Incorporating tone when modeling wordlikeness judgments

Abstract

Various phonotactic models have been proposed to predict speakers' wordlikeness judgments but most have focused primarily on segments. This article aims to model speakers' wordlikeness judgments incorporating tone. We first show how the two major determinants of wordlikeness judgments, namely phonotactic probability and neighborhood density, can be applied to tone languages. To test the role of the two determinants to wordlikeness judgments in a tone language, judgment data are obtained from speakers of Cantonese. The results are then used to model speakers' judgments, showing that phonotactic probability, but not neighborhood density, modulates wordlikeness judgments and that the phonotactic probabilities involving nucleus and coda are most relevant to wordlikeness judgments. We also show that phonotactic probability affects the tendency to judge items being absolutely perfect or more or less wordlike, while it does not affect the judgments that an item is absolutely not-wordlike. Implications of these results for phonotactic modeling and processes involved in wordlikeness judgments are discussed.

Keywords: wordlikeness, phonotactic probability, neighborhood density, tone

1. Introduction

Wordlikeness denotes the degree to which a sound sequence is considered typical in a language (Bailey & Hahn, 2001). Native speakers have consistent intuitions about which sound sequences are more wordlike. They can not only tell which existing sequences sound more like typical words (e.g., 'bag' [bæg] is more typical than 'squad' [skwad] in English) but also make a similar judgement for nonwords (e.g., 'bnick' [bnɪk] sounds more wordlike than 'bdick' [bdɪk] in English). In the research of phonotactics, one core interest has been on finding sources of such wordlikeness judgments. Previous literature has shown evidence for various sources of wordlikeness judgments, including phonotactic probability (Coleman & Pierrehumbert, 1997; Dankoviþová *et al*., 1998; Frisch *et al*., 2000; Gathercole & Martin, 1996; Vitevitch *et al*., 1997), lexical neighborhood density (Bailey & Hahn, 2001; Gathercole & Martin, 1996; Greenberg & Jenkins, 1964), and orthotactic probability (Bailey & Hahn, 2001). 'Phonotactic probability' denotes the probability of finding a phoneme's substring (e.g., how likely to have the sequence [st] in the English). 'Neighborhood density' is the degree to which a sound sequence overlaps with existing words in a lexicon. 'Orthotactic probability' refers to written letters, not sounds, calculated similarly to phonotactic probability.

By incorporating these sources, proposals have been made to model wordlikeness judgments, including syllabic parser (Coleman & Pierrehumbert, 1997), Generalized Neighborhood Model (Bailey & Hahn, 2001), Phonotactic Probability Calculator (Vitevitch & Luce, 2004), Phonotactic Learner (Hayes & Wilson, 2008), Featural Bigram Model (Albright, 2009), Simple Bigram Model (Jurafsky & Martin, 2009), and Generative Phonotactic Learner (Bailey & Hahn, 2001; Futrell *et al*., 2017). See Daland *et al*. (2011) for an overview. Work on wordlikeness judgment, however, has so far focused primarily on segments and studies incorporating suprasegmental features is relatively limited. Some suprasegmental features, including stress and tone, are used to create lexical contrasts cross-linguistically. Therefore, in order to understand the system of wordlikeness, the determinants of wordlikeness judgments should include suprasegmental features. Especially for the

languages where lexical contrasts are created with suprasegmental features, understanding the system of wordlikeness incorporating suprasegmental features is particularly crucial.

Some previous work examined how prosodic features related to stress should be incorporated into phonotactic models (Bird & Ellison, 1994; Coleman & Pierrehumbert, 1997; Hayes & Wilson, 2008; Olejarczuk & Kapatsinski, 2018). As for tone, which is the focus of this paper, there is previous work on wordlikeness judgments of tone languages which considers only segmental phonotactics, omitting tone. Gong (2017) compared speakers' acceptability and reaction times on lexical decisions involving systematic gaps and accidental gaps in Mandarin. They considered the role of phonotactic probability and neighborhood density to predict the results. It was found that the two had independent influence on acceptability, but neighborhood density was the only significant factor for reaction times. Gong also used Hayes & Wilson's (2008) phonotactic probability calculator, but without considering tone. As an extension of Gong (2017), Gong & Zhang (2020) did consider tonal neighbors, i.e. syllables that differ only by tone. In their investigation of lexical neighbors, however, cases where both a segment and a tone differ were not included, thus it was not possible to know the relative contribution of segments and tones in determining lexical neighbors in Mandarin. Myers (2015) is another work on a tone language, Mandarin, that considered only segmental phonotactics without tone. Myers focused on comparing the effect of lexical typicality and typological frequency on acceptability judgments. Lexical typicality was defined as to how many lexical syllables in Mandarin share the item's onset consonant, and typological frequency was defined in terms of the number of phoneme inventories that exhibit this consonant across languages. They looked at onset frequency within Mandarin and consonant frequency within The UCLA Phonological Segment Inventory Database (UPSID: Maddieson & Precoda, 1989) to find that both typological frequency and Mandarin-specific lexical typicality had effects which items speakers judge more wordlike.

Modeling work incorporating tone includes Myers & Tsay (2005), Kirby & Yu (2007), and Shoemark (2013). These studies differ with each other as to what they aimed to predict (e.g., judgements on real words, systematic gaps, accidental gaps, etc,), but a main goal was to identify the role of phonotactic probability and neighborhood density in predicting native speakers' wordlikeness judgments. Myers & Tsay (2005) examined the role of neighborhood density in predicting the typicality judgments in Mandarin and reported that the judgments of real Mandarin words can be predicted by neighborhood density but nonwords are inversely correlated to neighborhood density. In Kirby & Yu (2007), their focus was to find out the role of phonotactic probability and neighborhood density in understanding systematic and accidental gaps in Cantonese. The results showed the role of neighborhood density to predict wordlikeness judgments. Phonotactic probability also played a role, although there was a less correlation between phonotactic probability and wordlikeness. They suggest that this may be because Cantonese does not permit complex onsets and codas and thus has a much smaller number of possible monosyllables, leading to a lower importance of phonotactic probability. Also because the possible monosyllables are limited due to strict phonotactic regulations, lexical items occupy a much larger portion of the space of possible monosyllables, resulting in a greater role of lexical density. This idea was further pursued by Shoemark (2013) where it was argued that strict phonotactic restrictions in Cantonese create denser phonological network, from which the role of neighborhood density becomes crucial.

While the findings are mixed, the overall results seem to suggest that neighborhood density has a greater effect on tone language speakers' wordlikeness judgments than phonotactic probability. However, the space of possible ways to incorporate tone into the modeling of wordlikeness judgments has yet to be fully explored. In order to incorporate tone in modeling wordlikeness judgments, we need to address the following two issues; first, how

the major determinants of wordlikeness judgments, such as phonotactic probability and neighborhood density, should be operationalized with tone, and second, how to evaluate the contribution of these factors to wordlikeness judgment test results. For the first, we provide a survey of a variety of methods, and for the second, we provide a Bayesian hierarchical modeling. Both methods and modeling results will be presented with Cantonese as an example of a tone language. Section 2 first introduces the basics of Cantonese phonotactics and overviews multiple methods of measuring the two determinants of phonotactic knowledge, i.e. phonotactic probability and neighborhood density. It shows that both determinants have been primarily limited to measuring segments. Section 3 shows how to measure phonotactic probability and neighborhood density when tone is involved. Our methodology shows that 'classic' phonotactic probability calculation methods, originally proposed for segments such as *n*-gram models, (see Section 2.1), can be applied to tone languages, but tonal probabilities need to be incorporated into the calculation by identifying the tonal representation from which we can predict speakers' wordlikeness judgments. We also show how neighborhood density models, such as Generalized Context Model (Nosofsky, 1986) and Generalized Neighborhood Model (Bailey & Hahn, 2001), can be constructed with tone: Neighborhood density models should be informed by correct measurements of phonological distances between words which should incorporate measurements of segmental distances as well as tonal distances and their relative weights. To identify the role of phonotactic probability and neighborhood density in predicting speakers' wordlikeness judgments in Cantonese, we run a wordlikeness judgment test, presented in Section 4. Our results show that phonotactic probability, but not neighborhood density, is a significant factor in predicting speakers' wordlikeness judgments in Cantonese. When the role of each syllabic component is considered, probabilities of nucleus and coda are shown to contribute to the wordlikeness judgments most. We also show that phonotactic probability can predict the gradient items that fall between the two extreme judgments (i.e. between very wordlike and not at all wordlike) and categorically perfect items (very wordlike) but not for categorically bad items (not at all wordlike). Section 5 discuss the implications of the current findings to the study of phonotactic modeling with tone and the processes involved in wordlikeness judgments when tone is included.

2. Background

2.1. Cantonese phonotactics

Cantonese belongs to the Sinitic branch of Sino-Tibetan/Trans-Himalayan language family. Phonemically, the language has 19 consonants (Table 1), and 8 monophthongs with 11 diphthongs (Table 2). In this study, we assume that Cantonese has six tones as in Table 3 (Bauer, 1985; Matthews & Yip, 2011). There is another possible analysis, following the historical phonology tradition, in which Cantonese is analyzed to have nine tones, with high level, mid level, and low level tones assumed here being treated as the checked tones in syllables ending with an oral stop coda. The oral and nasal stops are then allophones in coda position. When the nine-tone system is assumed, one needs to suppose that codas change into oral stops when the tone is short (K. H. Cheung, 1986).

Consonants								
		Bilabial	Labio-dental	Alveolar	Palatal	Velar		Glottal
						Plain	Labialized	
Plosive	**Plain**	p		t		k	k^w	
	Aspirated	p^h		t^h		k^h	k^{wh}	
Nasal		m		n		ŋ		
Fricative			f	s				h
Affricate	**Plain**			ts				
	Aspirated			ts^h				
Approximant					j	w		
Lateral approximant				l				

Table 1. The inventory of Cantonese consonants.

Vowels										
Monophthongs										
	Front				Central				Back	
	Unrounded			Rounded			Short		Long	
Close	iː			yː						uː
Mid	ɛː			œː						ɔː
Open							ɐ		aː	
Diphthongs										
		Main vowel								
		aː	**ɐ**	**ɛː**	**e**	**œː**	**ɵ**	**ɔː**	**o**	**iː**	**uː**
Terminal	**i**	aːi	ɐi		ei						
	y						ɵy	ɔːy			uːy
	u	aːu	ɐu	ɛːu					ou	iːu	

Table 2. The inventory of Cantonese vowels.

Lexical tones						
Tone name	**High level**	**High rising**	**Mid level**	**Low falling**	**Low rising**	**Low level**
Tone letter	˥	˧˥	˧	˩˧	˨˩	˩
Jyutping tone number	1	2	3	4	5	6

Table 3. Lexical tones in Cantonese.

Cantonese has a maximal syllable structure of (C)V(C) or (C)V(V)[1] with strict restrictions on what segments or tones are allowed in certain syllabic positions (S. L. Cheung, 1991; Kirby & Yu, 2007; Yip, 1989). Apart from vowels, the syllabic nasal may take the nucleus position as there are a few syllables that consist of a nasal consonant alone ([ŋ] or [m]). All consonants are allowed in onsets, and in the current analysis, the secondary articulation /w/ is treated as a part of the onset, rather than as part of the nucleus. Compared to onset, only unreleased stops, nasals, and high vowels are allowed in codas, suggesting stricter phonotactic restrictions on coda than onset positions. Additional phonotactic restrictions are found from among the relations between two syllabic positions. The onset and coda of a syllable cannot be both labial (**pap, *mim*) (S. L. Cheung, 1991; Kirby & Yu, 2007; Yip,

[1] Note that consonant codas never co-occur with diphthongs, from which it was proposed that the second vowel in a diphthong should be considered as the coda of the syllable (Bauer & Benedict, 1997, pp. 13–14). Our study follows this idea.

1989). Rounded vowels cannot be followed by labial codas (**-uːm, *- ɔːp*) and front rounded vowels cannot be preceded by labial onsets *(*my-, *pœː-)*. The onset and coda of a syllable with a back vowel as the nucleus cannot be both coronal *(*nɔːn, *tuːt)* and coronal onsets cannot be followed by nucleus /uː/ or /uːy/ *(*tuː, *nuːy)*. Syllables ending in unreleased stops can only take the three level tones (tones ˥, ˧, and ˨). Syllables with unaspirated stops or affricates in onset do not bear tones ˩ and ˧, while syllables with aspirated stops or affricates in onset do not bear tone ˨ (Kirby & Yu, 2007). Exceptions to the aforementioned phonotactic regulations exist in loanwords and ideophones (Bauer, 1985). The results of a wordlikeness test showed that Cantonese native speakers' phonotactic knowledge reflects systematic and accidental gaps found in the lexicon (Kirby & Yu, 2007).

A morphological aspect that is relevant to the current study is the status of monosyllabicity in Cantonese. Cantonese favors disyllabicity, and many modern Cantonese monosyllabic morphemes are generally not used as independent words, but only appear in compounds (Bauer & Benedict, 1997). Some sociolinguistic sound changes in Cantonese are relevant to the current study as well: Cantonese shows ongoing sound changes including the initial [n-] and [l-] merger, the coda [-t] and [-k] merger, and [-n] and [-ng] merger (Bauer & Benedict, 1997), and the merges between tones ˥ and ˧, ˧ and ˨, and ˩ and ˨ (Mok *et al*., 2013).

2.2.Determinants of wordlikeness judgment

Our main question is on what basis native speakers make wordlikeness judgments in tone languages. For example, how do Cantonese native speakers tell that a novel sound sequence with coda /f/ is less wordlike than the one with coda /m/? As mentioned in Section 1, previous work suggests that there are mainly two sound-related determinants of wordlikeness, namely phonotactic probability and neighborhood density (see review in Bailey & Hahn, 2001). Phonotactic probability and neighborhood density models are often correlated, but they quantify different aspects of wordlikeness. Phonotactic probability decomposes strings of sounds into substrings and aggregate the probabilities of those substrings to create measures of wordlikeness (Albright, 2009). It is an analytical approach in that it decomposes words into pieces and calculates probabilities. Neighborhood density models count the number of words that are similar in a lexicon, by certain metrics which we will discuss in Section 2.2, to the sound sequence in question, sometimes weighted by some criteria like frequency (e.g., GNM of Bailey & Hahn, 2001). It is a holistic approach in that the calculation is based on the whole lexicon. In the following section, we first introduce the two determinants in detail and Section 2.3 considers how the two determinants can be measured in tone languages, such as Cantonese.

2.2.1. Phonotactic probability

Numerous ways to compute phonotactic probability have been proposed. These methodological decisions can fall into 'researcher degrees of freedom' (Roettger, 2019; Simmons *et al*., 2011) that can critically affect the results. Although the diverse methodological approaches all have a similar goal of generating good predictors of wordlikeness judgements (or performance in some other experimental task, such as spoken word recognition or non-word repetition) and are often quite strongly correlated, there is considerable variation in the underlying philosophy. Here we identify three main aspects in which the implementations of phonotactic probability may vary: (a) types of probabilities, (b) methods of estimating probabilities, and (c) methods of aggregating estimated probabilities.

Type of probabilities. Phonotactic probability is generally calculated over *n*-phones (*n*-grams) of segments, where *n* is the length of the substring of segments considered. A unigram is a single segment, a bigram/biphone consists of two contiguous segments, etc. Usually, the largest substring considered in phonotactics studies is the triphone. For Cantonese, when segment sequences are considered, the application of unigram to trigram calculation is straightforward as its maximal syllable structure is CVC or CVV, and only one phoneme is allowed in each syllabic position. There are also models that, instead of considering the probabilities of *n*-phones directly, consider the probabilities of *n*-grams of 'natural classes' as well as the probabilities of individual phonemes given the natural class (Albright, 2009; Albright & Hayes, 2003). Hybrid models of these also exist, generally based on syllable structure. The 'syllable part' approach (Bailey & Hahn, 2001) computes probabilities over the onset, nucleus and coda of a syllable, which may vary in length in some languages like English. Similarly, their 'syllable rime' approach computes probabilities over onsets and rimes, calculating the probabilities of onsets and rimes as single units, and treating them as independent. In Chinese, it has sometimes been argued that there is no need to decompose the rime into nucleus and coda, and instead the rime is treated as a single unit, with each rime being a *rimeme* (Chao 1934, Light 1977). The distinction of rime and rimeme is important for the analysis of different syllables in different Chinese dialects, but for the purpose of our paper, they can be understood comparable. If a rimeme or a rime is assumed to be a single unit, the syllable part approach cannot be pursued, and instead the syllable rime approach must be used. For Cantonese phoneme sequences, no complication is involved in applying the syllable structure approaches, whether it is a syllable part or a syllable rime approach, due to its phonotactic restriction to allow only one phoneme in each syllabic position. A main issue in measuring phonotactic probability in Cantonese is to determine the position of tone in relation to onset, nucleus, and coda. We further discuss this issue in Section 2.3.

Once we determine the representation of the syllable structure, the next step is to compute probabilities within a syllable. There are two types of probabilities computed. First, positional probability computes the probabilities that a segment or *n*-phone appears at a certain position in a word, e.g. the probabilities that the phone [a] is the second segment in a word, or that the biphone [pl] is the second and third segments in a word. Transitional probability computes the probabilities that a segment appears, given the $n - 1$ previous segments, where $n = 2$ for biphones, $n = 3$ for triphones, etc. Word boundaries, denoted #, are often considered 'segments' in these approaches, so that the probability distribution of the actual first sound in a phoneme sequence is the conditional distribution of it given the first 'segment', namely the word-initial boundary. Due to a restriction in Cantonese that allows a single phoneme in each syllabic position, measuring positional and transitional probabilities of Cantonese segmental sequences is simple. Like before, the issue is to determine the position of tone in relation to phonemes.

