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 6, and, correspondingly, the lowest % H. L., presumably due to the sparser lexical diversity of IARPA (see above).

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 6
HKCAC	0.0369	0.0366	0.0106	0.000010	6.05	19.18
HKCanCor	0.0412	0.0295	0.0136	0.000017	2.93	15.06
IARPA	0.0661	0.0306	0.0128	0.000001	2.00	19.53

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

singular pronoun; [tsek55] 即, 'then, namely'; [tsɐu22] 就, 'then; [tou55] 都, 'also'; [ka:33] 㗎, sentence final particle (modal); [tsek55hɐi22] 即係, 'then it is, which is to say'; [mou23] 冇, 'don't have'; [jɐu23] 有, 'exist, have'; [o55] 哦, interjection; [hoei22] 去, to go.

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. We first converted the IARPA entries to traditional Chinese characters with the Python package HanziConv (Yue 2016), so that they could be matched with traditional Chinese entries in HKCAC and HKCanCor. The counts of matched lexical items are shown in Table 7 and correlations in relative word frequency (i.e., probabilities) between these shared items are given in Table8.[4] The first observation is that the three lexicons do not overlap very much. For example, HKCAC and HKCanCor have lexicons with 7,711 and 6,209 words (for stacked words), respectively, but only 2,387 shared items between them, or an overlap of roughly 30.96% (of HKCAC) and 38.44% (of HKCanCor). The second observation is that, among shared items, the corpora are well-correlated in word frequency, 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 (30.96, 38.44)	
IARPA	3,065 (39.75, 15.60)	3,391 (53.61, 17.25)

[4] While the three corpora are tagged with part-of-speech information (see section 2.4), we depart from that practice here and leave words "stacked" together (i.e., undifferentiated by part of speech). This is due to the fact that HKCAC and IARPA contain types that have an 'unclassified' POS tag (X), which will lead to an underreported percentage of shared items.

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

	HKCAC	HKCanCor
HKCAC		
HKCanCor	0.7444	
IARPA	0.6050	0.8403

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 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	65.94	46.61	61.06
HKCAC, IARPA	60.46	39.43	55.52
HKCanCor, IARPA	68.14	42.92	53.98

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.

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 type frequencies (or frequencies in the lexicon). 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 and Yip 2011).

Table 10. Word size in syllables, token frequencies

	1	2	3	4	5	n □ 6	Total
HKCAC	62,142	34,606	2,266	368	24	14	99,420
HKCanCor	85,674	31,840	1,910	364	51	16	119,855
IARPA	664,779	172,554	18,531	2,966	180	30	859,040

Table 11. Word size in syllables, type frequencies

	1	2	3	4	5	n □ 6	Total
HKCAC	1,797	6,910	1084	271	22	13	10,097
HKCanCor	1,984	4,324	740	285	41	12	7,386
IARPA	5,789	18,652	6,121	1,304	103	27	31,996

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, 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	603 (53.74)	494 (44.03)	463 (41.27)
HKCanCor	574 (51.16)	499 (44.47)	446 (39.75)
IARPA	596 (53.12)	561 (50.00)	544 (48.48)

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 still undershoot it by 147 syllable types. The most comprehensive syllabary based on lexicons (i.e., derived from adjusted types) is that of IARPA, which is missing 206 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 18.08%, though the drop in the IARPA corpus is far less (8.72%), 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

[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 Leung et al. (2004) documented many casual speech phenomena that include both new segments (e.g., *əʔ*) and new combinations due to reduction, assimilations, and casual speech phonology.

syllables of adapted loanwords (chiefly English loans), and a large number of impossible syllables, that is, 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 the 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	34	603
	HKCanCor	532	36	1	5	574
	IARPA	555	37	2	2	596
Token	HKCAC	122,120	16,515	195	265	139,095
	HKCanCor	140,486	16,392	9	13	156,900
	IARPA	982,253	96,101	8	73	1,078,435
Type	HKCAC	19,286	510	19	101	19,916
	HKCanCor	14,041	225	3	9	14,278
	IARPA	65,725	1,603	4	28	67,360

The above patterns reflect 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 because 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 (table15), presumably because of the loss of many syllables in high frequency words.

Table 14. Correlations of syllable frequencies across corpora

Comparison	Token	Type
HKCAC, HKCanCor	0.8130	0.9037
HKCAC, IARPA	0.8795	0.8830
HKCanCor, IARPA	0.8451	0.9092

Table 15. Correlations between syllabaries from type and token frequencies

	Token, Type
HKCAC	0.6744
HKCanCor	0.5216
IARPA	0.5964

Finally, we further probe syllable frequencies by investigating syllable shapes across corpora. Table 16 and Table 17give 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 frequencies.

Table 16. Syllable shape token frequencies across corpora

	CV	CVV	CVN	CVS	N
HKCAC	45,035 (32.21)	45,156 (32.30)	32,952 (23.57)	12,678 (9.07)	4,008 (2.87)
HKCanCor	49,719 (31.69)	52,674 (33.57)	33,523 (21.37)	16,973 (10.82)	4,011 (2.56)
IARPA	370,385 (34.34)	371,491 (34.45)	219976 (20.40)	82,490 (7.65)	34,093 (3.16)

Table 17. Syllable shape type frequencies across corpora

	CV	CVV	CVN	CVS	N
HKCAC	3,829 (19.01)	5,766 (28.63)	7,428 (36.88)	3,004 (14.91)	115 (0.57)
HKCanCor	2,417 (16.93)	4,196 (29.39)	5,236 (36.67)	2,306 (16.15)	123 (0.86)
IARPA	10,870 (16.14)	21,280 (31.60)	24,765 (36.77)	9,423 (13.99)	1022 (1.52)

4.3 Consonants

We now turn to the distributions of consonants across the three corpora.[6] As noted in section 2.1, some consonants (stops and nasals) can appear in both onset and coda positions, 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,

[6] Given the importance of segmental transcription, we have spot-checked the three data sets for transcription accuracy, and the first author (a native speaker of Cantonese) has confirmed that the transcripts are accurate.