Method of estimating probabilities. The probabilities themselves are concepts (or 'parameters' in statistics) that are unknown and thus must be estimated using a corpus. Note that the estimation is independent of phonotactics of individual languages and estimating probabilities with or without tone is not an issue here. Different researchers differ with respect to the estimation methods used. One popular method, especially among psycholinguists, is to use log frequencies in the computation of phonotactic probability (Jusczyk *et al*., 1994; Vitevitch & Luce, 2004). To calculate positional probability of an *n*-phone, for instance, the log frequency of that *n*-phone in a certain position is divided by the log of the total number of words that contain in the position; to calculate the transitional probability of an *n*-phone, the log frequency of the *n*-phone is divided by the log of the total number of words where the first $n - 1$ segments of the *n*-phone appears. The underlying

assumption is that log frequencies are better measures of 'perceived' frequencies than raw frequencies. A second approach is to use maximum likelihood estimation in calculating the probabilities (e.g. Albright, 2007, 2009). Here, raw counts are used instead of log frequencies in the numerator and denominator; otherwise, the calculations are identical as the log frequency approach. Some probabilities are likely to be zero due to accidental or systematic gaps. A third approach is intended to better deal with such zero probabilities. It modifies the maximum likelihood estimation by adding a smoothing parameter to avoid overfitting (e.g. Dautriche *et al*., 2017; see also Jurafsky & Martin, 2009, for a more detailed description of the method as applied to word *n*-grams). In methods that use log-frequencies, zero counts are particularly problematic, as they would result in undefined log-frequencies and hence undefined probability estimations. Some methods using log frequencies can deal with issues of zero counts, though in somewhat 'ad hoc' ways: For example, Vitevich & Luce's (2004) phonotactic probability replaces the undefined probabilities for unattested *n*-phones, which have log 0 in the numerator, with 0 probabilities.

Apart from the methods of estimating probabilities, there is also a question of whether the frequencies used should be based on type frequencies or token frequencies (Daland *et al*., 2011; Denby *et al*., 2018; Richtsmeier, 2011). The former is counted with an entire lexicon, whereas the latter can be computed using a frequency wordlist or a corpus.

Method of aggregating estimated probabilities. Once we estimate individual probabilities, we need to combine them together. As with estimation, the methods of combining the probabilities are independent of the involvement of tone. There are two main ways of combining the estimated probabilities computed into single measures of phonotactic probability: taking the sums (i.e. adding probabilities) or taking the products (i.e. multiplying probabilities). Note though that simply adding or multiplying probabilities may produce a measure that is not a true probability, but they are still frequently used. For both methods of taking the sums (i.e. adding probabilities) and taking the products (i.e. multiplying probabilities), many variations exist. First, the probabilities may be logged before combining them; they may be combined before being logged; or they may not be logged at all. (Note that the sum of the log of the probabilities is the same as the logs of the products.) Second, the probabilities may be normalized to account for word length; the arithmetic mean of the probabilities may be taken if we are summing the probabilities, and the geometric mean if we are multiplying them. Many of these methods were explored in Bailey & Hahn (2001) and Vitevitch and Luce (2004).

2.2.2. Neighborhood density

The other major predictor of wordlikeness judgments is neighborhood density, the degree to which an item under consideration resembles other items in the lexicon. Like phonotactic probability, neighborhood density can be measured in many different ways. The simplest and most common measure is the number of lexical neighbors, where a word is a neighbor of another word if one word can be obtained from another by adding, deleting or changing one segment. For example, /kæt/ is a neighbor of /kæts/, /æt/ and /bæt/. In this approach, a neighbor is a categorical concept: Two words are either neighbors or not. Due to the assumption that a lexical neighbor is a categorical concept, counting the number of lexical neighbors including tone does not differ from the one only with phonemes. In Cantonese, for instance, /kaː˨/ is a neighbor of /kaː˧/, /kʰaː˨/, /kanː˨/ and /saː˨/. This is the method used in Kirby & Yu (2007) when measuring neighborhood density in Cantonese.

A more sophisticated measure of lexical neighborhood allows for gradience. For example, we would expect that /kæt/ is a closer neighbor to /kæts/ than to /bræts/, but also /kæt/ is

closer to /bræts/ than to /brɪts/. To arrive at such 'gradient' neighborhood models, we need to construct phonological distance measures for the exact distances between words. The literature on such measures is large, primarily for the ones for segments.[2] The most common method is to determine the distances between two corresponding phonemes first, then combine them to find distances between phoneme strings. When measuring distance between two phonemes, Bailey & Hahn (2001) used the natural class distance in Frisch *et al*. (1997). In the distance measurement in (1), the number of non-shared natural classes between two phonemes is divided by the total number of natural classes, i.e. shared natural classes + non-shared natural classes, across the two phonemes. In other words, the distance between two phonemes is defined by the proportion of non-shared natural classes that the phonemes belong to.

(1) $Distance_{NC} = \frac{Non-shared\ natural\ classes}{Total\ number\ of\ natural\ classes}$

The Levenshtein distance (Jurafsky & Martin, 2019) between the two phoneme strings is then computed: An algorithm is used to find the way of adding a segment, deleting a segment or substituting one segment for another that minimizes the 'cost' of these operations, cost being the distance between the two corresponding phonemes involved. The distance between two phoneme strings then becomes the average cost of the operation. When tone is involved in the distance calculation, the distance between two tones should be included in the calculation. We elaborate this point in Section 2.3.

Once a distance measure between two words, $d(w_i, w_j)$, has been constructed, incorporating phoneme distances and tone distances, it is used to measure the lexical density of words. One possibility is the Generalized Context Model (GCM) of Nosofsky (1988). GCM is an exemplar model, where categorization of a lexical item is based on its similarity towards all relevant stored exemplars, i.e. lexical neighbors. In GCM, the neighborhood density of a word is calculated by summing up the exponent of the negative distance of every word in the lexicon from the word itself. In (2), L denotes the lexicon, i.e. the set of all words in the language. Because of the negation sign, words that are far away from the word under consideration are weighted less whereas words that are close to the word are weighted heavily. In order to measure how far a word is to other words including tone, we need to identify the relative contribution of segmental distance and tonal distance in determining the distance between two words. Section 3.2 shows methods of identifying the relative weightings of segmental and tonal distances.

(2) $GCM(w_i) = \sum_{w_j \in L} \quad e^{-d(w_i, w_j)}$

Although GCM does consider gradient similarity of all relevant words in the lexicon, one disadvantage of GCM is that lexical frequencies are ignored. To address this issue, Bailey & Hahn (2001) propose the Generalized Neighborhood Model (GNM), where the contribution of a word depends not only on its distance to the word under consideration but also its frequency of occurrence. So, under GNM, frequency of a word' occurrence affects its 'weights'. In (3), log frequency of occurrence is denoted by f_j, and A, B, C, and D are free parameters. These parameters give the relative contribution of the quantity weighted by the square of the log frequency (A), the quantity weighted by the raw frequency itself (B), the non-frequency weighted (i.e. GCM) quantity (C), along with a 'sensitivity parameter' D that is multiplied to each distance, as in (3). GNM modeling with tone is similar to the GCM

[2] Readers are pointed to Kessler (2005) for an overview.

modeling with tone, except that the relative weight of tones and segments, which incorporates frequency information, should be included in the distance measure. Due to the additional parameters, it is mathematically more complicated than GCM, but the core components needed for GNM modeling with tone are similar to those for GCM modeling: We need to identify both segmental and tonal distances and their relative contributions in determining the distances between words.

(3) $GNM(w_i) = \sum_{w_j \in L} \quad \left(Af_j^2 + Bf_j + C\right)e^{-D \cdot d(w_i, w_j)}$

As an example of the GCM and the GNM applications to phoneme strings without tone, consider a miniature language below. The language has five words taken from English, *strata* [streɪtə], *spray* [spreɪ], *star* [stɑr], *tar* [tɑr] and *states* [steɪts], and each word appears in the corpus 8, 15, 16, 5 and 20 times respectively. We consider the problem of determining the neighborhood density of *star* [stɑr]. The distance between *star* [stɑr] and the other four words are shown in Figure 1 on the lines joining them with *star* /stɑr/. Under GCM, the neighborhood density of *star* [stɑr] is $e^{-1} + e^{-2} + e^{-3} + e^{-4} = 0.571$. Under the GNM with $A = 1, B = -2, C = 3, D = 4$, the neighborhood density of *star* [stɑr] is $(1 \times 5^2 - 2 \times 5 + 3)e^{-4\times1} + (1 \times 8^2 - 2 \times 8 + 3)e^{-4\times2} + (1 \times 20^2 - 2 \times 20 + 3)e^{-4\times3} + (1 \times 15^2 - 2 \times 15 + 3)e^{-4\times4} = 0.349$. To build GCM and GNM models with tone here, the distances of a word to the other words in Figure 1 (1, 2, 3, and 4) should be measured with tonal distances as well, and log frequency of occurrence, f_j, and the four parameters in (3), *A*, *B* and *C,* and *D,* should be informed by lexicon including tonal information.

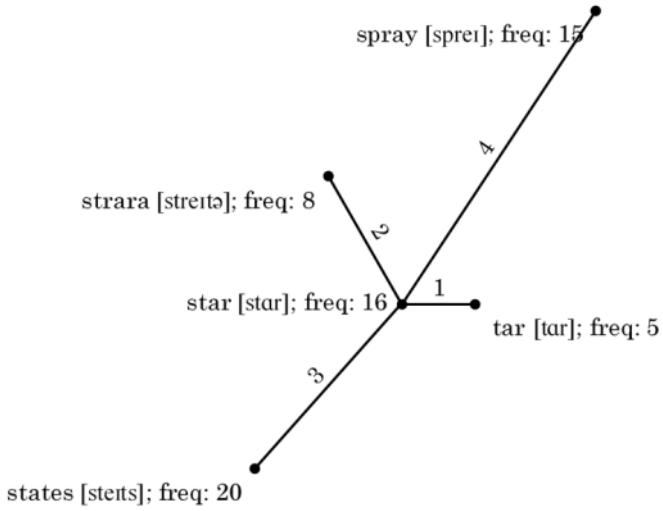


Figure 1. The distance between *star* [stɑr] and the other four words with their frequency.

As we will discuss below, our experimental design is constructed with reference to the 'number of neighbors' measure, as in Kirby & Yu (2007), as well as the GCM and GNM models. In Section 3, we show how these methods can be adapted to incorporate tone in calculations, with Cantonese as a case study.

3. Phonotactic modeling with tone

3.1. Phonotactic probability

As introduced in Section 2, there are multiple methods for computing phonotactic probability, involving a large number of decisions concerning the type of probabilities, the estimation methods, and the methods of aggregating the probabilities. For the computation of phonotactic probability, a main guiding principle is to create a theoretically well-grounded measure of the joint probability of the entire syllable including tone.

In our study with Cantonese, we use traditional bigram probabilities due to its best performance in Kirby & Yu (2007). Alternatively, unigram or trigram probabilities can be applied without much difference from bigram probabilities, because only one segment or tone is allowed in each syllabic position in Cantonese. Bigram probabilities are calculated on the basis of token frequencies from Hong Kong Cantonese corpus (Luke & Wong, 2015), whose performance was better than when the calculation was based on type frequency in Kirby & Yu[3]. We adopt the 'syllable parts' approach described by Bailey & Hahn (2001), which computes probabilities over the onset, nucleus and coda of a syllable. We do not adopt an approach which assumes a rime as a single unit, such as a syllable rime approach or a rimeme approach (Chao 1934, Light 1977). See our justification for this decision in Section 4. As Cantonese syllable structure only allows one phoneme in each of syllable part slots (assuming that the second vowels in diphthongs are considered to be a coda), we compute P(onset)[4]—conceptually equivalent to P(onset|#) for models that consider word boundaries—P(nucleus|onset), and P(coda|nucleus), then multiply the three together as P(segments). In other words, P(segments) is calculated by multiplying probabilities of the syllabic components' *n*-grams. We assume that the second vowels in diphthongs are codas for the following reason. In Cantonese, there are strict phonotactic restrictions on diphthong-coda sequences. For example, falling-sonority diphthongs like *ei* never co-occur with nasal or oral stop coda. So it has been proposed that the second component of the diphthongs be considered part of the coda (Bauer & Benedict, 1997). When computing these probabilities, additive smoothing is performed to prevent zero probabilities, since this would result in an undefined log-probability (See section 2.2). For smoothing, we simply add 1 to all counts for simplicity (i.e. add-one smoothing), and do not pursue more complicated methods.

To calculate probabilities of tone given the string of segments, we compute P(tone|segments) using a multinomial logistic regression model with the nnet package (Ripley & Venables, 2016) in R (R Core Team, 2020). An assumption here is that the probability of a syllable having a certain tone is dependent on the identities of all segments in the syllable. Dummy variables representing onset, nucleus and coda are included in the model; we excluded interaction effects to ensure that the probabilities of tone can be calculated for an unattested segment string as well. The probability of a monosyllable is then the joint probability of segments with tone given the segments, P(segments) P(tone|segments). Then we take the natural logarithm of this joint probability as a linear predictor of wordlikeness in our model. As an alternative, we also consider P(tone|coda)

[3] We believe that a better performance based on token probabilities is largely due to language-specific properties of lexicon. In Cantonese, words are predominantly disyllabic (73%) and the length of almost all of the words lies between 1 to 3 syllables (97%) (Lai & Winterstein, 2020). Due to this, there are many homophones and many characters have multiple pronunciations, which can be indicative of a more crucial role of token rather than type frequencies in judgment tests.

[4] Throughout this paper, we will use notation such as P(onset) to denote the probability of a particular onset (e.g. /p/ or /l/), not the probability that any onset will appear.

rather than P(tone|segments), because, as introduced in Section 2.1, there is a strong co-occurrence restriction in most Cantonese words whereby oral stop finals [p, t, k] may only co-occur with tones ˥, ˧, and ˨ (Bauer & Benedict, 1997). Additionally, we also consider P(tone|onset), because Cantonese exhibits some restrictions on the relation between onset and tone, i.e. syllables with unaspirated stops or affricates in onset do not bear tones ˩ and ˧˥ while syllables with aspirated stops or affricates in onset do not bear tone ˨ (Bauer & Benedict, 1997). Three other conceptually possible probabilities are tested as well, including tone conditioned on nucleus, tone not conditioned on segments, and the one only with segments without tone. Our case of Cantonese thus resulted in six log-probability measurements shown in (4) with different assumptions with regards to the relationship between tone and segments.

(4) Six types of measurements of tonal probabilities

Types	Abbreviations	Definitions
P(tone|segments)	P(T|S)	Tone conditioned on all segments
P(tone|coda)	P(T|C)	Tone conditioned on coda only
P(tone|onset)	P(T|O)	Tone conditioned on onset only
P(tone|nucleus)	P(T|N)	Tone conditioned on nucleus only
P(tone)	P(T)	Tone unconditioned on segments
No tonal probability	NoT	Segmental probabilities only without tone

3.2.Neighborhood density

As mentioned in Section 2.2, it is straightforward to count the number of neighbors in tone languages if a tonal neighbor is assumed to be a categorical concept. For examples, *ka5* [kʰɐ˧˥] is a neighbor of *ka6* [kʰɐ˨] (tone substitution), just as it is to *ki5* [kʰiː˧˥] (segment substitution), *a5* [ɐ˧˥] (segment deletion), or *kat5* [kʰɐt̄˧˥] (segment addition). On the other hands, modeling neighborhood density in tone languages using GCM or GNM models is more complicated, mainly because (a) tonal distance should be measured and incorporated into modeling and (b) relative contributions of segmental and tonal distances should be identified in determining the distance between words.

First, we demonstrate how to construct distance metrics between two words, $d(w_i, w_j)$, in the notation introduced in Section 2.2, including tonal distance. Instead of directly following Bailey & Hahn (2001) in calculating segmental Levenshtein distances (Jurafsky & Martin, 2019), we adopt a way that was proposed to measure phonological distance in Cantonese taking both segments and tones into account. A study from Do & Lai (forthcoming) reported how to measure phonological distances of words when tone is involved using Cantonese as an example. In their study, the distances of segments and tones were first calculated separately assuming various phonological representations of segments and tones. The assumed representations included binary and multivalued feature representations for segments and the Chao tone letters, autosegmental, and (onset)-contour-(offset) representations of tones (see Do & Lai for the justifications for each representation). They collected phonological distance judgement data between two items by asking how similar the two items are, such as between *se4* [sɛː˩] ‘snake’ and *te6* [tʰɛː˨], within the scale of 0 (totally different) and 100 (identical). Various models were compared with the participants’ data to find out the optimal way to measure segmental and tonal distances. For segmental distance, Do & Lai found that a distance measure represented by a multivalued, mostly articulatory-based featural representation based on the one in Ladefoged (1975) worked best. There are several ways to calculate segmental distances between such segmental representations, but the one that worked optimally in their study was with Hamming distance (Nerbonne &

Heeringa, 1997). Hamming distance measures the number of features that are not shared between two phonemes and divides them into the total number of phonological features, i.e. shared and non-shared phonological features between the two. Given its optional performance in the previous study, the current study adopts Hamming distance measures for multivalued representations to measure segmental distances in Cantonese. For the tonal distance, Do & Lai found that Hamming distance measure with representing tone in terms of contour and offset was optimal in predicting native speakers' phonological distance judgements. This result echos the results of perception studies in Cantonese, where tonal contours are found to be an important perceptual cue (e.g. Khouw & Ciocca, 2007; Xu *et al*., 2006), which in fact is more important than tonal heights (Gandour, 1981). Reflecting its good performance, the current study also adopts the same distance measure and the tonal representation. The six tones of Cantonese are represented in Table 4 following the contour-offset representation. The distance between the tones was 1 if both contour and offset were different (e.g., tone ˥ vs. tone ˩), 0.5 if either contour or offset was different (e.g., tone ˥ vs. tone ˧), and 0 if both were same (i.e. same tone).