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	∅	9,529	618	7,959	209	117,627	1,516
	p	3,092	826	3,680	671	27,239	3,044
	pʰ	956	272	995	253	5,798	1,125
	t	11,459	1,556	14,223	1,022	95,117	5,400
	tʰ	2,910	627	3,200	544	26,204	2,537
	k	21,762	2,275	21,787	1,305	139,382	6,440
	kʰ	3,872	343	3,762	257	28,164	1,325
	kʷ	427	154	1,732	214	2,997	610
	kʷʰ	71	40	60	28	365	135
	f	3,089	902	2,619	549	19,973	3,069
	s	8,851	2,401	9,564	1,864	63,181	7,982
	h	12,556	1,119	18,603	870	116,689	5,019
	ts	13,677	2,209	13,830	1,459	79,972	6,505
	tsʰ	3,327	978	4,047	956	31,004	4,337
	m	5,916	1,017	6,432	695	64,544	3,226
	n	296	126	7,841	269	4,214	260
	ŋ	1,308	215	3,515	152	216	135
	w	4,329	624	5,117	425	35,930	2,228
	l	14,480	1,373	8,193	923	92,397	5,149
	j	13,161	2,112	15,730	1,490	93,329	6,296
Coda	p	899	389	1,085	346	7,303	1,234
	t	5,599	1,237	6,086	862	34,207	3,362
	k	5,814	1,367	9,802	1,098	40,980	4,827
	m	6,209	754	6,930	588	36,445	2,521
	n	12,434	3,095	13,583	2,244	94,760	10,651
	ŋ	14,219	3,519	13,010	2,404	88,771	11,593
Nucleus	m	4,008	115	3,850	76	34,062	1,005
	ŋ	19	14	161	47	31	17

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	0.91069	0.98458
HKCAC, IARPA	0.94610	0.98492
HKCanCor, IARPA	0.88478	0.98692

Table 20. Correlations in token and type frequency in consonants

	Token, Type
HKCAC	0.80568
HKCanCor	0.70155
IARPA	0.69935

4.4 Vowels

We now turn to the distribution of vowels across the three corpora. Recall from Section 2.1 that monophthongs can appear in either 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 the predication marker [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	17,517	3,510	20,287	2,657	128,624	10,245
	e	11,299	527	9,619	340	49,157	2,003
	y	3,159	808	3,320	606	17,504	2,798
	œ	4,036	969	3,970	657	22,713	2,942
	u	5,227	1,583	5,007	1,102	32,315	5,331
	o	17,828	1,988	18,447	1,052	144,664	5,846
	aː	15,226	2,101	21,993	1,749	185,362	9,208
	ɐ	15,653	2,561	17,572	1,796	101,322	7,667
Diphthongs	ei	8,386	898	8,437	644	62,934	2,671
	œi	4,788	489	4,860	342	40,138	1,416
	ui	1,069	224	1,090	92	3,153	456
	oi	1,597	361	1,285	250	8,895	1,312
	ɐi	10,355	785	14,761	693	85,538	3,124
	ai	2,069	479	2,849	352	20,561	2,070
	iu	1,756	405	1,835	266	14,723	1,423
	eu	5	3	3	3	0	0
	ou	7,921	1,030	9,880	749	73,975	3,792
	ɐu	6,578	870	6,801	654	47,373	3,207
	au	599	196	873	151	5,391	827

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	0.98519
HKCAC, IARPA	0.92989	0.97153
HKCanCor, IARPA	0.97352	0.98009

Table 23. Correlations between type and token frequency in vowels

	Token, Type
HKCAC	0.84940
HKCanCor	0.84170
IARPA	0.86371

4.5 Tone

Finally, we report on the distribution of the 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	Total
CV(V)	18,986	15,925	21,497	10,291	17,128	22,577	106,404
CVN	9,295	10,102	2,573	6,257	786	4,510	33,523
CVS	8,241	130	3,157	1	0	5,444	16,973
Total	36,522	26,157	27,227	16,549	17,914	32,531	

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	28,777 (20.92)	22,627 (16.45)	26,527 (19.29)	16,885 (12.28)	14,841 (10.79)	27,879 (20.27)
HKCanCor	36,522 (23.28)	26,157 (16.67)	27,227 (17.35)	16,549 (10.55)	17,914 (11.42)	32,531 (20.73)
IARPA	321,924 (29.85)	141,034 (13.08)	174,181 (16.15)	101,637 (9.42)	138,101 (12.81)	201,558 (18.69)

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	6,016 (27.57)	3,149 (14.43)	3,338 (15.30)	3,803 (17.43)	944 (4.33)	4,570 (20.94)
IARPA	18,378 (27.28)	10,093 (14.98)	10,531 (15.63)	11,861 (17.61)	4,004 (5.94)	12,493 (18.55)

Frequency distributions for tone, however, are affected by context, and this needs to be factored into calculations of the impact of frequency on 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. By contrasting the percentage occurrence in $\square_1$ versus $\square_2$, 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 is more muted in the latter. As far as how these trends relate to Cantonese tone, 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 words (Chen 2000; Yip 1980), the opposite of what is found here. On the other hand, recent research has uncovered a default 'high-low' T1-T4 pattern in incorporated English loans (see Mok and Lee (2018)), which could account for the drop in frequency of T1 from the first to the second syllable.