Tone letter	˥	˧	˦	˩	˨	˩
Tone number	1	2	3	4	5	6
Contour	Lv	R	Lv	F	R	Lv
Offset	H	H	M	L*	M	L

(Lv: level; R: rising; F: falling; H: high; M: mid; L: low; L*: Extra low)
Table 4. Six tones in Cantonese in the contour-offset representation.

Second, once we identify optimal distance measures for segments and tones, the issue now is to combine the two measures together. One straightforward way to do this is to simply add them together. However, this does not allow for different weightings for segments and tones, which are experimentally evidenced from perception (Cham, 2003), word recognition (Cutler & Chen, 1997; Keung & Hoosain, 1979), word reconstruction (Wiener & Turnbull, 2016), and phonological distance studies (Do & Lai, forthcoming; Yang & Castro, 2008). To model empirically informed weights of segments and tones, we decided to choose the weights that can predict native speakers' phonological distance judgements in Do & Lai using the fixed intercept and coefficients for segmental and tonal distance.

The computation of a GCM model is straightforward so far as we identify segmental and tonal distances and their relative weights. However, there are additional complications with GNM. It is mainly because GNM modeling incorporates 'frequency', which is ignored in GCM modeling. As mentioned in Section 2.2, there are four free parameters in Bailey & Hahn's (2001) model, A (the quantity weighted by the square of the log frequency), B (the quantity weighted by the raw frequency itself), C (a free parameter that gives the relative contribution of the non-frequency weighted quantity), and D (a sensitivity parameter that is multiplied to each distance). Bailey & Hahn mentioned that they computed these coefficients in GNM by regression. However, as they did not specify the details of the implementation, we devised our own method of estimating the parameters. In our modeling, to simplify calculations, we fixed a sensitivity parameter D at 1 but inferred A, B and C empirically from the results of our wordlikeness test. This greatly simplifies the process of finding the values of A, B and C, as the GNM without a sensitivity parameter will become a linear combination of three quantities—the sum of the exponent of the negative distance from each word to the word under question, weighted by the square of the frequency, weighted frequency and unweighted (i.e. GCM), respectively—with A, B and C as coefficients. Frequency weighting was based on token frequencies from Hong Kong Cantonese corpus (Luke & Wong, 2015).

Now, with frequency information and segmental and tonal distances as well as their relative weights, what is needed is wordlikeness judgment data from native speakers. Section 4 collects wordlikeness judgment data, from which we build the GCN and GNM models in Section 5.

4. Wordlikeness judgment test[5]

Previous sections introduced ways to measure phonotactic probability and neighborhood density. We also showed our measurement decisions for Cantonese, incorporating tone. With the two predictors measured with tone, we test their roles in predicting Cantonese native speakers' wordlikeness judgments. In our experiment, participants were asked to judge how wordlike given words are within the range of 0 (not at all wordlike) to 100 (very wordlike).

4.1. Test

Participants. The experiment was built by using an online survey software Qualtrics (Qualtrics, 2020) and was distributed through social media to the public. Self-reported native speakers of Hong Kong Cantonese participated in the experiment and they received 100 HKD (13 USD) compensation upon the completion of the experiment. In total, 145 participants were recruited. They were within the age range of 18 and 60. Among all the participants, 44 of them did not complete the experiment and 4 of the participants provided more than three incorrect answers out of 12 in the pretest (see procedure for the specifics of the pretest), and thus did not proceed to the main test. Data from the participants who did not complete the test (n=44) and who did not pass the pretest (n=4) were excluded from the analysis. Consequently, 97 participants' data were analyzed.

Design. In creating the experimental stimuli, we first calculated the phonotactic probability and neighborhood density following our measurement decisions presented in Section 3 for every logically possible combination of possible onsets, nuclei, codas and tones in Cantonese. The list was shortened to exclude the real words that are present in the Hong Kong Cantonese corpus (Luke & Wong, 2015). We then chose 288 items from the list. This includes all possible onsets, nuclei and codas, and it was made sure that every single possible phoneme at each syllabic position appears in the stimuli list. After creating the list, the second author examined the items to identify syllables that exist in Cantonese but were not present in the corpus by accident and modified one syllabic component to create non-existing syllables. For example, the string *gep6* [kɛːp̄ˌ] (which is present in Cantonese as the colloquial reading of 夾 'press from both sides') was replaced by non-existing syllable *get6* [kɛːt̄ˌ]. The stimuli list is provided in Supplementary Materials A.

A female Cantonese native speaker from Hong Kong recorded the stimuli. An examination of the speaker's natural speech revealed that she was not affected by ongoing sound changes in Cantonese including the initial [n-] and [l-] merger, the coda [-t] and [-k] and [-n] and [-ng] mergers, and the merges between tones ˥ and ˧, ˦ and ˨, and ˩ and ˨. The stimuli were recorded in a sound-attenuated booth in the first author's institute with Marantz PMD661MKII Handheld Solid State Recorder and Sennheiser MKE2-P-K Clip-On Lavalier Condenser Microphone. All stimuli were recorded as WAV format in mono with 16 bit-

[5] The data and the code for the experiment are available at https://osf.io/3j2se/?view_only=911bd65bb6f54db6ae1083e98937e543.

resolution at a sampling rate of 44.1kHz and were normalized by using the built-in normalize command in Praat (Boersma & Weenink, 2019).

Procedure. The experiment began with an introduction and an electronic consent form, followed by a demographic questionnaire related to participants' language background. See Supplementary Materials B for the questionnaire. Participants had to complete a pre-test before entering the main experiment session. The pre-test was in the form of AXB test to ensure that participants could perceptually distinguish between [n] and [l] initials, [t] and [k] finals, and [n] and [ng] finals, which have merged in some Cantonese speakers (Bauer & Benedict, 1997). The AXB test also included items to check whether they could distinguish between tones ˥ and ˧, ˦ and ˨, and ˩ and ˨, which are merging in some Cantonese speakers (Mok *et al*., 2013). If participants submitted more than three incorrect answers to the 12 questions, the experiment stopped.

In the main session, the experimental items were randomly presented to participants, one at a time. The main session lasted on average 40 minutes. Participants were asked to rate how likely each item would be a Cantonese word from 0 to 100 by using a slider. They were allowed to listen to it multiple times. Afterwards, the results were divided by 100 to lie between 0 and 1 for ease of interpretation.

4.2. Results

Data exploration. Before turning to modelling and statistical inference, we first provide a descriptive analysis of the data. To do so, graphs of the wordlikeness judgement data against the two assumed determinants, i.e. phonotactic probability and neighborhood density, are provided. The plots here show the percentage of categorically 'wordlike' judgements, percentage of categorically 'not at all wordlike' judgements, as well as the average gradient judgement against each of the predictors that we use. Scatterplots of the raw data are given in Supplementary Materials C.

First, in Figure 2, the *x*-axis denotes log-probabilities of test items and the *y*-axis is wordlikeness judgments converted to the range of 0 (not at all wordlike) to 1 (very wordlike). The size of the points on the graph indicate the number of points with the same *x* and *y* values. We chose the version of log-probabilities where tone is conditioned on all segments for illustration, but the graphs are very similar across different types of log-probabilities. For log-probabilities, there is a clear relationship with the wordlikeness judgements, as seen in Figure 2. Aside from an outlier on the far left, the higher the log-probability, the greater the chances that participants rate the wordlikeness as 1 or in the higher rate regions. In particular, ratings above 0.5 are quite sparse before the log-probabilities of −23, and they become much more common afterwards, especially after around −17. Ratings below 0.5 become quite rare for the three stimuli with the very highest log-probabilities. However, aside from the three items with the top log-probabilities, the rest of the items all have a similar number of 0 rating judgements. There are also stimuli, such as those around −23 log-probabilities, where there are frequent ratings of 1, but ratings in the higher regions are still sparse. All this suggests that there is great degree of variations among the participants' wordlikeness judgments, and the categorical judgments of 0s and 1s and the gradient judgements may not be produced by the same process; in particular, the judgements of 0s do not seem heavily affected by log-probabilities, whereas visually, the trend is clearer for gradient judgements and the judgments of 1s.

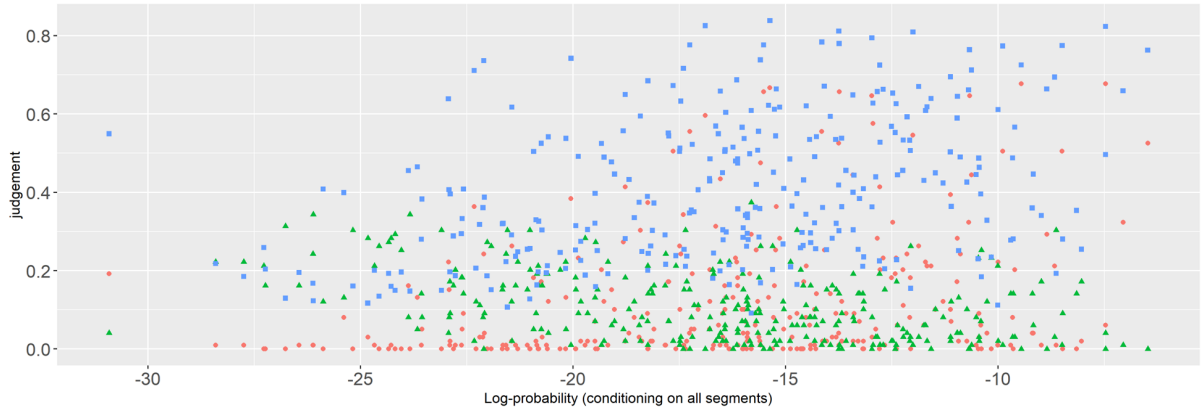


Figure 2. A plot of the proportion of 0 judgements (green triangles), proportion of 1 judgements (red circles) and average gradient judgement (blue squares) against the log-probability (x-axis).

Second, the descriptive results for neighborhood density measures are provided in Figures 3–6. Each figure shows the judgment data against the number of neighbors (NN, Figure 3), GCM (Figure 4), and GNM (Figures 4-6) respectively. As shown in Figure 3, the judgments of 0s (not at all wordlike) tend to be somewhat more common on the left side of the graph, lower NN regions. However, items with higher NN were not clearly judged better, indicating no predictive power of NN for the wordlikeness judgments.

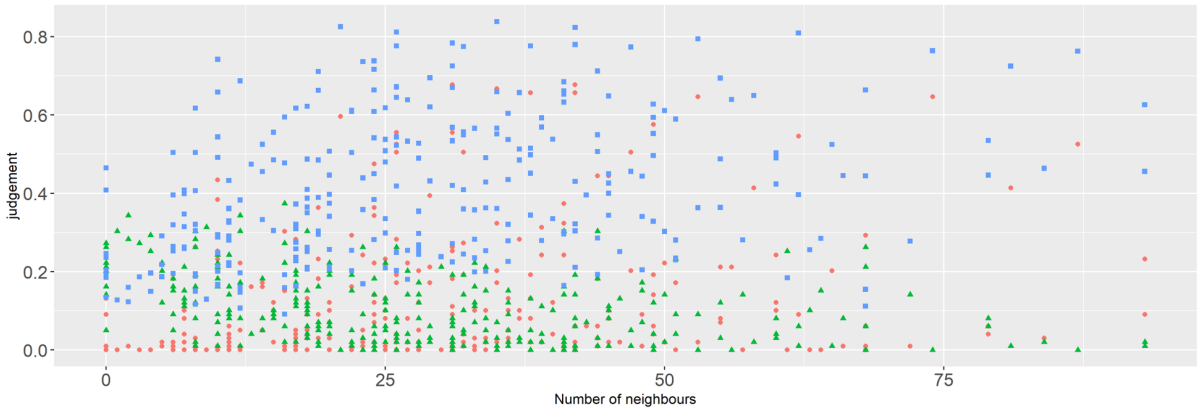


Figure 3. A plot of the proportion of 0 judgements (green triangles), proportion of 1 judgements (red circles) and average gradient judgement (blue squares) against the number of neighbors (x-axis).

When the neighbors' gradience is taken into account, there is no clear pattern for any of the three terms, the one with A (i.e. the quantity weighted by the square of the log frequency), with B (i.e. the quantity weighted by the raw frequency itself), and with C (i.e. the unweighted GCM quality). We start by examining the GCM value, i.e. with the coefficient C. As in Figure 4, the judgments of 1s and 0s, categorical judgments, tend to be concentrated in the middle of the GCM values. For intermediate judgements, there are items skewed towards high and low values across the entire *x*-axis. The addition of frequency weighting does not seem to create clear patterns, either. Figure 5 is for the raw frequency, i.e. with the coefficient B. For intermediate judgements, there is a tendency in which items in the far lower part of the graph tend to disfavor the judgments of 1s. But there is no clear tendency for the rest of the graph. Figure 6 shows the data for the square frequencies, i.e. the coefficient A. We see a

similar pattern, whereby items with value below around 50 disfavor 1s, but we do not see very clear tendencies for the judgments of 0s or for the other sections of the graph.

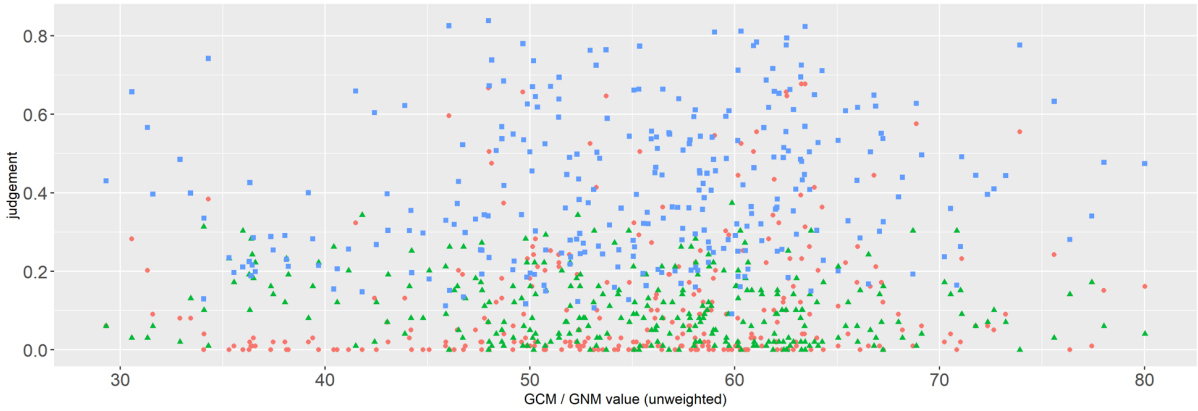


Figure 4. A plot of the proportion of 0 judgements (green triangles), proportion of 1 judgements (red circles) and average gradient judgement (blue squares) against the GCM values (x-axis).

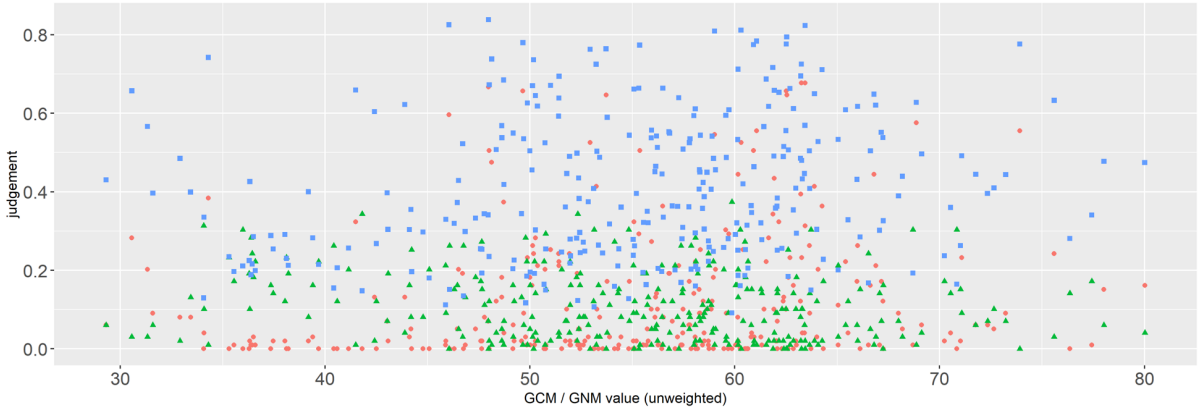


Figure 5. A plot of the proportion of 0 judgements (green triangles), proportion of 1 judgements (red circles) and average gradient judgement (blue squares) against the frequency-weighted GNM values (x-axis).

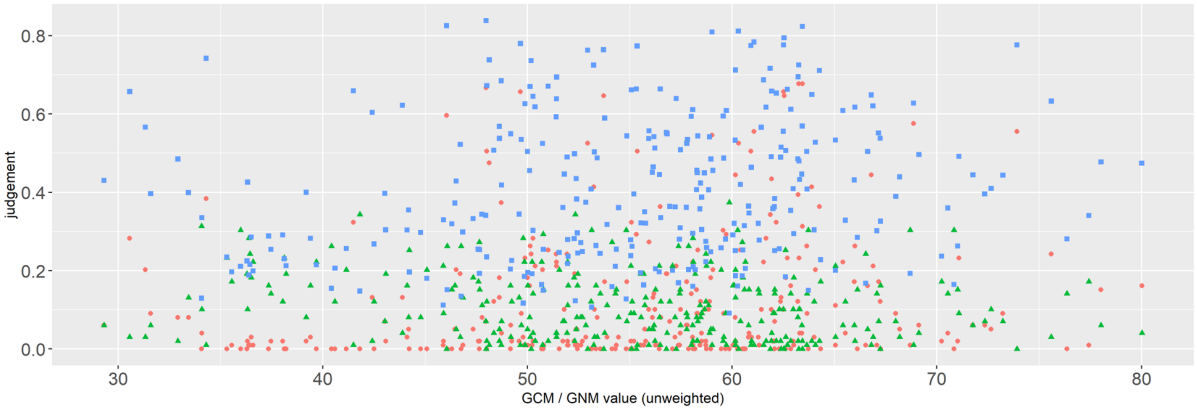


Figure 6. A plot of the proportion of 0 judgements (green triangles), proportion of 1 judgements (red circles) and average gradient judgement (blue squares) against the square frequency-weighted GNM values (x-axis).

Descriptive data seem to suggest that log-probability is relevant to wordlikeness judgments in Cantonese, but the effect of neighborhood density, if at all present, is weak: whether it be on categorical measure like NN nor gradient measures like GCM or GNM. Modeling results in Section 4.3 concur with these descriptive observations.

4.3. Modeling

Our modeling decisions were made based on the descriptive data in Section 4.2. As we noted above, there is a clear tendency for the categorical judgements to behave differently from gradient judgements. For example, log-probability seems to have little effect on the judgments of 0s while it seems to be correlated with gradient judgments. Its role to the judgments of 1s is less clear but items with low log-probability were rarely rated very wordlike. Based on the observations showing the distinctive patterns among the judgments of 0s, 1s, and gradient ones, we chose a model that allows us to separate judgements of 0s, 1s, and gradient ones. Specifically, we employ a mixed-effect Zero-One-Inflated Beta regression model (ZOIB: Ospina & Ferrari, 2012), which is similar to a beta regression model, but with extra components that allow the response to take on values of 0 or 1, modelled separately from judgement between 0 and 1. More 'familiar' models will not be appropriate for modeling our data. For example, linear regression models assume the residuals to be normally distributed, an assumption that is difficult to justify for the current case because of the multimodality prevalent throughout the data, which can be clearly seen in the scatterplots of the raw data in Supplementary Materials C. Beta regression models only cover the *open* interval (0, 1), and to use beta regression, we need to artificially turn the categorical judgements into values like 0.001 and 0.999, which is not ideal. Thus, ZOIB was our choice for the current data type. It is an 'inflated' regression model in that the distribution of the dependent variable is assumed to contain frequent 0s and 1s, which is consistent with our data. The ZOIB model was fit using the package *brms* version 2.13.0 (Bürkner, 2017a, 2017b), which employs Bayesian inference. We chose to use a Bayesian analysis because most implementations of ZOIBs are Bayesian. Additionally, *brms* is the most accessible package for modelling ZOIBs that we are aware of, as it makes use of a syntax very similar to the familiar *lme4* package. Moreover, Bayesian analyses allow us to put weakly informative priors on the coefficients, which allows easier convergence in the optimization process. Details of the model settings and prior choices are given in Supplementary Materials D, including an explanation about the basics of the ZOIBs.

Our modeling decision so far is empirically driven. It may seem to go against some other empirical findings that both words and nonce words lie on a continuum of acceptability (e.g. Albright, 2009; Bailey & Hahn, 1998; Coleman & Pierrehumbert, 1997; Hay *et al*., 2004; Hayes & Wilson, 2008; Shademan, 2006 among many) and the claim about the gradient nature of phonotactic wellformedness (Borowsky, 1989; Chomsky & Halle, 1968; Clements & Keyser, 1983; S. Myers, 1987). Thus, if we accept that the grammar plays a role in wordlikeness judgments (Berent *et al*., 2001; Frisch & Zawaydeh, 2001), as opposed to treating the gradient judgments as the product of mere performance (see studies reviewed in Hayes, 2000 and Schütze, 1996), we should first check theoretical significance of our data (i.e. one that justifies the choice of the ZOIB model). Specifically, we need to check if the nature of grammar that generates wordlikeness judgments is both gradient and categorical. Gorman (2013) argued that large number of gradient judgments reported in previous literature may not be due to the gradient nature of wordlikeness system. Instead, the grammar may consist of categorical and gradient components, but gradient judgments were observed more frequently due to the nature of gradient rating tasks. More directly evidence showing both categorical and gradient nature of the grammar comes from Coetzee (2009, computer

science). The study tested wordlikeness judgments from Hebrew and English speakers. Two types of tests were conducted, a wordlikeness rating test on a gradient scale and a comparative wordlikeness test which forced participants to choose a more wordlike item between two grammatical or two ungrammatical ones. The results showed that grammatical and ungrammatical items were rated categorically in a wordlikeness rating test, but the comparative test elicited gradient wordlikeness distinctions from the participants. The results suggest that there are two independent cognitive processes involved in wordlikeness judgments and speakers use their grammar both gradient and categorical ways. Both processes may not be used in all types of wordlikeness tasks, but crucially the nature of the grammar that generates wordlikeness judgments are not only gradient but also categorical, supporting the current choice of the ZOIB model.

As to the modeling decision on syllabic structure, recall that there are two possible options; one is the syllable part approach (Bailey & Hahn, 2001) which decomposes a syllable into onset, nucleus, coda, and tone, and another is a syllable rime approach which decomposes a syllable into onset, rime, and tone. If a syllable rime approach has its psychological reality in Cantonese speakers' mind, thus it should be pursued, we would expect that syllables with unattested rimes tend to be frequently judged categorically bad. This was not borne out in the present data. Refer to the following graph showing the relationship between log-probabilities (with tone dependent on all segments) but color-coded to show whether the rime is attested or not: Grey items have unattested rimes whereas orange items have attested rimes in Cantonese in Figure 7.

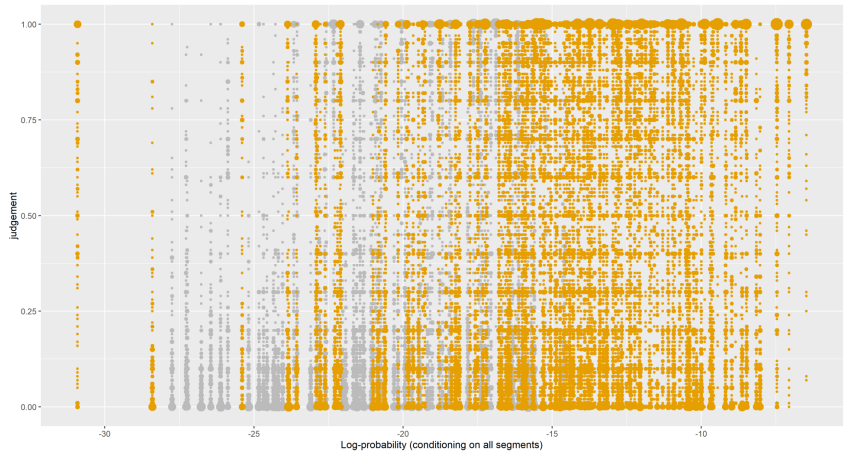


Figure 7. Scatterplot of the log probability (x-axis) against the wordlikeness (y-axis) depending on the attested of rimes (grey: attested rimes vs. orange: unattested rimes). The sizes of the circles indicate the number of items at that log-probability value.

As we can see here, the items with unattested rimes (grey dots) were overall rated less wordlike than the items with attested rimes (orange dots). However, ones with unattested rimes (grey dots) have comparable judgement distributions as ones with attested rimes (orange dots) in the mid log-probability range, e.g. probabilities between -25 and -15, and many of grey dots have a fair number of responses judging the items as categorically wordlike as well. Based on this observation, we decided not to include a syllable rime analysis and pursue a syllabic part analysis where a syllable is decomposed of onset, nucleus, and coda, and tone.

Comparison between phonotactic probability and neighborhood density measures. To examine the role of phonotactic probability and neighborhood density in predicting the wordlikeness judgment data, the ZOIB model we constructed was fit using the combined

measures of phonotactic probability and neighborhood density. This was to identify the relative contributions of phonotactic probability and neighborhood density in predicting the wordlikeness judgments. We fit the models using each of possible pairings of phonotactic probability measures in (4) (six measures: log-probabilities with tonal probability conditioned on (a) all segments, (b) onset only, (c) nucleus only, (d) coda only, (e) tonal probability not conditioned on segments, and (f) with no tonal component) and three neighborhood density measures ((a) NN, (b) GNM, and (c) GCM), along with models that only have phonotactic probability measures (6 in total) or only have neighborhood density measures (3 in total). In total, there were 27 models, i.e. 6×3 combined models + 6 phonotactic probability models + 3 neighborhood density models. We then compared the model fits using WIDELY APPLICABLE INFORMATION CRITERION (Vehtari *et al*., 2017), an approximation of the Akaike Information Criterion which is used as a measure of models' out-of-sample predictive power, i.e. how good the model will be to predict data beyond the current sample.

The full model for the ZOIBs, in principle, includes population-level coefficients for the predictors along with item-level and participant-level random intercepts and participant-level random slopes for all predictors. Due to the limitation of modeling capacity, it was impossible to fit the full ZOIB model for all the combinations we tested. So we fit all the 27 models initially with random intercepts only. Once we identified the optimal model only with random intercepts, then we re-fit the optimal model with both random slopes and intercepts. There were additional modeling complications for GNM. Because of the large sample size, and more crucially, because of highly correlated nature of the three GNM-related predictors, it was not possible to fit the models containing all three GNM values in a timely manner, even with random slopes removed. Thus, for the models involving GNM, we decided to first fit a model with only the three GNM-related quantities as predictors with no random effects. Then we used those values, normalized to sum up to 1, to derive GNM quantities for each syllable, which were used to fit the GNM models including random effects.

Table 5 shows the performance of the models with random effects. Phonotactic probability effect is on the horizontal line and neighborhood density effect is on the vertical line. The performance of each model was based on WAIC values. The WAIC values in each cell in Table 5 indicate the model performance for the combination of the two assumed determinants. For example, the WAIC value of 11437.8 in the top left cell is when phonotactic probabilities are calculated assuming that tone is conditioned on onset and neighborhood density is measured by the number of neighbors. Lower WAIC values indicate better predictive power of the data. Standard errors of the WAIC values are provided in parenthesis in each cell.

	T|O	T|N	T|C	T|S	T	NoT	No phonotactic probability
NN	11437.8 (316.8)	11438.1 (316.9)	11439.8 (317.0)	11440.7 (316.9)	11437.4 (317.0)	11438.2 (316.9)	11439.4 (316.9)
GCM	11439.7 (316.9)	11439.3 (316.9)	11443.4 (317.0)	11442.4 (317.0)	11435.0 (316.9)	11438.1 (316.9)	11442.8 (316.9)
GNM	11434.3 (316.8)	11435.6 (316.8)	11434.5 (316.8)	11437.4 (316.9)	11433.9 (316.8)	11436.8 (316.8)	11436.1 (316.8)
No lexical neighborhood	11437.6 (316.9)	11437.0 (316.9)	11443.6 (317.0)	11441.3 (316.9)	**11433.1** (316.9)	11435.8 (316.9)	

Table 5. WAIC values of the different models and standard errors of the WAICs, with columns indicating the measure of phonotactic probability, and rows indicating the measure of neighborhood density.

The model with only log-probabilities using unconditional tonal probabilities (T) has the lowest WAIC values, indicating its best performance (boldface). However, it only has a slight edge over some other models, especially those with GNM; T with GNM, T|O with GNM, and T|C with GNM. Note, however, though that the WAIC values from the current GNM modeling are not exactly comparable with the WAIC values from the other models. The GNM models did correctly incorporate frequency effects, including the square of the log frequency and the raw frequency itself. The exact weights of the GNM quantities were also calculated from the current wordlikeness judgment data. However, this was done in the 'first round' of the fitting process, and was not factored into the calculation of WAIC in the final model. Recall that this was an inevitable modeling decision, due to the large sample size and intrinsically high correlation of the three GNM-related predictors. In principle, the WAIC values from our 'simplified' GNM modeling are underestimated. Thus, we refrain from interpreting the precise WAIC values from the GNM models, but instead infer the GNM performance from GCM: Given that the GNM is based on GCM but it adds extra complications (rather highly correlated to the GCM value—both extra components have correlation coefficients over 0.99 with the GCM value), we interpret that if there is no evidence that GCM is better than other models, it is unlikely that GNM, with much more parameters but not adding much extra information, performs better. Thus, in the following reports, we compare the models excluding the WAIC values of the GNM (grey-colored in Table 5), and we infer the performance of GNM models from the performance of GCM models.

When the GNM is excluded from the comparison of the exact WAIC values across the models, the best model, i.e. the one with only log-probabilities using unconditional tonal probabilities (T), is still comparable to some other models, such as the one with log-probabilities without tone (NoT), and the one with log-probabilities of unconditional tonal probabilities along with GCM (T with GCM). Given the closeness of their WAIC values and given that the WAIC differences are around the similar size as the standard error differences, it would be inappropriate to choose the optimal model from these values alone. We decided to first identify whether the performance differences of the six types of measurements of tonal probabilities (i.e. the horizontal line in Table 5 except for No phonotactic probability) are meaningful. In other words, we examined if one tonal representation has a better predictive power than others. The WAIC value differences between the optimal model (T) and the best models for each tonal representation were compared: These best models include tone given onset with no lexical effect, tone given nucleus with no lexical effect, tone given coda with NN, tone given all segments with NN, and a representation excluding tone with no lexical effect. The WAIC differences, along with standard errors, are shown in Table 6.

	T|O	T|N	T|C	T|S	NoT
lexical effect	None	None	NN	NN	None
WAIC diff.	−1.3	−1.0	−2.4	−2.8	−0.4
SE	2.0	1.4	1.9	2	1.4

Table 6. Differences in WAIC between the unconditioned tone model and the other models among other tonal representations.

In Table 6, the WAIC differences among the models are quite small considering their standard errors: Most of the times, the magnitude of the difference is smaller than the

standard error. Even for the greatest difference, i.e. the one with tone condition on all segments (T|S), the WAIC difference is only slightly greater than the standard error. From this observation that the optimal models per each tonal representation do not significantly vary in their performance, we conclude that there are no grounds for preferring one tonal representation over others. Therefore, we report the results based on the optimal model (T) in the following discussion for the purpose of presentation, but it should be understood that the results are comparable across different tonal probabilities we tested.

To understand the exact relation between log-probability and the wordlikeness judgments, we examined the coefficient estimates of the models. We fit the best-performing model with full random effects for all predictors in three regions considered, i.e. gradient judgments (between zero and one) and two categorical judgments (judgments of 0s and 1s). This was done using the model with unconditioned tonal representation (T) and no neighbourhood effects. Then, we examined point and interval estimates of the coefficients. Since the different models have roughly similar performance in terms of WAIC, we also ran a robustness check known as multiverse analysis (e.g. Steegen et al., 2016). This was performed using the models without random effects in Table 5. That is, we examined a variety of logically possible ways to do the analysis, in this case all the different tonal representations and neighbourhood representations, and then examined the coefficient estimates in each one.

The results for the optimal model are in Table 7. The 95% credible intervals (CI)[6] indicate the range of values for which we can be 95% sure that the coefficient lies in. 95% CI that excludes zero would indicate strong evidence that the coefficient is nonzero, meaning that the evidence is sufficient to support the considered effect.

	Estimate	**Estimated Error**	**Lower 95% CI**	**Upper 95% CI**
Beta regression component (between zero and one)				
Intercept	0.769	0.144	0.490	1.058
Logprob$_T$	0.079	0.009	0.061	0.097
Logistic regression component (ones)				
Intercept	8.480	1.125	6.403	11.96
Logprob$_T$	0.509	0.069	0.375	0.647
Logistic regression component (zeros)				
Intercept	−1.324	0.265	−1.853	−0.826
Logprob$_T$	0.039	0.018	0.003	0.073

Table 7. The overall results of the optimal model.