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	9,655 (28.08)	6,149 (17.88)	4,254 (12.37)	5,306 (15.43)	3,245 (9.44)	5,773 (16.79)
	σ_2	6,601 (19.10)	5,604 (16.21)	7,459 (21.58)	4,151 (12.01)	992 (2.87)	9,754 (28.22)
HKCanCor	σ_1	10,162 (31.92)	6,174 (19.40)	3,284 (10.31)	5,458 (17.13)	2,407 (7.56)	4,355 (13.68)
	σ_2	6,277 (19.68)	4,863 (15.27)	4,107 (12.90)	3,644 (11.44)	2,235 (7.02)	10,724 (33.67)
IARPA	σ_1	54,004 (31.88)	37,455 (21.71)	16,413 (9.51)	24,251 (14.05)	13,330 (7.73)	26,101 (15.13)
	σ_2	41,274 (23.92)	24,873 (14.42)	23,102 (13.39)	23,746 (13.76)	12,750 (7.39)	46,809 (27.13)

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	1,845 (27.06)	1,074 (15.75)	1,173 (17.21)	970 (14.23)	537 (7.88)	1,218 (17.87)
	σ_2	1,483 (21.57)	1,006 (14.63)	1,429 (20.78)	974 (14.17)	351 (5.10)	1,633 (23.75)
HKCanCor	σ_1	1,233 (28.52)	660 (15.26)	683 (15.80)	736 (17.02)	255 (5.90)	757 (17.51)
	σ_2	913 (21.11)	741 (17.14)	723 (16.72)	747 (17.28)	206 (4.76)	994 (22.99)
IARPA	σ_1	5,430 (29.11)	2,627 (14.08)	3,032 (16.26)	3,029 (16.24)	1,154 (6.19)	3,380 (18.12)
	σ_2	4,732 (25.37)	2,919 (15.65)	2,981 (15.98)	3,271 (17.54)	1,059 (5.68)	3,690 (19.78)

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 and White 2013; Prince and 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. We list seven constraints in Table 29 that have played important roles in generative phonological accounts of sound structure in Cantonese (Cheng 1991; Yip 1997). The constraints are assumed to constitute systematic gaps in Cantonese sound structure, and indeed many have a logic to them that relate to cross-linguistic constraints on feature structure, like the avoidance of identical place

specifications. Each constraint is stated in featural terms relative to syllabic roles, with specific banned phoneme sequences given on the last line.[7] The frequencies reported here show that most constraints are respected by all corpora, and in many cases (excluding HKCAC), 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). This finding accords with the linguistic research, which qualifies some of these constraints (e.g., (a) Table 29, Cheng 1991) by stating that violations do occasionally occur in loans and onomatopoeic expressions.

[7] 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*, and cor(onal) for coronal sounds using the front and tip of the tongue.

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 p^h m f k^w k^{wh}/	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 p^h m f k^w k^{wh}/+/y œ/	3	2	0	0	0	0
d. *Ons Nuc Ons cor [+bk, +rd] cor */t t^h s n l/ + /o, u/ + cor	11	7	0	0	0	0
e. *Nuc Coda e lab/cor *em en ep et	4	4	4	1	0	0
f. *Ons Nuc cor u */t t^h s n l/ + u	2	2	0	0	0	0
g. *Nuc Coda [+high] dorsal *ik iŋ yk yŋ uk uŋ	5	2	0	0	0	0

5. DATA VALIDATION

One critical application of frequency norms is its predictive power on psycholinguistic processes. For instance, numerous studies have shown that high frequency words are processed more quickly than low frequency words in word recognition and lexical decision tasks in various languages (Bates et al. 2003; Levelt and Wheeldon 1994; Oldfield and Wingfield 1965). Previous reports of Cantonese frequency norms, such as Leung et al. (2004), have not validated their frequency norms against behavioral

data. In the subsections below, we attempt to do so with word frequency against two data sets, word recognition reaction times and accuracy (Section 5.1) and incidence of speech errors (Section 5.2). By conducting novel data validation tasks, we hope to assess the applicability of Cantonese spoken corpora in experimental settings, as well as provide a comparison between frequency norms of different sizes, populations, and sources (e.g., conversations versus written text). These two tasks, one on word comprehension, another on word production, also illustrate the uses of our frequency norms in psycholinguistic studies.

5.1 Word Recognition Data

We validated our lexical frequencies against written Chinese word recognition data elicited in Tse et al. (2017), and adopted many of the methodological assumptions from this megastudy. Tse et al. collected reaction times and response accuracies from native Cantonese speakers participating in a lexical decision task, where participants decided if two-character strings constituted a Cantonese word, typically a compound. They analyzed their experimental data by computing the R^2 values for the correlations between reaction time and accuracy scores against frequency norms derived from several large written Chinese or Mandarin corpora.

Aside from investigating the predictive power of our frequency norms, our analysis below also extends the empirical scope of Tse et al. (2017) to spoken Cantonese corpora. The frequency norms from Tse et al. (2017) were all based on written Chinese or Mandarin corpora, not Cantonese corpora, and they concluded that a large corpus of film and television subtitles that are not specific to Cantonese (Cai and Brysbaert 2010) was the best predictor of the word recognition data. We are interested in asking whether our frequency norms derived from spoken Cantonese corpora are better predictors of the word recognition variance. We hypothesize that since spoken Cantonese corpora bear more resemblance to the daily speech of the participants of the lexical decision task than spoken Mandarin or written Chinese corpora, it should have stronger predictive power. Tse et al.'s results therefore provide us with a baseline to determine whether our frequency norms were effective in predicting word recognition variance.

Tse et al.'s database, the repository of word recognition data, has 25,281 items, but these words were winnowed down to 22,808 items that have at least 70% accuracy in the lexical decision task, and reduced further to 15,759 items that are shared across the six corpora investigated. To standardize the data, the authors first log-transformed (base 10) all frequency counts, then subtracted each log-transformed value by the mean value across all available items that have larger or equal to 70% accuracy in that norm (i.e., non-zero counts within the 22,808 items). R^2 (variance explained by predictor; Pearson correlation squared) values were then derived from a combination of the response variables (zRT or accuracy). It is important to emphasize that, while the six corpora investigated were not from Cantonese spontaneous speech, they are extremely large corpora. For example, the frequency norms from Cai and Brysbaert (2010) are based on a corpus of 33.5 million words, which is an order of magnitude larger than any of the corpora we examined here.

To compare our data, we performed two similar analyses with five-word frequency norms, namely norms based on HKCAC, HKCanCor, IARPA, all Cantonese corpora (i.e., HKCAC + HKCanCor + IARPA), and all Hong Kong based corpora (i.e., HKCAC + HKCanCor). Since Tse et al.'s (2017) frequency norms are not specified by part of speech, our counts for each item in the word recognition database is the sum of the frequencies of all lexical items with the same orthographic form. Analysis 1 below was restricted to a set of lexical items in the word recognition database that occurred in all six frequency norms from Tse et al. (2017), as well as our five norms, which yielded 840 items. Analysis 2 removed the restrictions imposed on the set of lexical items, generating larger sets for our R^2 calculations. Instead of requiring non-zero counts on all 11 measures, we ran separate analyses for each of our five norms, though each of these were required to have lexical items in all of the six Tse et al. (2017) norms. The size of each sub-analysis is reported in Table 30 below. The values reported for Analysis 2 for the Tse et al. (2017) frequency counts are averaged across the five sub-analyses.

Contrary to expectation, our Cantonese frequency norms were not better predictors than the corpora reported in Tse et al. (2017) for zRT or accuracy scores. Thus, in both analyses, the Cantonese language frequency norms achieved far lower R^2 scores than the Tse et al. (2017)

frequency norms (t = -3.31, df = 33.381, $p < 0.0025$). Within the five Cantonese measures, the total frequencies norm (i.e., HKCAC + HKCanCor + IARPA) performed the best in both zRT and accuracy R^2 scores, suggesting that corpus size matters.

Table 30. Percentage of zRT and accuracy variance explained by 11 frequency norms

	Analysis 1 (n = 840)		**Analysis 2**		
Corpus	**zRT (%)**	**Accuracy (%)**	**zRT (%)**	**Accuracy (%)**	**Size**
Da (2004) News	8.53	1.67	15.382	4.198	12.5m
Da (2004) Fiction	14.33	1.36	22.932	4.502	15.8m
Shaoul et al. (2016)	15.95	2.58	24.924	6.772	358b
Google frequencies	8.50	0.97	25.972	8.14	*NA*
Cai & Brysbaert (2010) word frequencies	25.26	3.55	35.004	9.426	33.5m
Cai & Brysbaert (2010) contextual diversity frequencies	24.72	3.52	34.968	9.482	33.5m
HKCAC	4.47	0.13	12.30	1.47	2,315
HKCanCor	3.97	0.14	9.14	1.25	2,094
IARPA	5.74	0.13	11.07	2.17	3,915
Total frequencies	6.47	0.17	13.74	2.13	3,336
HK Cantonese frequencies	5.18	0.12	15.17	3.21	5,404

The discrepancy between the six Tse et al. frequency norms and ours can be explained by two hypotheses. The first is that the size of the corpora had an effect on the R^2 scores—as a corpus grows in size, its variance and explanatory power grows as well. To run a preliminary test on this hypothesis, we calculated the correlation between the frequency sum (after log-transformation and scaling) of the 840 lexical items in each measure

and its zRT and accuracy R^2. We found that both zRT and accuracy R^2 were strongly correlated with corpus size, with $r = 0.78$ and $r = 0.84$, respectively.

Another possible explanation is that the lexical items employed in the word recognition study included mostly words that are in Standard Written Chinese, which are more similar to the written Chinese and spoken Mandarin corpora than the spoken Cantonese corpora (Bauer 2018), yielding higher predictive power for this large subset of data. To test this alternative hypothesis, we further sub-divided the original set of 840 lexical items manually into 54 "exclusively Cantonese" words that are unique to the Cantonese language and the remaining 786 words that are not (these aggregated lexical items are available from the GitHub project page). If it is the case that spoken Cantonese frequency norms are better predictors for these exclusively spoken words, we should observe increased R^2 values in the Cantonese-exclusive condition for our five Cantonese frequency norms and decreased R^2 values for the same words with the non-Cantonese norms.

As shown in Table 31, we observe that in the exclusive condition, the R^2 scores of Cantonese frequency norms have increased and are on par with the Mandarin/written Chinese corpora in explaining accuracy rates. The correlations with zRT have also decreased in the exclusive condition for non-Cantonese norms, but they still far exceed the five Cantonese norms in percentage explained. On the other hand, we found that Mandarin/written Chinese corpora performed better in the written-spoken condition. The results of this analysis are therefore consistent with our second hypothesis in terms of accuracy R^2, but only partly so for zRT R^2. The reason behind this unexpected difference is unclear, but we suspect that sample and frequency count size had an effect on R^2, just as we conjectured in the first analysis.