Higher log-probabilities lead to intermediate judgements (between zeros and ones) being higher in general, since the coefficient of the log-probabilities in the beta regression component excludes zero (0.061, 0.097). Evidence is sufficient that higher log-probabilities substantially enhance the chances of items being judged as 1 (i.e. very wordlike), since the 95% credible interval for its coefficient in the logistic regression component for ones does not include zero (0.375, 0.647). Also, we have some evidence for the logistic regression component for zeroes (i.e. not at all wordlike) being affected by log-probabilities, since the coefficient excludes zero (0.003, 0.073), though in an unexpected direction whereby higher log-probabilities makes zero judgements *more* frequent. This aligns with our descriptive

[6] This is different from frequentist confidence intervals (also abbreviated as CIs), which cannot be interpreted in such terms.

observations in Section 4.2 such that the judgments of 0s seem to be less affected by log-probabilities, compared to those of 1s and intermediate judgements. So we do have clear evidence that log-probability does contribute to the determination of sound sequences as categorically legitimate Cantonese words and more or less wordlike at a gradient level.

Next, we performed a multiverse analysis as a robustness check for the effects obtained above. We examine the CIs for the coefficients for phonotactic probability and neighborhood density under each possible pairing of phonotactic probability and neighborhood density measures in Table 5. The examination of the CIs was conducted for each of the three regions, 0s, 1s, and gradient judgments. CIs that exclude 0, i.e. do not touch the red line, indicate evidence that the predictor is effective. First, the effect of phonotactic probability is examined in Figures 8a–c. The exact numerical values are given in Supplementary Materials E.

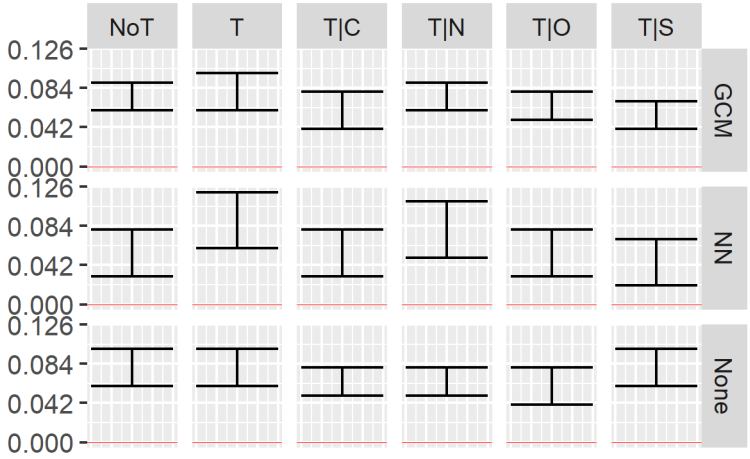


Figure 8a. Multiverse results for the 95% CI of the effect of log-probabilities on gradient judgements.

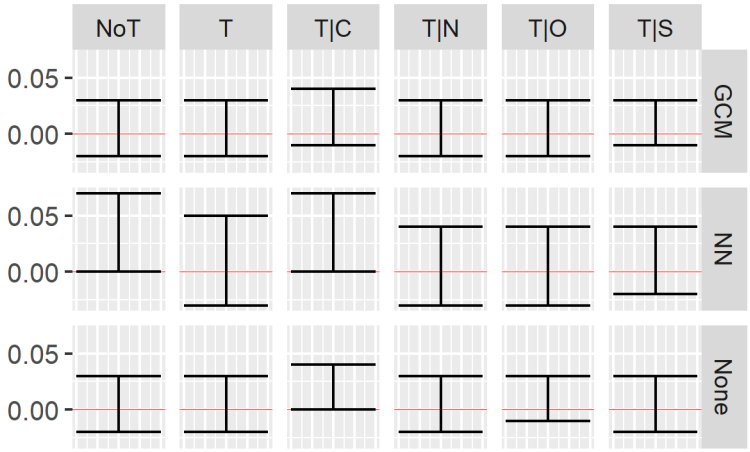


Figure 8b. Multiverse results for the 95% CI of the effect of log-probabilities on the judgments of 0s.

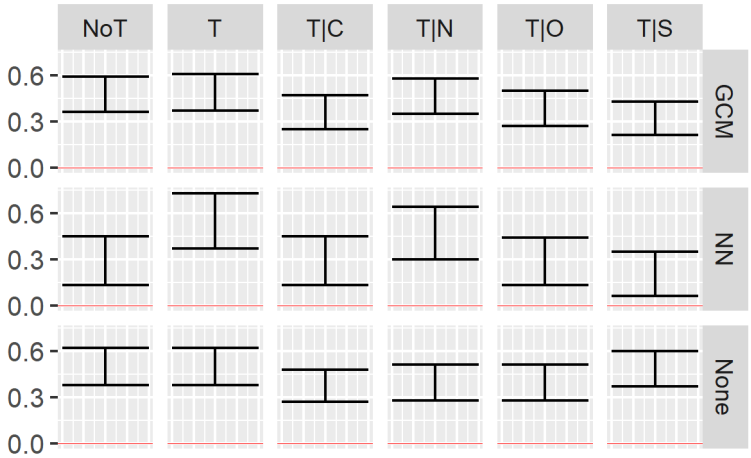


Figure 8c. Multiverse results for the 95% CI of the effect of log-probabilities on the judgments of 1s.

As clearly seen from the above figures, in the case of gradient judgements (Figure 8a) and the judgments of 1s (Figure 8c), the effect of phonotactic probability on judgments is completely robust regardless of the tonal representations. However, only two possible decisions lead to a small effect on the judgments of 0 (Figure 8b), and both are very marginal. Given the results, we confirm that the effect of log-probability is only on the judgments of 1s and gradient judgements, but not on the judgments of 0s. Now the analysis of neighborhood density is given in Figures 9a–c.

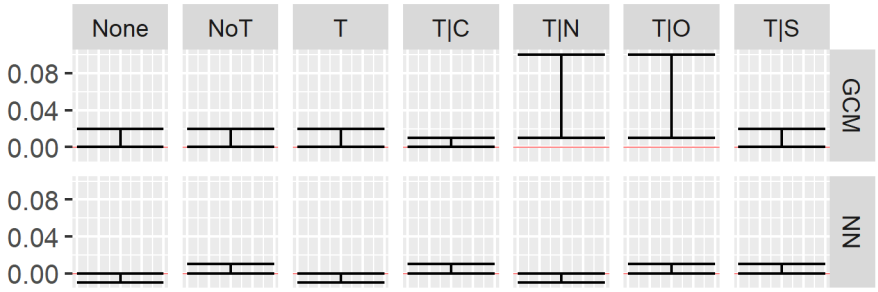


Figure 9a. Multiverse results for the 95% CI of the effect of neighborhood density on gradient judgements

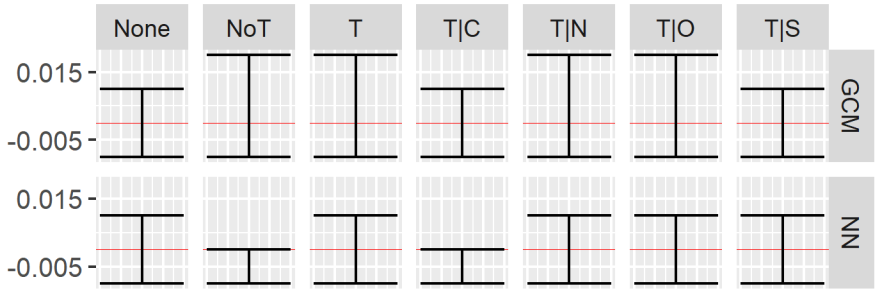


Figure 9b. Multiverse results for the 95% CI of the effect of neighborhood density on the judgments of 0s.

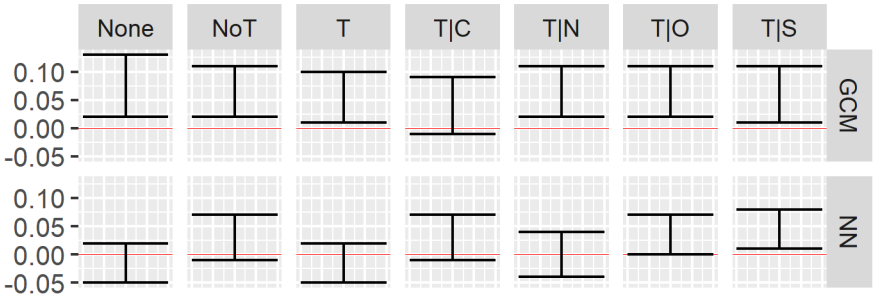


Figure 9c. Multiverse results for the 95% CI of the effect of neighborhood density on the judgments of 1s.

For the judgments of 0s (Figure 9b), every model shows CIs including 0, suggesting no neighborhood density effect of on the judgements of 0s. For the effects on gradient judgements (Figure 9a) and judgments of 1s (Figure 9c), the presence of a CI excluding 0, i.e. neighborhood density effect, hinges crucially on the choices between NN and GCM. Specifically, models with GCM almost always have a GCM coefficient excluding 0. Models with NN rarely have a NN coefficient excluding 0. This suggests that between NN and GCM, GCM is a better predictor of the wordlikeness judgements. However, the models with GCM have *not* been shown to perform better than those without GCM in terms of WAIC values in Table 5. Recall we note in Table 5 that under *no* choice of log-probability measure does the model using GCM outperform the model with no neighborhood density measures. Therefore, while GCM appears to be a better predictor than NN from these graphs, our data still do not provide sufficient evidence for a GCM effect. In sum, our modeling results suggest that log-probability, regardless of tonal representations, does play a role in predicting the wordlikeness data for the categorically wordlike items and gradient items, but not for categorically not wordlike items. We do not have sufficient evidence for the role of neighborhood density.

To summarize, log-probability predicts the wordlikeness judgments within gradient judgment areas and categorical judgments for 1s (very wordlike), but not for categorical judgments for 0s (not at all wordlike). This effect is robust across different ways that tonal probabilities depend on segmental probabilities that we have considered, and we have little evidence in favor of one tonal representation over the rest. The number of neighbors is not a predictor of the current wordlikeness judgement data. Although we have suggestive, yet by no means conclusive, evidence that GCM may be involved in the prediction of gradient judgements and the judgments of 1s, we can still be confident that it rarely plays a role in the judgements of 0s. Given the limited GCM effect, we infer that the GNM effect will be extremely minor, if it ever exists.

Comparison of the relative contribution of syllable components. So far, we have only investigated the effect on phonotactic judgements of the log-probability of the items as a whole. An assumption behind such consideration is that the different parts of a syllable that make up these probabilities are equally important. Recall that we also aimed to examine whether the different components of the syllable (onset, nucleus, coda, tone) differ in their importance in determining wordlikeness judgements. To determine the relative roles of different syllable components towards the prediction of wordlikeness judgements, we separated the log-probabilities in Table 7 into the four syllable-component probabilities—onset, nucleus, coda, and tone—and allowed the model to assign separate coefficients to each

syllable component.[7] The results in Table 10 are based on the assumption that tone is conditioned on onset, but the general trends are same for other tonal representations that we examined as well.[8]

	Estimate	Estimated Error	Lower 95% CI	Upper 95% CI
Beta regression component (between zero and one)				
Intercept	0.583	0.170	0.249	0.919
ln P(Onset)	0.055	0.029	0.000	0.112
ln P(Nucleus|Onset)	0.131	0.016	0.099	0.163
ln P(Coda|Nucleus)	0.048	0.012	0.025	0.072
ln P(Tone|Onset)	−0.002	0.017	−0.035	0.030
Logistic regression component (zeros)				
Intercept	−1.625	0.266	−2.142	−1.109
ln P(Onset)	0.039	0.039	−0.036	0.117
ln P(Nucleus|Onset)	0.013	0.021	−0.030	0.055
ln P(Coda|Nucleus)	−0.002	0.016	−0.033	0.029
ln P(Tone|Onset)	−0.001	0.023	−0.047	0.044
Logistic regression component (ones)				
Intercept	6.957	1.157	4.739	9.269
ln P(Onset)	0.407	0.175	0.060	0.762
ln P(Nucleus|Onset)	0.700	0.103	0.503	0.911
ln P(Coda|Nucleus)	0.349	0.074	0.207	0.495
ln P(Tone|Onset)	−0.088	0.105	−0.297	0.118

Table 10. The results with the decomposition of syllabic components.

First, for the beta regression component (gradient judgments), we have strong evidence for effects of nucleus and coda, because their CIs exclude zero; we also have sufficient evidence that the effect of nucleus given onset is greater than that of coda given nucleus in the beta regression component (estimated difference between nucleus and coda: 0.08; SE: 0.02; 95% CI: (0.05, 0.12)). Moreover, we have some evidence that the probability of the onset matters.

[7] One might question whether the estimates of the coefficient values for different syllable components are comparable, because they are logs of different types of probabilities. Note that under the assumptions that the following syllable component only depends on the preceding one, e.g. nucleus only depends the onset, the coda only depends on the nucleus, each of the covariates, like ln P(Nucleus|Onset), or ln P(Coda|Nucleus), may be considered −1 times of the surprisal (Levy & Gibson, 2013) of the corresponding syllable component. The only difference from a usual definition of surprisal is that the surprisal is measured in nats (Cover & Thomas, 2012) instead of bits in our case, because we are taking natural instead of base-2 logarithm. Since the different predictors take the same unit (nats), their coefficients should be considered comparable, as the coefficients can be interpreted in the same way. For the beta regression component, the coefficients should be interpreted as 'for each unit increase of the surprisal in nats, the mean judgement decreases by the coefficient'. For the logistic regression component, the coefficients should be interpreted as 'the log-odds ratio of the probability two syllables being judged as 0/1, the first one having a surprisal value of 1 less than the second one, is the value of the coefficient.'

[8] For some representations, we could not get the models to converge.

This effect is also smaller than the nucleus effect (estimated difference between onset and nucleus: −0.08, SE: 0.03, 95% CI: (−0.13, −0.02)). We have no evidence for a difference between onset and coda (estimated difference: 0.01, SE: 0.03, 95% CI: (−0.04, 0.06)). We have no evidence for a tonal effect either. Second, in the logistic regression component for ones (very wordlike), we see the same situation with the coefficients of all three segmental probabilities excluding 0. We do not have evidence that their coefficients are different: The estimated difference between onset and nucleus is 0.03 (SE: 0.04; 95% CI: (−0.05, 0.1)), the one between onset and coda is 0.04 (SE: 0.04; 95% CI: (−0.03, 0.11)), and the one between nucleus and coda is 0.01 (SE: 0.03, 95% CI: (−0.03, 0.06)). Third, we have no significant predictors for the judgments of zeroes, consistent with the results in the previous subsection. Thus, the general tendency such that the log-probability can predict gradient judgements and the categorical judgments of 1s, but not those for 0s, is consistent across the model assuming a syllable as a whole (Table 6) and the model with the decomposition of syllabic components (Table 10).

To summarize, we find evidence that phonotactic log-probability is a good predictor of the current wordlikeness judgement data. We do not find evidence that neighborhood density contributes to the prediction of the current data patterns. Moreover, if we split up the log-probability into its syllabic component parts, we find that the conditional probabilities of nucleus and coda play a crucial role; onset is less important but still it plays a role to a certain degree, but tone does not. Finally, such effects tell us how likely participants are to rate an item as perfect (1) or, if their rating falls between 0 and 1, how likely they are to rate it higher; but the predictors do not affect the rate at which participants consider the items not at all wordlike.

5. General discussion

5.1. Phonotactics vs. lexical neighborhoods

Our finding that mainly phonotactic log-probability, but not neighborhood density, is important in predicting wordlikeness judgements goes against some previous studies, including Kirby & Yu (2007) who also tested wordlikeness in Cantonese, although their research focus was specifically on lexical gaps. Recall that Kirby & Yu (2007) found the relative *weakness* of phonotactic probability and a *stronger* effect of neighborhood density. Since our study and theirs both tested Cantonese, despite differences of the exact research questions, it is worth considering the different results more in detail. Kirby & Yu attributed their findings to the fact that Cantonese makes use of larger space of possible monosyllabic words than some other languages like English. Because of strict phonotactic restrictions of Cantonese, possible phonotactic combinations are more limited compared to other languages like English. Due to this, proportionally, large portions of limited phonotactic space are taken by real words in Cantonese. If so, native speakers rely more on the lexicon in making wordlikeness judgments. Kirby & Yu also pointed out that due to the high rate of words in the limited phonotactic space, many nonwords have their lexical neighbors which might encourage speakers' reliance on lexical neighbors in making wordlikeness judgments. This idea was further pursued by Shoemark (2013), who argued that because the connectivity of Cantonese phonological networks is denser than those of English, great proportion of the Cantonese lexicon is activated by any nonword. Beyond the work on Cantonese, our results are also against Bailey & Hahn (2001) and Myers (2016), who found independent effects of lexicon and phonotactics in English and Mandarin respectively. Gorman (2013) also reported a major role of neighborhood density in English, which is different from the current results.

Our results are, however, in line with Frisch *et al*. (2000) and Albright (2009). Frisch *et al*. (2000) reported that English native speakers' wordlikeness judgments of multisyllabic nonwords was better predicted by phonotactic probability than by neighborhood density, although the difference was only marginal. Albright (2009) found that, although judgements are correlated with lexical neighborhood measures at a descriptive statistical level, they were not found to be significant in the regression model. There are several ways to account for such discrepancies across different studies.