Table 31. Percentage of zRT and accuracy variance, sub-divided by Cantonese exclusive and non-exclusive words

	Cantonese Exclusive (n = 54)		**Written + Spoken (n = 786)**	
Corpus	zRT (%)	Accuracy (%)	zRT (%)	Accuracy (%)
Da (2004) News	0.79	1.02	6.25	1.8
Da (2004) Fiction	0.88	0.27	14.48	1.51
Shaoul et al. (2016)	6.43	0.23	13.17	2.79
Google frequencies	7.54	0.23	6.82	0.79
C & B (2010) word frequencies	16.79	2.88	22.92	3.09
C & B contextual diversity	15.29	2.54	22.61	3.11
HKCAC	1.33	1.33	5.39	0.09
HKCanCor	0.52	2.77	6.16	0.11
IARPA	2.81	2.14	7.73	0.11
Total frequencies	2.58	2.49	8.81	0.14
HK Cantonese frequencies	0.87	1.42	7.03	0.1

In sum, while our data did not outperform Tse et al.'s (2017) frequency norms in general, we note that the disparity in size and the small amount of exclusively Cantonese words had an effect on the amount of variance explained. For future work, we recommend further aggregation of both spoken Cantonese and written Chinese data sets, leading to larger baselines for the latter, as we have observed an increase in R^2 scores by aggregating our frequency norms.

5.2 Speech Error Data

Speech errors involving mis-selections of sounds are more likely to occur in low frequency words than high frequency words (Dell 1990; Stemberger and MacWhinney 1986). While this finding has been

investigated in English, it has not been documented extensively in other languages. In a first of its kind for a Chinese language, we examine the impact of frequency norms on the incidence of speech errors in Cantonese. This contribution both provides an opportunity to validate our word frequency norms, and contributes new data to research on the impact of frequency on language production.

The speech error data were drawn from SFUSED Cantonese 1.0 (Alderete and Chan 2018), a large corpus of speech errors collected from natural conversations. We extracted 840 speech errors from this collection involving sound substitutions in content words (i.e., noun, verb, adjective, or adverb). Following a procedure from Stemberger and MacWhinney (1986), we extracted all content word items from each corpus and determined the midpoint of the frequency distribution of all tokens. Let the frequency of the lexical item with the midpoint token be *m*. All lexical items with a frequency larger or equal to *m* are considered high frequency, and all items with a frequency lesser than *m* are low frequency. We then perform one analysis per corpus, where we first subset words that occur in both the sound substitution set and the corpus, then count the number of high and low frequency items within that subset. Results, together with frequency group (i.e., high vs. low), are summarized in Table 32. For each frequency norm, we performed a Chi-squared test with one degree of freedom.

Table 32. High versus low frequency speech errors, sub-divided by midpoint frequency.

	HKCAC	**HKCanCor**	**IARPA**
# of low frequency tokens	29654	30961	272497
# of high frequency tokens	29693	31068	272815
Midpoint frequency	77	183	1330
# of low frequency errors	249	261	268
# of high frequency errors	52	38	34
$\chi^2(1)$	$\chi^2=128.93$	$\chi^2=166.32$	$\chi^2=179.89$
p-value	$p < 0.00001$	$p < 0.00001$	$p < 0.00001$

For all three corpora, the low frequency group had significantly more errors than the high frequency group. This finding corroborates the

findings of Stemberger and MacWhinney (1986) for English speech errors. While a partition between low and high frequency counts is a relatively coarse measure for a data validation task, it supports the validity of the three frequency norms. To further explore our data, we employed a sampling technique from Vitevitch (1997) to test for the difference between low and high frequency items. In particular, we sampled at random (without replacement) from each of our corpora a set of words that is 10 times larger than the amount of error items matched in the error validation. For each sample, we counted the amount of high frequency items with the same criteria as previously described. The counts for each sample, compared against the speech error sampling is shown in Figure 2. Consistent with our findings above, we find that there were far more high frequency items in a random sample (chance estimate) than there are speech errors in high frequency words. This further supports the predictive power of our frequency norms, especially in spoken Cantonese phenomena.

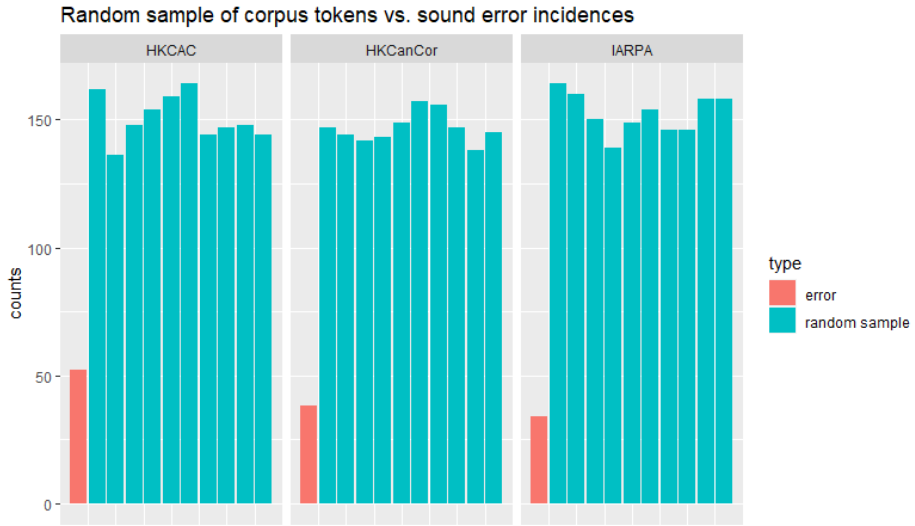

Figure 2. Distribution of high frequency tokens sampled from sound error data set (red) and corpus tokens (blue)

6. DISCUSSION

6.1 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 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 considered.