First, the differences might be due to a research design, specifically the inclusion of real words in some experiments. For example, Bailey & Hahn (2001) included real words, and in Kirby & Yu (2007), over one third of test items were real words (162 out of 432 items). This was not the case in our experimental design where only nonwords were tested. In fact, some studies which reported no neighborhood density effect did not include real words either (Albright, 2009; Frisch *et al*., 2000). As argued by Vitevitch & Luce (1998, 1999) and Shademan (2006), the processing of real words is highly dominated by lexical influences. This idea is supported by Myers and Tsay (2005) who found the lexical effects to the judgments of real words in Mandarin but no such effect to nonwords. The inclusion of real words may encourage lexical access, and therefore strong lexical effects might have observed. However, note also that some studies which included only nonwords did report strong effects of neighborhood density (Gorman, 2013; Myer, 2016). This suggests that the method of a stimulus selection might affect the results and the conclusions drawn from different wordlikeness studies but the method itself is not a sole factor determining the lexical effect to wordlikeness judgments.

Second, another possibility is related to the size of syllable inventory, which differs across languages. Compared to some languages like English, Cantonese has highly restricted phonotactics, allowing no consonant clusters and a fairly limited set of codas, including only an oral stop series, a nasal stop series, and /i/ and /u/. Myers (2016) argued that for languages involving a small syllable inventory due to their strict phonotactic restrictions (e.g., Mandarin and Cantonese), lexical neighborhoods are more important than phonotactic probability, because the small numbers involved in the inventory makes syllables easier to access from rote memory. Along the same logic, in languages like English, where there are larger number of syllables, speakers rely less on lexical neighborhoods, because there are too many syllables they need to access, which makes the process too complicated. This idea is similar to Kirby & Yu (2007) where the strict phonotactic restrictions were argued to encourage the lexical effect because a language with strict phonotactic restrictions makes use of a larger proportion of limited phonotactic possibilities. This line of logic predicts that our study on Cantonese wordlikeness judgments should have observed a strong lexical effect, but it was not the case. We believe that there is an alternative way to consider the relation between the level of phonotactic restrictions and the role of neighborhood density to wordlikeness judgments. As Kirby & Yu (2007) and Myers (2016) argued, if phonotactic restrictions are very strict in a certain language, phonotactically possible patterns are limited. This will not only result in limited syllable inventory, but also result in relatively little *variation* in lexical density, compared to languages that allow varying degree of phonotactic combinations (e.g., complex onsets and codas) such as English. For example, English syllables can go up to seven segments (as in the word *strengths*), while Cantonese syllables only go up to three segments and a tone. So, in principle, the range of values covered by Cantonese phonological space is only from 0 to 4 (including tone), when a distance of 1 is assume for each syllabic component, whereas in English, it varies more widely, from 0 to 7. Due to this limited phonological space taken by Cantonese words, lexical neighborhood effects may not be as significant as in languages like English where phonotactic patterns are more varying and complicated, thus taking wider phonological space. Even if there were the lexical effect, it is

possible that the effect would be difficult to estimate, since the range of the independent variable is too narrow. Further investigations are needed to identify the exact relation between the degree of variation in lexical density or in phonological space in languages and the role of lexical effects in wordlikeness judgments. Crucially, considering that wordlikeness judgment tests using the same languages frequently yielded contrastive results, including English (e.g. Albright, 2009; Bailey & Hahn, 2001) and Cantonese (e.g. Kirby & Yu, 2007; the current study), language-specific phonotactic factors are important but they should not be treated deterministic in predicting wordlikeness judgments for specific languages.

Third, another possible explanation is related to speakers' different perception of nonwords depending on different morphological systems of different languages. As introduced, many modern Cantonese monosyllabic morphemes are generally not used as independent words *per se*, but only appear in compounds (Bauer & Benedict, 1997). This is different from some other languages like English. For instance, the morpheme 則 *zak1* [tsɐk̚˥] is used in many common words such as 規則 *kwai1zak1* [kʷʰɐi˥tsɐk̚˥] 'rules, regulations', 守則 *sau2zak1* [sɐu˧˥tsɐk̚˥] 'regulation, code of conduct', etc., but the monosyllabic morpheme does not really mean anything on its own.[9] Previous work consistently suggested that the syllables, not the individual phonemes, are the fundamental unit in Chinese languages: Alpatov (1996) described the syllables in Chinese as 'the most important psycholinguistic units' and O'Seaghdha *et al*. (2010) called the syllables in Mandarin 'proximal unit'. So there is little doubt that Cantonese speakers can easily recognize and process monosyllabic items, which are the basic unit in their mind. However, they may not regard monosyllabic items as independent words which can potentially have corresponding Chinese characters bearing their own meanings, because of monosyllables' frequent involvements in compounds. In fact, Chan *et al*. (2011) casted doubt on testing Cantonese speakers with monosyllabic nonwords, from which they argued that nonwords created based on one language's phoneme inventory and phonotactic regulations are different from nonwords created based on other languages. A clinical work by Stokes *et al*. (2006) reported the failure of monosyllable-based nonword repetition test to discriminate children with specific language impairment in Cantonese while studies in English found evidence that the monosyllabic nonword repetition test serves as a meaningful clinical marker (see a meta-analysis in Estes *et al*., 2007). These may suggest that monosyllabic nonwords in Cantonese may have different status as those in English, a factor to which the current results might be attributed. The current study conducted neighborhood analysis based on syllables, where an example like 則 *zak1* [tsɐk̚˥] was counted as a neighbor of a stimulus. Considering that such syllable is not used as an independent word, it is conceivable that the results would differ when the neighborhood analysis is based on words. Future work involves identifying which neighborhood analysis matches better with speakers' judgments in Cantonese. Additionally, a cross-linguistic exploration of nonwords processing is needed differing in their morphological systems.

We suggested the effects of the stimulus selection methods, language-specific phonotactic complexity, and language-specific morphological systems in determining the predictors of wordlikeness judgments. Crucially, the conclusions drawn from similar methods or from same languages differ from each other. This suggests that the predictors of wordlikeness judgments should be considered in a comprehensive way including research design-specific factors as well as language-specific factors and that the exact correlations of each factor should be further identified to correctly model wordlikeness judgments.

[9] 則 *zak1* [tsɐk̚˥] is a function word in Standard Written Chinese and Classical Chinese, and so is an independent morpheme when written texts in these languages are read using Cantonese readings of characters, but it is not common in spoken Cantonese.

5.2. The roles of syllabic components

When splitting up the phonotactic log-probabilities into syllabic components, we found that the conditional probabilities of nucleus and coda matter most, those of onset marginally matter, but those of tone do not play a role. Note that our experiment only used permissible phonemes and tones in Cantonese. Compared to nuclei and codas, where the judgment can be based on their concurrence with the previous phonemes, onsets can be judged on their own given their initial position. So, it is not surprising that permissible onsets were simply treated all 'acceptable', resulting in low weight in onsets when judging items' wordlikeness. What is surprising is no major role of conditional probabilities of tone, given that there is at least one highly robust generalization about Cantonese phonotactics whereby oral stop codas are only compatible with tones ˥, ˨ and ˩ (and ˧, where there is a tone change from one of these tones). This result, though, goes with the previous studies on lexical access in Mandarin showing that tone is playing a minor role than other syllabic components (Taft & Chen, 1992), especially among monosyllables (Lin, 2016). Why do results from different experimental paradigms and across two languages consistently show that tone is less important than segments in lexical processing? We consider that the results can be accounted for when lexical predictability of syllabic components is taken into consideration. One way to measure how 'predictable' a component is in a lexicon is measuring its functional load. In Cantonese, for example, it has been found that onsets and tones have higher functional load than nuclei and codas (Do & Lai, forthcoming), where functional load is defined as the entropy of the language contrasts in a syllable component divided by the actual entropy of the language (e.g. Hockett 1966). This suggests that nuclei and codas are lexically more predictable (i.e. restricted) than onsets and tones in Cantonese, and hence play a smaller role in discriminating between lexical items. Thus, our results may tentatively be predicted as saying that lexically less predictable aspects of an item are more likely to contribute to wordlikeness judgements.

The weaker reliance on tone is additionally in line with the results from Cantonese perception studies. Phonological awareness tests in Cham (2003) reported that Cantonese speakers performed poorer in tone awareness tasks compared to segment awareness tasks, suggesting that tones are perceptually less salient than segments for Cantonese speakers. Studies on spoken word recognition also showed that word recognition is more challenging when tone differences were involved (e.g. Cutler & Chen, 1997; Keung & Hoosain, 1979), which imply listeners' lower sensitivity to tone differences than segment differences. If so, the current study may suggest that speakers rely less on the syllabic component that is perceptually less salient and more on perceptually more salient ones.

5.3. Categorical vs. gradient judgments

Our final finding is that log-probability of syllabic components affects only the tendency to judge words being absolutely perfect or between the two extremes. They do not affect the probability that the participants will judge items as absolutely unacceptable. This suggests that potentially there are two different cognitive processes involved in wordlikeness judgements, for only one of which we have a solid predictor. The other one, one for categorically bad judgements, remains poorly understood. It is not surprising though that phonotactically illicit items are processed in a different way from absolutely grammatical and gradient items. Large amount of evidence showed difficulties in processing phonotactically illicit sequences in perception (e.g. Berent *et al*., 2007; Dupoux *et al*., 1999; Kabak & Idsardi, 2007) and in production (Davidson, 2005, 2006a, 2006b; Rose & King, 2007; Vitevitch & Luce, 1998, 2005), which may suggest speakers' limited ability to process the representations

of illicit sequences (Gorman, 2013). While the gradient nature of wordlikeness judgments has been widely recognized (Bailey & Hahn, 2001; Coleman & Pierrehumbert, 1997; Frisch *et al*., 2000; Hayes, 2000; Ohala & Ohala, 1986), the exact processes involved in the judgments for absolutely perfect vs. absolutely bad vs. gradient are yet to be known. We refrain from speculating as to why this is the case and whether this is generalizable to other languages and tasks.

Modeling work on wordlikeness judgments has shown that phonotactic probability and neighborhood density are crucial determinants of speakers' judgments (Bailey & Hahn, 1998, 2001; Frisch *et al*., 2000). However, the full understanding of speakers' phonotactic knowledge has yet to be obtained, given the lack of research focus on suprasegmental features in phonotactic modeling work. Our paper was an attempt to model wordlikeness judgments incorporating tone. Future work is to test other tonal languages, on the basis of the methodologies presented in the current study, so that the determinants of speakers' wordlikeness judgments can be understood inclusive of segments and tones.

References

Albright, Adam (2007). *Gradient phonological acceptability as a grammatical effect*. Ms, MIT.

Albright, Adam (2009). Feature-based generalisation as a source of gradient acceptability. *Phonology* **26**. 9–41.

Albright, Adam & Bruce Hayes (2003). Rules vs. analogy in English past tenses: a computational/experimental study. *Cognition* **90**. 119–161.

Alpatov, Vladimir Mikhaylovich (1996). On the notion of submorph and its applicability to the languages of East and South-East Asia. In *Pan-Asiatic Linguistics: Proceedings of the Fourth International Symposium on Languages and Linguistics, January 8-10, 1996*. Thailand: Institute of Language and Culture for Rural Development, Mahidol University at Salaya. 1799–1805.

Bailey, Todd M. & Ulrike Hahn (1998). Determinants of wordlikeness. In Morton Ann Gernsbacher & Sharon J. Derry (eds.) *Proceedings of the twentieth annual conference of the Cognitive Science Society: August 1-4, 1998, University of Wisconsin-Madison*. Mahwah, N.J.; London: Lawrence Erlbaum. 90-95.

Bailey, Todd M. & Ulrike Hahn (2001). Determinants of wordlikeness: phonotactics or lexical neighborhoods? *Journal of Memory and Language* **44**. 568–591.

Bauer, Robert S. (1985). The expanding syllabary of Hong Kong Cantonese. *Cahiers de Linguistique Asie Orientale* **14**. 99–111.

Bauer, Robert S. & Paul K. Benedict (1997). *Modern Cantonese phonology*. Berlin; New York: Mouton de Gruyter.

Berent, Iris, Joseph Shimron & Vered Vaknin (2001). Phonological constraints on reading: evidence from the Obligatory Contour Principle. *Journal of Memory and Language*, **44**. 644–665.

Berent, Iris, Donca Steriade, Tracy Lennertz & Vered Vaknin (2007). What we know about what we have never heard: evidence from perceptual illusions. *Cognition* **104**. 591–630.

Bird, Steven & T. Mark Ellison (1994). One-level phonology: autosegmental representations and rules as finite automata. *Computational Linguistics* **20**. 55–90.

Boersma, Paul & David Weenink (2019). *Praat: doing phonetics by computer* (version 6.0.49). http://www.praat.org/.

Borowsky, Toni (1989). Structure preservation and the syllable coda in English. *Natural Language & Linguistic Theory* **7**. 145–166.

Bürkner, Paul-Christian (2017a). brms: an R package for Bayesian multilevel models using stan. *Journal of Statistical Software* **80**. 1–28.

Bürkner, Paul-Christian (2017b). Advanced Bayesian multilevel modeling with the R package brms. arXiv:1705.11123v2 **[stat.CO]**.

Cham, Hoi Yee (2003). *A cross-linguistic study of the development of the perception of lexical tones and phones*. BSc thesis, The University of Hong Kong.

Chan, Erica, Peter Skehan & Gwendolyn Gong (2011). Working memory, phonemic coding ability and foreign language aptitude: potential for construction of specific language aptitude tests – the case of Cantonese. *Ilha Do Desterro: A Journal of English Language, Literatures in English and Cultural Studies* **60**. 45–73.

Chao, Yuen Ren (1934). The non-uniqueness of phonemic solutions of phonetic systems. *Bulletin of the Institute of History and Philology Academia Sinica* **4**. 363–398.

Cheung, Kwan Hin (1986). *The phonology of present—Day Cantonese*. PhD dissertation, University of London.

Cheung, Sik Lee (1991). The notion of "result" in Cantonese children. *Papers and Reports on Child Language Development* **30**. 17–24.

Chomsky, Noam & Morris Halle (1968). *The sound pattern of English*. New York: Harper & Row.
Clements, George N. & Samuel Jay Keyser (1983). *CV phonology: a generative theory of the syllable*. Cambridge, MA: MIT Press.
Coetzee, Andries W. (2009). Grammar is both categorical and gradient. In Stephen George Parker (ed.) *Phonological argumentation: essays on evidence and motivation*. Oakville, CT: Equinox Pub. Ltd. 9–42.
Coleman, John & Janet Pierrehumbert (1997). Stochastic phonological grammars and acceptability. arXiv:cmp-lg/9707017v1.
Cover, Thomas M. & Joy A. Thomas (2012). *Elements of information theory* (2nd ed.). Somerset: John Wiley & Sons.
Cutler, Anne & Hsuan-Chih Chen (1997). Lexical tone in Cantonese spoken-word processing. *Perception & Psychophysics* **59**. 165–179.
Daland, Robert, Bruce Hayes, James White, Marc Garellek, Andrea Davis & Ingrid Norrmann (2011). Explaining sonority projection effects. *Phonology* **28**. 197–234.
Dankoviþová, Jana, Paula West, John Coleman & Andrew Slater (1998). *Phonotactic grammaticality is gradient*. Poster presented at the 6th International Conference on Laboratory Phonology, University of York, 2–4 July 1998.
Dautriche, Isabelle, Kyle Mahowald, Edward Gibson, Anne Christophe & Steven T. Piantadosi (2017). Words cluster phonetically beyond phonotactic regularities. *Cognition* **163**. 128–145.
Davidson, Lisa (2005). Addressing phonological questions with ultrasound. *Clinical Linguistics & Phonetics* **19**. 619–633.
Davidson, Lisa (2006a). Phonotactics and articulatory coordination interact in phonology: evidence from nonnative production. *Cognitive Science* **30**. 837–862.
Davidson, Lisa (2006b). Phonology, phonetics, or frequency: influences on the production of non-native sequences. *Journal of Phonetics* **34**. 104–137.
Denby, Thomas, Jeffrey Schecter, Sean Arn, Svetlin Dimov & Matthew Goldrick (2018). Contextual variability and exemplar strength in phonotactic learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **44**. 280–294.
Do, Youngah & Ryan Ka Yau Lai (forthcoming). Accounting for lexical tones when modeling phonological distance. *Language*.
Dupoux, Emmanuel, Kazohiko Kakehi, Yuki Hirose, Christophe Pallier & Jacques Mehler (1999). Epenthetic vowels in Japanese: a perceptual illusion? *Journal of Experimental Psychology: Human Perception and Performance* **25**. 1568–1578.
Estes, Katharine Graf, Julia L. Evans & Nicole M. Else-Quest (2007). Differences in the nonword repetition performance of children with and without specific language impairment: a meta-analysis. *Journal of Speech, Language, and Hearing Research* **50**. 177–195.
Frisch, Stefan A., Michael Broe & Janet Pierrehumbert (1997). *Similarity and phonotactics in Arabic*. Available as ROA-223 from the Rutgers Optimality Archive.
Frisch, Stefan A., Nathan R. Large & David B. Pisoni (2000). Perception of wordlikeness: effects of segment probability and length on the processing of nonwords. *Journal of Memory and Language* **42**. 481–496.
Frisch, Stefan A. & Bushra Adnan Zawaydeh (2001). The psychological reality of OCP-Place in Arabic. *Language* **77**. 91–106.
Futrell, Richard, Adam Albright, Peter Graff & Timothy J. O'Donnell (2017). A generative model of phonotactics. *Transactions of the Association for Computational Linguistics* **5**. 73–86.