6.2 Applications to Experimental Design

In Section 5, we gave two concrete examples of how the frequency norms can be applied to word comprehension and production studies. Here, we ask more generally how they apply to experimental designs and decisions about experimental stimuli. The question of how to use the

frequency norms is more important for word frequency than for sound structure 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 with part-of-speech tagging. Although all three corpora are part-of-speech-tagged, HKCanCor does not have unclassified tags (as in a small percentage of IARPA and HKCAC words), due to the original dataset being tagged and manually verified. 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 it has 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).

When these corpora are used to select specific stimuli, chosen for the frequency characteristics discussed above, researchers must of course be mindful of the dialect differences that exist and that not all lexical items from, for example, Hong Kong Cantonese will be recognized by native speakers of Guangzhou Cantonese. We hope that the data supplements we

created will give researchers an abundance of stimuli as potentially fitting their research designs to avoid this problem.

6.3 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 was 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 also speak English 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 (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.

REFERENCES

Alderete, John, and Queenie Chan. 2018. Simon Fraser University Speech Error Database - Cantonese 1.0 (First release). Retrieved January 01, 2021, from: https://www.sfu.ca/people/alderete/sfused.html.

Ambridge, Ben, Evan Kidd, Caroline F Rowland, and Anna L Theakson. 2015. The ubiquity of frequency effects in first language acquisition. *Journal of Child Language* 42: 239-272.

Anand, Pranav, Sandra Chung, and Matthew Wagers. 2011. *Widening the net: Challenges for gathering linguistic data in the digital age* (NSF SBE 2020. ID 121). National Science Foundation Directorate of Social, Behavioral, and Economic Sciences.

Andrus, Tony, Eyal Dubinski, Jonathan Fiscus, Breanna Gillies, Mary Harper, T.J. Hazen, Brook Hefright, Amy Jarrett, Willa Lin, Jessica Ray, Anton Rytting, Wade Shen, Evelyne Tzoukermann, and Jamie Wong. 2016. IARPA Babel Cantonese Language Pack IARPA-babel101b-v0.4c LDC2016S02. Retrieved June 01, 2019, from: https://catalog.ldc.upenn.edu/LDC2016S02.

Atkins, Sue, Jeremy Clear, and Nicholas Ostler. 1992. Corpus design criteria. *Literary and Linguistic Computing* 7: 1-16.

Baayen, R. Harald, R Piepenbrock, and L Gulikers. 1996. CELEX2. Retrieved June 01, 2019, from: https://catalog.ldc.upenn.edu/LDC96L14.

Bates, Elizabeth, Simona D'amico, Thomas Jacobsen, Anna Székely, Elena Andonova, Antonella Devescovi, Dan Herron, Ching Ching Lu, Thomas Pechmann, Csaba Pléh, Nicole Wicha, Kara Federmeier, Irini Gerdjikova, Gabriel Gutierreze, Daisy Hung, Jeanne Hsu, Gowri Iyer, Katherine Kohnert, Teodora Mehotcheva, Araceli Orozco-Figueroa, Angela Tzeng, and Ovid Tzeng. 2003. Timed picture naming in seven languages. *Psychonomic Bulletin & Review* 10: 344-380.

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

Bauer, Robert S. 2018. Cantonese as a written language in Hong Kong. *Global Chinese* 4: 103-142.

Bauer, Robert S., and Paul K. Benedict. 1997. *Modern Cantonese phonology*. Berlin: Mouton de Gruyter.

Bauer, Robert S., Kwan-hin Cheung, and Pak-man Cheung. 2003. Variation and merger of the rising tones in Hong Kong Cantonese. *Language Variation and Change* 15: 211-225.

Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. Sebastopol: O'Reilly Media, Inc.

Bod, Rens, Jennifer Hay, and Stefanie Jannedy. 2003. *Probablistic linguistics*. Cambridge: The MIT Press.

Bybee, Joan. 2001. *Phonology and language use*. Cambridge: Cambridge University Press.

Cai, Qing, and Marc Brysbaert. 2010. SUBTLEX-CH: Chinese word and character frequencies based on film subtitles. *PLoS ONE* 5(6): e10729.

Chao, Yuen Ren. 1930. A system of tone letters. *Le Maître Phonétique* 45: 24-27.
Chen, Matthew. 2000. *Tone sandhi: Patterns across Chinese dialects*. Cambridge: Cambridge University Press.
Cheng, Lisa Lai-Shen. 1991. Feature geometry of vowels and co-occurrence restrictions in Cantonese. In *Proceedings of the 9th West Coast Conference on Formal Linguistics 9*, California, pp.107-124.
Chin, C. O., and A. M. Tweed. 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, April 13, 2019, The Hong Kong Polytechnic University, Hong Kong.
Cohen Priva, Uriel, Emily Strand, Shiying Yang, Abigail Creighton, Justin Bai, Rebecca Mathew, Allison Shao, Jordan Schuster, and Daniela Wiepert. 2021. The cross-linguistic phonological frequencies (XPF) corpus. Retrieved June 01, 2021, from: https://www.urielcohenpriva.com/resources/xpf_manual07.pdf.
Connine, Cynthia M., Larissa J. Ranbom, and David J. Patterson. 2008. Processing variant forms in spoken word recognition: The role of variant frequency. *Perception and Psychophysics* 70: 403-411.
Dell, Gary S. 1990. Effects of frequency and vocabulary type on phonological speech errors. *Language and Cognitive Processes* 5: 313-349.
Dell, Gary S., Cornell Juliano, and Anita Govindjee. 1993. Structure and content in language production: A theory of frame constraints in phonological speech errors. *Cognitive Science* 17: 149-195.
Dell, Gary S., Nazbanou Nozari, and Gary M. Oppenheim. 2014. Word production: Behavioral and computational considerations. In *The Oxford handbook of language production*, eds. M. Goldrick, V. Ferreira, and M. Miozzo, pp.88-104. Oxford: Oxford University Press.
Ellis, Nick C. 2002. Frequency effects in language processing. *Studies in Second Language Research* 24: 143-188.
Fletcher, Paul, C. S. S Leung, S Stokes, and Z Weizman. 2000. *Cantonese pre-school language development. A guide*. Hong Kong: Hong Kong: Department of Speech and Hearing Sciences.
Frisch, Stefan A. 1996. *Similarity and frequency in phonology*. Evanston: Northwestern University dissertation.
Frisch, Stefan A., Nathan R. Large, and David S. Pisoni. 2000. Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. *Journal of Memory and Language* 42: 481-496.
Fung, R, and B Bigi. 2015. Automatic word segmentation for spoken Cantonese. Paper presented at the International Conference Oriental COCOSDA, October 28-30, 2015, Jiao Tong University, Shanghai, China.
Goldrick, Matthew. 2004. Phonological features and phonotactic constraints in speech production. *Journal of Memory and Language* 51: 586-603.
Gordon, Barry. 1983. Lexical access and lexical decision: Mechanisms of frequency sensitivity. *Journal of Verbal Learning and Verbal Behavior* 22: 24-44.

Gries, Stefan Th. 2015. Quantitative designs and statistical techniques. In *The Cambridge handbook of English corpus linguistics*, eds. D. Biber, and R. Reppen, pp.50-71. Cambridge: Cambridge University Press.

Gries, Stefan Th. 2016. *Quantitative corpus linguistics with R: A practical introduction*. New York: Routledge.

Hay, Jennifer, Janet Pierrehumbert, and Mary Beckman. 2004. Speech perception, well-formedness and the statistics of the lexicon. In *Phonetic interpretation: Papers in laboratory phonology VI*, eds. J. Local, R. Ogden, and R. Temple, pp.58-74. Cambridge: Cambridge University Press.

Hayes, Bruce, and James White. 2013. Phonological naturalness and phonotactic learning. *Linguistic Inquiry* 44: 45-75.

Hong Kong Polytechnic, University. 2015. PolyU Corpus of Spoken Chinese. Retrieved June 1, 2019, from: http://www4.lt.cityu.edu.hk/~tswong/corpus.htm.

Huang, Parker Po-Fei. 1970. *Cantonese dictionary: Cantonese - English, English - Cantonese*. New Haven and London: Yale University Press.

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

Johnson, Khia, Molly Babel, Ivan Fong, and Nancy Yiu. 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*, Marseille, pp.4082-4088.

Kessler, Brett, and Rebecca Treiman. 1997. Syllable structure and the distribution of phonemes in English syllables. *Journal of Memory and Language* 37: 295-311.

Kilgarriff, Adam. 2001. Comparing corpora. *International Journal of Corpus Linguistics* 6: 97-133.

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

Lee, Jackson L. 2015. *PyCantonese Python package*. Chicago: University of Chicago.

Lee, Thomas Hung-Tak, and Colleen Wong. 1998. CANCORP: The Hong Kong Cantonese Child Language Corpus. *Cahiers de Linguistique Asie Orientale* 27: 211-228.

Leung, Man Tak, and Sam-Po Law. 2001. HKCAC: The Hong Kong Cantonese Adult Language Corpus. *International Journal of Corpus Linguistics* 6: 305-326.

Leung, Man Tak, Sam-Po Law, and Suk-Yee Fung. 2004. Type and token frequencies of phonological units in Hong Kong Cantonese. *Behavior Research Methods, Instruments, and Computer* 36: 500-505.

Levelt, Willem J. M, Ardi Roelofs, and Antje S Meyer. 1999. A theory of lexical access in speech production. *Behavioral and Brain Sciences* 22: 1-75.

Levelt, Willem J. M, and Linda Wheeldon. 1994. Do speakers have access to a mental syllabary?. *Cognition* 50: 239-269.

Levitt, Andrea, and Alice Healy. 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, Juliana. 2017. Cantonese v Mandarin: When Hong Kong languages get political. *BBC News*. Retrieved January 01, 2021, from: https://www.bbc.com/news/world-asia-china-40406429

Luce, Paul A., and David B. Pisoni. 1998. Recognizing spoken words: The neighborhood activation model. *Ear Hear* 19: 1-36.

Luke, Kang-Kwong, and May Lai-Yin Wong. 2015. The Hong Kong Cantonese Corpus: Design and uses. *Journal of Chinese Linguistics* 25: 309-330.

MacDonald, Maryellen C. 2016. Speak, act, remember: The language-production basis of serial order and maintenance in verbal memory. *Current Directions in Psychological Science* 25: 47-53.

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

Matthews, Stephen, and Virginia Yip. 2011. *Cantonese: A comprehensive grammar*. London: Routledge.