Gandour, Jack (1981). Perceptual dimensions of tone: evidence from Cantonese. *Journal of Chinese Linguistics* **9**. 20–36.
Gathercole, Susan E. & Amanda J. Martin (1996). Interactive processes in phonological memory. In Susan E. Gathercole (ed.) *Models of short-term memory*. Hove: Psychology Press. 73–100.
Gelman, Andrew (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* **1**. 515–533.
Gong, Shuxiao (2017). Grammaticality and lexical statistics in Chinese unnatural phonotactics. *UCL Working Papers in Linguistics* **29**. 1–23.
Gong, Shuxiao & Jie Zhang (2020). Gradient acceptability in Mandarin nonword judgment. In *Proceedings of the 2019 Annual Meeting on Phonology* **8**.
Gorman, Kyle (2013). Categorical and gradient aspects of wordlikeness. *NELS* **43**.
Greenberg, Joseph. H. & James J. Jenkins (1964). Studies in the psychological correlates of the sound system of American English. *WORD* **20**. 157–177.
Hay, Jennifer, Janet Pierrehumbert & Mary E. Beckman (2004). Speech perception, well-formedness and the statistics of the lexicon. In John Local *et al*. (eds.) *Phonetic interpretation: papers in Laboratory Phonology VI*. Cambridge: Cambridge University Press. 58–74.
Hayes, Bruce (2000). Gradient well-formedness in optimality theory. In Joost Dekkers *et al*. (eds.) *Optimality theory: phonology, syntax, and acquisition*. Oxford; New York: Oxford University Press. 88–120.
Hayes, Bruce & Colin Wilson (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* **39**. 379–440.
Hockett, Charles Francis (1966). *The quantification of functional load: a linguistic problem* (RM-5168-PR). Available at https://eric.ed.gov/?id=ED011649/.
Jurafsky, Dan & James H. Martin (2009). *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed.). Upper Saddle River, N.J.: Pearson Prentice Hall.
Jurafsky, Dan & James H. Martin (2019). *Speech and language processing* (3rd ed. draft). Stanford Stanford University.
Jusczyk, Peter W., Paul A. Luce & Jan Charles-Luce (1994). Infants′ sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language* **33**. 630–645.
Kabak, Bariş & William J. Idsardi (2007). Perceptual distortions in the adaptation of English consonant clusters: syllable structure or consonantal contact constraints? *Language and Speech* **50**. 23–52.
Kessler, Brett (2005). Phonetic comparison algorithms. *Transactions of the Philological Society* **103**. 243–260.
Keung, Tsang & Rumjahn Hoosain (1979). Segmental phonemes and tonal phonemes in comprehension of Cantonese. *Psychologia: An International Journal of Psychology in the Orient* **22**. 222–224.
Khouw, Edward & Valter Ciocca (2007). Perceptual correlates of Cantonese tones. *Journal of Phonetics* **35**. 104–117.
Kirby, James P. & Alan C. L. Yu (2007). Lexical and phonotactic effects on wordlikeness judgements in Cantonese. In Jürgen Trouvain & William J. Barry (eds.) *Proceedings of the International Congress of the Phonetic Sciences XVI*. Saarbrücken, Germany: Universität des Saarlandes. 1389–1392.
Ladefoged, Peter (1975). *A course in phonetics*. New York: Harcourt Brace Jovanovich.
Lai, Regine & Grégoire Winterstein (2020). Cifu: a frequency lexicon of Hong Kong Cantonese. In Nicoletta Calzolari *et al*. (eds.) *Proceedings of the 12th Language*

Resources and Evaluation Conference. Marseille, France: European Language Resources Association. 3069–3077.
Levy, Roger & Edward Gibson (2013). Surprisal, the PDC, and the primary locus of processing difficulty in relative clauses. *Frontiers in Psychology* **4**. 229.
Lewandowski, Daniel, Dorota Kurowicka & Harry Joe (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis* **100**. 1989–2001.
Light, Timothy (1977). The Cantonese final: an exercise in indigenous analysis. *Journal of Chinese Linguistics* **5**. 75–102.
Lin, Yu-Leng (2016). *Sonority effects and learning bias in nasal harmony*. PhD dissertation, University of Toronto.
Luke, Kang Kwong & May L. Y. Wong (2015). The Hong Kong Cantonese corpus: design and uses. *Journal of Chinese Linguistics Monograph Series* **25**. 312–333.
Maddieson, Ian & Kristin Precoda (1989). Updating UPSID. *The Journal of the Acoustical Society of America* **86**. S19–S19.
Matthews, Stephen & Virginia Yip (2011). *Cantonese: a comprehensive grammar* (2nd ed). London; New York: Routledge.
Mok, Peggy P. K., Donghui Zuo & 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**. 341–370.
Myers, James (2015). Markedness and lexical typicality in Mandarin acceptability judgments. *Language and Linguistics* **16**. 791–818.
Myers, James (2016). Meta-megastudies. *The Mental Lexicon* **11**. 329–349.
Myers, James & Jane Tsay (2005). The processing of phonological acceptability judgments. In *Proceedings of Symposium on 90–92 NSC Projects*. Taipei: National Science Council. 26–45.
Myers, Scott (1987). Vowel shortening in English. *Natural Language & Linguistic Theory* **5**. 485–518.
Nerbonne, John & Wilbert Heeringa (1997). Measuring dialect distance phonetically. In *Computational Phonology. Third Meeting of the ACL Special Interest Group in Computational Phonology*. Somerset, NJ: Association for Computational Linguistics. 11–18.
Nosofsky, Robert M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General* **115**. 39–57.
Nosofsky, Robert M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **14**. 700–708.
Ohala, John J. & Manjari Ohala (1986). Testing hypotheses regarding the psychological manifestation of morpheme structure constraints. In John J. Ohala & Jeri J. Jaeger (eds.) *Experimental Phonology*. Orlando: Academic Press. 239–252.
Olejarczuk, Paul & Vsevolod Kapatsinski (2018). The metrical parse is guided by gradient phonotactics. *Phonology* **35**. 367–405.
O'Seaghdha, Padraig G., Jenn-Yeu Chen & Train-Min Chen (2010). Proximate units in word production: phonological encoding begins with syllables in Mandarin Chinese but with segments in English. *Cognition* **115**. 282–302.
Ospina, Raydonal & Silvia L. P. Ferrari (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis* **56**. 1609–1623.
Qualtrics. (2020). *Qualtrics*. http://www.qualtrics.com/.
R Core Team. (2020). *R: a language and environment for Statistical Computing*. https://www.R-project.org/.

Richtsmeier, Peter T. (2011). Word-types, not word-tokens, facilitate extraction of phonotactic sequences by adults. *Laboratory Phonology* **2**. 157–183.
Ripley, Brian & William Venables (2016). *R package "nnet"* (version 7.3-12).
Roettger, Timo B. (2019). Researcher degrees of freedom in phonetic research. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* **10**. 1-27.
Rose, Sharon & Lisa King (2007). Speech error elicitation and co-occurrence restrictions in two Ethiopian Semitic languages. *Language and Speech* **50**. 451–504.
Schütze, Carson T. (1996). *The empirical base of linguistics: grammaticality judgments and linguistic methodology*. Chicago, Ill.: University of Chicago Press.
Shademan, Shabnam (2006). Is phonotactic knowledge grammatical knowledge? *WCCFL* **25**. 371-379.
Shoemark, Philippa (2013). *Cross-linguistic network structure effects on nonword acceptability judgements*. MA dissertation, The University of Edinburg. Paper presented at the 3rd International Conference of Undergraduate Linguistics Association of Britain.
Simmons, Joseph P., Leif D. Nelson & Uri Simonsohn (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* **22**. 1359–1366.
Steegen, Sara, Francis Tuerlinckx, Andrew Gelman & Wolf Vanpaemel (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science* **11**. 702–712.
Stokes, Stephanie F., Anita M-Y., Fletcher Paul & Laurence B. Leonard (2006). Nonword repetition and sentence repetition as clinical markers of specific language impairment: the case of Cantonese. *Journal of Speech, Language, and Hearing Research* **49**. 219–236.
Taft, Marcus & Hsuan-Chih Chen (1992). Judging homophony in Chinese: the influence of tones. *Advances in Psychology* **90**. 151–172.
Vehtari, Aki, Andrew Gelman & Jonah Gabry (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* **27**. 1413–1432.
Vitevitch, Michael S. & Paul A. Luce (1998). When words compete: levels of processing in perception of spoken words. *Psychological Science* **9**. 325–329.
Vitevitch, Michael S. & Paul A. Luce (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language* **40**. 374–408.
Vitevitch, Michael S. & Paul A. Luce (2004). A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, & Computers* **36**. 481–487.
Vitevitch, Michael S. & Paul A. Luce (2005). Increases in phonotactic probability facilitate spoken nonword repetition. *Journal of Memory and Language* **52**. 193–204.
Vitevitch, Michael S., Paul A. Luce, Jan Charles-Luce & David Kemmerer (1997). Phonotactics and syllable stress: implications for the processing of spoken nonsense words. *Language & Speech* **40**. 47–62.
Wiener, Seth & Rory Turnbull (2016). Constraints of tones, vowels and consonants on lexical selection in Mandarin Chinese. *Language and Speech* **59**. 59–82.
Xu, Yisheng, Jackson T. Gandour & Alexander L. Francis (2006). Effects of language experience and stimulus complexity on the categorical perception of pitch direction. *The Journal of the Acoustical Society of America* **120**. 1063–1074.

Yang, Cathryn & Andy Castro (2008). Representing tone in Levenshtein distance. *International Journal of Humanities & Arts Computing: A Journal of Digital Humanities* **2**. 205–219.
Yip, Moria. (1989). Cantonese morpheme structure and linear ordering. *WCCFL* **8**. 445–456.

Supplementary Materials A

List of Stimuli

No.	Stimuli string	Jyutping	IPA	No.	Stimuli string	Jyutping	IPA
1	mik2	mik2	mɪk˧	145	bom4	bom4	pɔːm˨
2	pYu3	pyuu3	pʰyːu˧	146	GOu5	gwoeu5	kʷœːu˩
3	Gek3	gwek3	kʷɛːk˧	147	hYk6	hyuk6	hyːk˨
4	juY3	juyu3	juːy˧	148	cOt4	ceot4	tsʰɵt˨
5	bai4	bai4	pɐi˨	149	zYu2	zyuu2	tsyːu˥
6	gAu5	gaau5	kaːu˩	150	wom5	wom5	wɔːm˩
7	kYi1	kyui1	kʰyːi˥	151	hAk2	haak2	haːk˧
8	lYk1	lyuk1	lyːk˥	152	fap4	fap4	fɐp̚˨
9	gOt2	geot2	kɵt˧	153	fYm1	fyum1	fyːm˥
10	GeY5	gweyu5	kʷey˩	154	cip4	cip4	tsʰiːp̚˨
11	cem5	cem5	tsʰɛːm˩	155	cOp6	ceop6	tsʰœːp̚˧
12	hem1	hem1	hɛːm˥	156	fei5	fei5	fei˩
13	ceY2	ceyu2	tsʰey˧	157	GOi6	gweoi6	kʷɵi˧
14	kOp2	keop2	kʰœːp̚˧	158	tAk6	taak6	tʰaːk˧
15	zek5	zek5	tsɛːk˩	159	dYi6	dyui6	tyːi˧
16	jou4	jou4	jou˨	160	kAp4	kaap4	kʰaːp̚˨
17	jeY3	jeyu3	jey˧	161	mOn5	meon5	mɵn˩
18	cei6	cei6	tsʰei˧	162	GAt1	gwaat1	kʷaːt̚˥
19	fen1	fen1	fɛːn˥	163	dOn4	deon4	tɵn˨
20	mAm1	maam1	maːm˥	164	mi06	mi6	miː˧
21	wet5	wet5	wɛːt̚˩	165	foi4	foi4	fɔːi˨
22	pOt1	peot1	pʰɵt̚˥	166	lY06	lyu6	lyː˧
23	hak3	hak3	hɐk̚˧	167	KAu4	kwaau4	kʷʰaːu˨
24	poi4	poi4	pʰɔːi˨	168	bak3	bak3	pɐk̚˧
25	KOY4	kweoyu4	kʷʰɵy˨	169	tik5	tik5	tʰɪk̚˩
26	GiY3	gwiyu3	kʷiːy˧	170	kek4	kek4	kʰɛːk̚˨
27	got2	got2	kɔːt̚˧	171	koY3	koyu3	kʰoy˧
28	fOt1	feot1	fɵt̚˥	172	pon3	pon3	pʰɔːn˧
29	wYN5	wyung5	wyː ŋ˩	173	jup3	jup3	juːp̚˧
30	weu1	weu1	wɛːu˥	174	cen1	cen1	tsʰɛːn˥
31	liY6	liyu6	liːy˨	175	seY3	seyu3	sey˧
32	hak4	hak4	hɐk̚˨	176	wYn2	wyun2	wyːn˧
33	sAp1	saap1	saːp̚˥	177	zYt4	zyut4	tsyːt̚˨
34	sai4	sai4	sɐi˨	178	KaN6	kwang6	kʷʰɐ ŋ˧
35	fOt3	feot3	fɵt̚˧	179	hYp5	hyup5	hyːp̚˩
36	heN3	heng3	hɛː ŋ˧	180	dum5	dum5	tuːm˩
37	jAm1	jaam1	jaːm˥	181	pet5	pet5	pʰɛːt̚˩
38	pYp6	pyup6	pʰyːp̚˧	182	jO02	joe2	jœː˧
39	tot5	tot5	tʰɔːt̚˩	183	cO05	coe5	tsʰœː˩
40	mYn2	myun2	myːn˧	184	bOY5	beoyu5	pɵy˩
41	GON6	gwoeng6	kʷœː ŋ˧	185	huk5	huk5	hʊk̚˩
42	dA04	daa4	taː˨	186	fip4	fip4	fiːp̚˨
43	gik3	gik3	kɪk̚˧	187	muY3	muyu3	muːy˧
44	lOn3	leon3	lɵn˧	188	jak4	jak4	jɐk̚˨
45	fYp3	fyup3	fyːp̚˧	189	wAt5	waat5	waːt̚˩

46	gAm6	gaam6	kaːm˨	190	suk2	suk2	sʊkˊ˧
47	bip5	bip5	piːp̄˧	191	za03	za3	tsɐ˨
48	cip6	cip6	tsʰiːp̄˨	192	KOY1	kweoyu1	kʷʰɵy˥
49	mYN4	myung4	myː ŋ˩	193	fok5	fok5	fɔːk̄˧
50	lan5	lan5	lɐn˧	194	fat4	fat4	fɐt̄˩
51	pe06	pe6	pʰɛːˊ	195	keN1	keng1	kʰɛː ŋ˥
52	dep4	dep4	tɛːp̄˩	196	dap4	dap4	tɐp̄˩
53	boY5	boyu5	poy˧	197	GYN1	gwyung1	kʷyː ŋ˥
54	wi05	wi5	wiː˧	198	moi3	moi3	mɔːiˊ
55	dAY3	daayu3	taːyˊ	199	pOt2	peot2	pʰɵt̄˧
56	zot1	zot1	tsɔːt̄˥	200	sOp1	seop1	sɵp̄˥
57	bum6	bum6	puːm˨	201	hOi2	heoi2	hɵi˧
58	tuk3	tuk3	tʰʊk̄˨	202	cum6	cum6	tsʰuːm˨
59	dou5	dou5	tou˧	203	Gek6	gwek6	kʷɛːk̄˨
60	tet5	tet5	tʰɛːt̄˧	204	wO05	woe5	wœː˧
61	deN4	deng4	tɛː ŋ˩	205	men3	men3	mɛːn˨
62	tuY4	tuyu4	tʰuːy˩	206	tei6	tei6	tʰeiˊ
63	gaN4	gang4	kɐ ŋ˩	207	ta01	ta1	tʰɐ˥
64	bu01	bu1	puː˥	208	dap3	dap3	tɐp̄˨
65	fik3	fik3	fɪk̄˨	209	ton3	ton3	tʰɔːn˨
66	gOt3	geot3	kɵt̄˨	210	fon3	fon3	fɔːn˨
67	KAN5	kwaang5	kʷʰaː ŋ˧	211	muN3	mung3	mʊ ŋ˨
68	GYp3	gwyup3	kʷyːp̄˨	212	cit6	cit6	tsʰiːt̄˩
69	soY2	soyu2	soy˥	213	dek1	dek1	tɛːk̄˥
70	cap5	cap5	tsʰɐp̄˧	214	gOn1	geon1	kɵn˥
71	GoY5	gwoyu5	kʷoy˧	215	jAu2	jaau2	jaːu˥
72	Kai3	kwai3	kʷʰɐiˊ	216	hYi5	hyui5	hyːi˧
73	bOu1	boeu1	pœːu˥	217	dOn5	deon5	tɵn˧
74	dAk1	daak1	taːk̄˥	218	bOp5	boep5	pœːp̄˧
75	jek2	jek2	jɛːk̄˥	219	cY01	cyu1	tsʰyː˥
76	mam2	mam2	mɐm˥	220	leY1	leyu1	ley˥
77	cu05	cu5	tsʰuː˧	221	bOn2	beon2	pɵn˥
78	sYu5	syuu5	syːu˧	222	Kan5	kwan5	kʷʰɐn˧
79	KAp4	kwaap4	kʷʰaːp̄˩	223	hYk3	hyuk3	hyːk̄˨
80	koN6	kong6	kʰɔː ŋ˨	224	jAu5	jaau5	jaːu˧
81	kom2	kom2	kʰɔːm˥	225	KOp2	kweop2	kʷʰœːp̄˥
82	get6	get6	kɛːt̄˩	226	kat5	kat5	kʰɐt̄˧
83	Ki01	kwi1	kʷʰiː˥	227	hYp2	hyup2	hyːp̄˥
84	sen3	sen3	sɛːn˨	228	pAY4	paayu4	pʰaːy˩
85	tek2	tek2	tʰɛːk̄˥	229	bYu3	byuu3	pyːu˨
86	faN6	fang6	fɐ ŋ˨	230	Kim4	kwim4	kʷʰiːm˩
87	pik6	pik6	pʰɪk̄˨	231	sOp6	seop6	sɵp̄˨
88	pAu5	paau5	pʰaːu˧	232	sO02	soe2	sœː˥
89	GOi1	gweoi1	kʷɵi˥	233	Gen5	gwen5	kʷɛːn˧
90	heN2	heng2	hɛː ŋ˥	234	wAu2	waau2	waːu˥
91	fiY5	fiyu5	fiːy˧	235	wOi3	weoi3	wɵiˊ
92	bem1	bem1	pɛːm˥	236	dek5	dek5	tɛːk̄˧
93	ceu5	ceu5	tsʰɛːu˧	237	sui3	sui3	suːiˊ
94	jO03	joe3	jœːˊ	238	zem3	zem3	tsɛːmˊ
95	fYu4	fyuu4		239	bYt2	byut2	