Mok, Peggy P.-K, and Albert Lee. 2018. The acquisition of lexical tones by Cantonese-English bilingual children. *Journal of Child Language* 45(6): 1-20.

Mok, Peggy P.-K., Dongui Zuo, and Peggy W.-Y. Wong. 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, and A Wingfield. 1965. Response latencies in naming objects. *Quarterly Journal of Experimental Psychology* 17: 273-281.

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

Pitt, Mark A, Laura Dilley, and Michael Tat. 2011. Exploring the role of exposure frequency in recognizing pronunciation variants. *Journal of Phonetics* 39: 304-311.

Prince, Alan, and Bruce Tesar. 2004. Learning phonotactic distributions. In *Fixing priorities: Constraints in phonological acquisition*, eds. R. Kager, and J. Pater, pp.245-291. Cambridge: Cambridge University Press.

Pulleyblank, Edwin. 1997. The Cantonese vowel system in historical perspective. In *Studies in Chinese phonology*, eds. W. Jialing, and N. Smith, pp.185-217. Mouton de Gruyter: Berlin.

Roland, Douglas, Frederic Dick, and Jeffrey Elman. 2007. Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language* 57: 348-379.

Shaw, Jason, and Shigeto Kawahara. 2018. Predictability and phonology: Past, present, and future. *Linguistic Vanguard* 4: 20180042.

So, L.K.H. 1992. Hong Kong spoken Cantonese database. Retrieved June 1, 2019, from: https://hub.hku.hk/handle/10722/113983.

Stemberger, Joseph P, and Brian MacWhinney. 1986. Frequency and the lexical storage of regularly inflected forms. *Memory & Cognition* 14: 17-26.

Sun, Junyi (Producer). 2020. "Jieba" Chinese text segmentation (Version 0.42) [Computer software]. Available: https://pypi.org/project/jieba/.

Treiman, Rebecca, Brett Kessler, Stephanie Knewasser, Ruth Tincoff, and Margo Bowman. 2000. English speakers' sensitivity to phonotactic patterns. In *Papers in laboratory phonology V: Acquisition and the lexicon*, eds. M.B. Broe, and J.B. Pierrehumber, pp.269-282. Cambridge: Cambridge University Press.

Tse, Chi-Shing, Melvin J. Yap, Yuen-Lai Chan, Wei Ping Sze, Cyrus Shaoul, and Dan Lin. 2017. The Chinese Lexicon Project: A megasudy of lexical decision performance for 25,000+ traditional two-character compound words. *Behavior Research Methods* 49: 1503-1519.

Vitevitch, Michael S. 1997. The neighborhood characteristics of malapropisms. *Language and Speech* 40: 211-228.

Wong, Wai Yi Peggy. 2006. *Syllable fusion in Hong Kong Cantonese connected speech*. Columbus: The Ohio State University dissertation.

Xu, L.J, and T Lee. 1998. *Parametric variation in three Chinese dialects, Cantonese, Shanghainese and Mandarin*. Hong Kong: Research Grant Council.

Yip, Moira. 1980. *The tonal phonology of Chinese*. Cambridge: MIT dissertation.

Yip, Moira. 1997. Consonant-vowel interaction in Cantonese. In *Studies in Chinese phonology*, eds. J. Wang, and N. Smith, pp.251-274. Berlin: Mouton de Gruyter.

Yip, Virginia, and Stephen Matthews. 2007. *The bilingual child: Early development and language contact*. Cambridge: Cambridge University Press.

Yue, Bernard (Producer). 2016. Simplified and traditional Chinese character conversion (Version 0.3.2) [Computer software]. Available: https://pypi.org/project/hanziconv/.

Zipf, George K. 1949. *Human behavior and the principle of least effort: An Introduction to Human Ecology*. Cambridge: Addison-Wesley.

[Received 24 July 2021; revised 26 November 2021; accepted 6 January 2022]

Jane Li
Department of Cognitive Science, Johns Hopkins University
3400 N. Charles Street Baltimore
sli213@jhu.edu

Heikal Badrulhisham
Department of Linguistics, Simon Fraser University
8888 University Dr W, Burnaby

John Alderete
Department of Linguistics, Simon Fraser University
8888 University Dr W, Burnaby

APPENDICES

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, part-of-speech category, and word size (i.e., numbers of syllables). 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’
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	唯 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’

粵語單詞、子詞、聲音結構頻率分析

Jane Li[1,2], Heikal Badrulhisham[1], John Alderete[1]
西門菲莎大學 [1]
約翰霍普金斯大學 [2]

這份報告首次詳細介紹了三個大型粵語語料庫中的單詞和子詞頻。此三個語料庫中的詞頻總體上具有相似的結構，但語料庫之間的成對比較顯示詞彙重疊較低，單個詞的頻率相關性較弱。相比之下，聲音結構的頻率，包括片段、音節和聲調，都具有良好的相關性，但由於類型/標記、正字法編碼、詞的位置和口語文本類型的區別，聲音結構的頻率相關性中仍然出現重要的差異。這些差異提醒廣東話的心理語言學研究學者，實驗條件標準必須包括頻率。此外，我們記錄了如何從正文中分割單詞、如何對單詞進行拼寫和語音編碼、以及從大型語料庫中提取標記和類型頻率的方法，從而提供對數據的進一步分析。最後，我們通過 預測粵語語音錯誤及單詞識別來驗證詞頻數據。所有這些發現都在開放數據集中進行了總結。

關鍵字：粵語、單詞／子詞頻率、聲音結構頻率、語料庫語言學