96	lOu3	loeu3	fy:u꜖	240	mip6	mip6	py:t̄˥
97	pek4	pek4	lœ:u˧	241	GOm2	gweom2	mi:p̄˧
98	pO04	poe4	pʰɛ:k̄꜖	242	got4	got4	kʷœ:m˥
99	fiY6	fiyu6	pʰœ:꜖	243	bom2	bom2	kɔ:t̄꜖
100	tot2	tot2	fi:y˧	244	hYm1	hyum1	pɔ:m˥
101	meu4	meu4	tʰɔ:t̄˥	245	Kok4	kwok4	hy:m̄
102	lik3	lik3	mɛ:u꜖	246	GAn2	gwaan2	kʷʰɔ:k̄꜖
103	tat6	tat6	lɪk̄˧	247	Gim3	gwim3	kʷa:n˥
104	dO03	doe3	tʰɐt̄˧	248	dYp1	dyup1	kʷi:m˧
105	bOi1	beoi1	tœ:˧	249	dei4	dei4	ty:p̄˥
106	mOk4	moek4	pɐi˥	250	wO02	woe2	tei˩
107	gun6	gun6	mœ:k̄꜖	251	zan5	zan5	wœ:˥
108	pek5	pek5	ku:n˧	252	tap4	tap4	tsɐn˩
109	Gon1	gwon1	pʰɛ:k̄˩	253	sAu5	saau5	tʰɐp̄꜖
110	gik4	gik4	kʷɔ:n˥	254	gYt4	gyut4	sa:u˩
111	jom5	jom5	kɪk̄꜖	255	wuk4	wuk4	ky:t̄꜖
112	bot3	bot3	jɔ:m˩	256	cuY6	cuyu6	wʊk̄꜖
113	bAp2	baap2	pɔ:t̄˧	257	tuY1	tuyu1	tsʰu:y˧
114	pak2	pak2	pa:p̄˥	258	pOY1	peoyu1	tʰu:y˥
115	wok4	wok4	pʰɐk̄˥	259	tem2	tem2	pʰɵy˥
116	KYi2	kwyui2	wɔ:k̄꜖	260	tem5	tem5	tʰɛ:m˥
117	fAu5	faau5	kʷʰy:i˥	261	wiY6	wiyu6	tʰɛ:m˩
118	KOi4	kweoi4	fa:u˩	262	jOu2	joeu2	wi:y˧
119	ka01	ka1	kʷʰɵi꜖	263	kOm3	keom3	jœ:u˥
120	hot2	hot2	kʰɐ˥	264	pen5	pen5	kʰœ:m˧
121	koY2	koyu2	hɔ:t̄˥	265	dYp6	dyup6	pʰɛ:n˩
122	pup4	pup4	kʰoy˥	266	Gen3	gwen3	ty:p̄˧
123	wi03	wi3	pʰu:p̄꜖	267	pO02	poe2	kʷɛ:n˧
124	wAu5	waau5	wi:˧	268	tan5	tan5	pʰœ:˥
125	laY4	layu4	wa:u˩	269	kAn3	kaan3	tʰɐn˩
126	zot4	zot4	lɐy꜖	270	feY4	feyu4	kʰa:n˧
127	joY2	joyu2	tsɔ:t̄꜖	271	Gak4	gwak4	fey꜖
128	zun1	zun1	joy˥	272	baN3	bang3	kʷɐk̄꜖
129	wOY2	weoyu2	tsu:n˥	273	pip6	pip6	pɐ ŋ ˧
130	fAm4	faam4	wɵy˥	274	tOn6	teon6	pʰi:p̄˧
131	jop3	jop3	fa:m˩	275	bap1	bap1	tʰɵn˧
132	fYu5	fyuu5	jɔ:p̄˧	276	mit2	mit2	pɐp̄˥
133	Kin4	kwin4	fy:u˩	277	Gom4	gwom4	mi:t̄˥
134	bYu1	byuu1	kʷʰi:n꜖	278	jak1	jak1	kʷɔ:m꜖
135	hAu5	haau5	py:u˥	279	kaY3	kayu3	jɐk̄˥
136	jok3	jok3	ha:u˩	280	ket3	ket3	kʰɐy˧
137	kap3	kap3	jɔ:k̄˧	281	KAt3	kwaat3	kʰɛ:t̄˧
138	gAt4	gaat4	kʰɐp̄˧	282	set2	set2	kʷʰa:t̄˧
139	koi2	koi2	ka:t꜖	283	maY5	mayu5	sɛ:t̄˥
140	dam4	dam4	kʰɔ:i˥	284	zYk3	zyuk3	mɐy˩
141	ki04	ki4	tɐm꜖	285	tan2	tan2	tsy:k̄˧
142	poN1	pong1	kʰi:˩	286	tOi2	teoi2	tʰɐn˥
143	cut1	cut1	pʰɔ: ŋ ˥	287	tum4	tum4	tʰɵi˥
144	sYu3	syuu3	tsʰu:t̄˥	288	pot3	pot3	tʰu:m꜖
			sy:u˧				pʰɔ:t̄˧

Supplementary Materials B
Demographic Questionnaire

1. 您嘅性別係？What is your gender?
   a. 男 Male
   b. 女 Female
   c. 其他 Other

2. 您屬於以下邊一個年齡層？Which age group are you belong to?
   a. 18–24
   b. 25–34
   c. 35–44
   d. 45–54
   e. 55–60

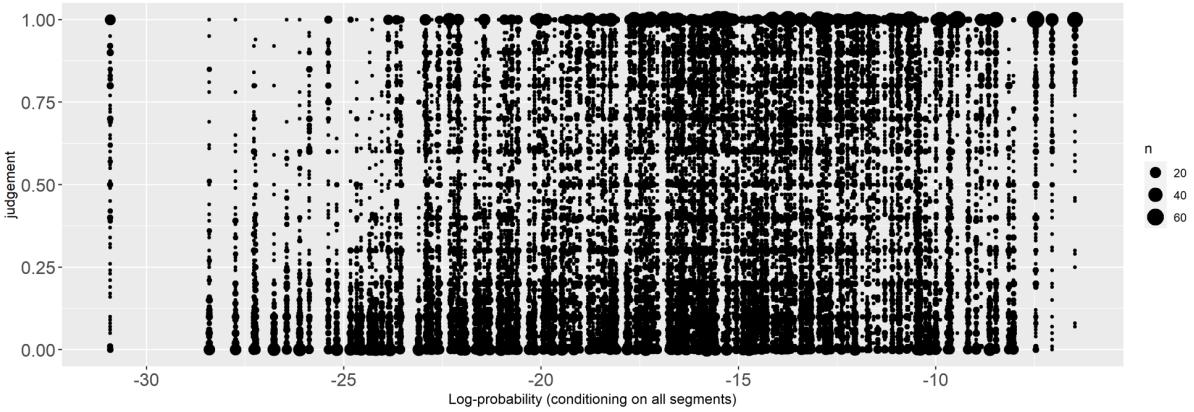
3. 您嘅母語係？（可選擇多於一項）
   What is(are) your mother language(s)? (More than one option can be selected)
   a. 廣東話 Cantonese
   b. 普通話/國語 Putonghua/Taiwanese Mandarin
   c. 英文 English
   d. 其他語言/方言，請註明：
      Other language/dialect, please indicate:

4. 喺十歲之前，您主要喺邊個地方生活？
   Where did you mainly live before the age of 10?
   a. 香港 Hong Kong
   b. 澳門 Macau
   c. 內地，請註明城市：
      Mainland, please indicate city:
   d. 其他，請註明國家/地區：
      Other, please indicate country/city:

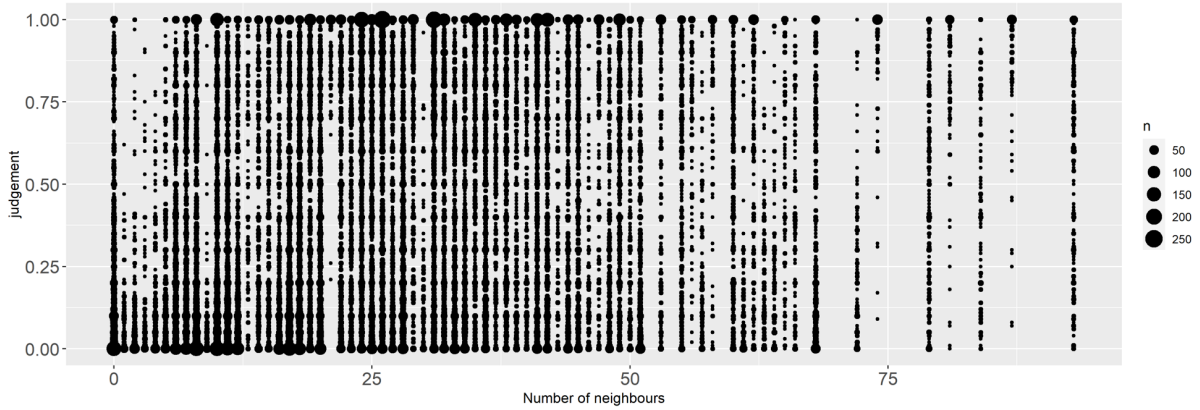
5. 您現居嘅城市係？
   Which country are you living in?
   a. 香港 Hong Kong
   b. 澳門 Macau
   c. 內地，請註明城市：
      Mainland, please indicate city:
   d. 其他，請註明國家/地區：
      Other, please indicate country/city:

Supplementary Materials C
Scatterplots of predictors against wordlikeliness judgements

In the figures below, size of the circles are proportional to the number of judgements.

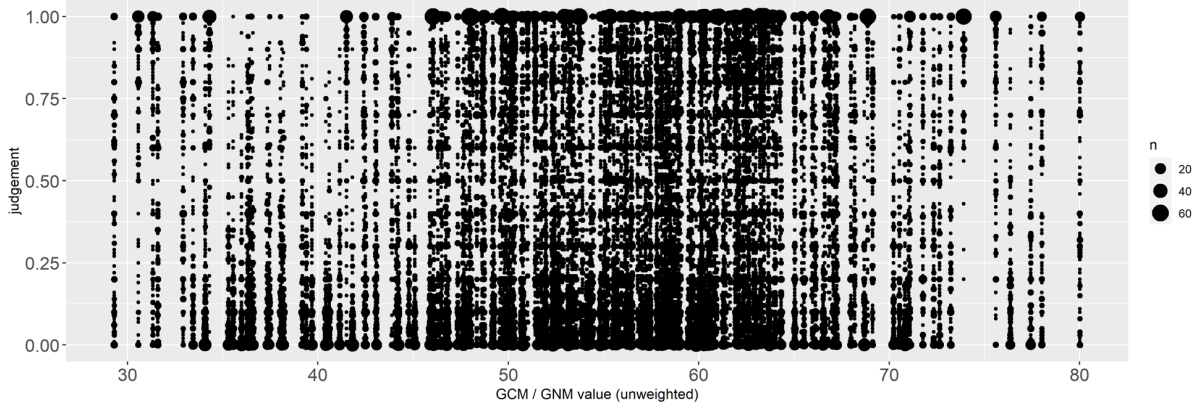


(a) Scatterplot of judged wordlikeness (*y*-axis) against the log-probability

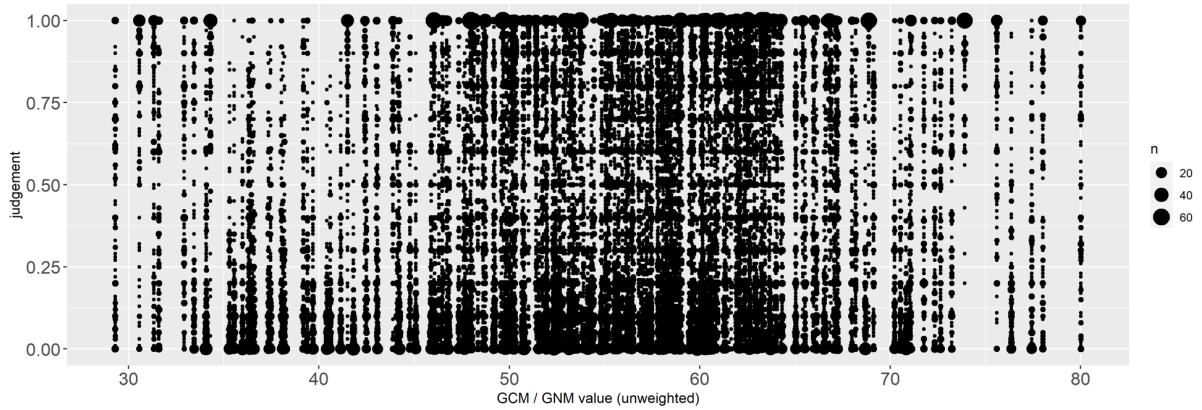


(b)

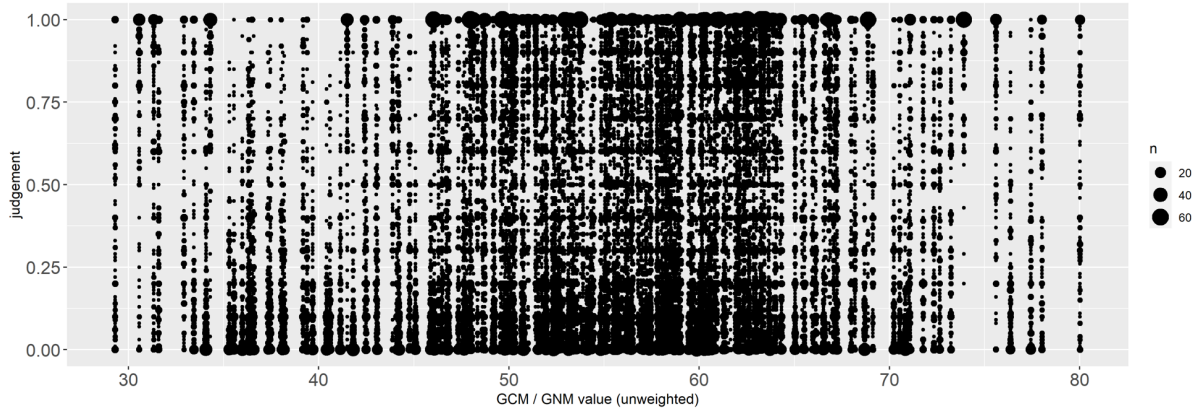
(c) Scatterplot of the judged wordlikeness (*y*-axis) against the number of neighbors (*x*-axis)



(d) Scatterplot of the judged wordlikeness (*y*-axis) against the the GCM value (*x*-axis), i.e. the third GNM quality insensitive to frequency



(e) Scatterplot of the judged wordlikeness (*y*-axis) against the second GNM quantity (*x*-axis), i.e. GCM weighted by frequency (with B as a coefficient).



(e) Scatterplot of the judged wordlikeness (*y*-axis) against the first GNM quantity(*x*-axis), i.e. GCM weighted by square frequency (with A as a coefficient).

Supplementary Materials D

Basics of ZOIB models

The ZOIB model has three components: A Bernoulli-distributed (i.e. discrete probability distribution) component for predicting whether the judgement is zero (absolutely impossible), another Bernoulli-distributed component for predicting whether the judgment is one (absolutely possible), and a beta-distributed (i.e. continuous probability distribution) component for modelling the density of the intermediate judgements (between 0 and 1). The three components' distributions are given below:

$$(1)\ I(Y_{ij}=0) \sim Bernoulli\left(logit^{-1}(\beta_{00}+(\beta_{01}+\alpha_{01i})x_{lp,j}+\alpha_{00i}+\gamma_{0j})\right)$$
$$I(Y_{ij}=1) \sim Bernoulli\left(logit^{-1}(\beta_{10}+(\beta_{11}+\alpha_{11i})x_{lp,j}+\alpha_{10i}+\gamma_{1j})\right)$$
$$Y_{ij} \mid Y_{ij} \in (0,1) \sim Beta\left(\phi\ logit^{-1}\left(\beta_{20}+(\beta_{21}+\alpha_{21i})x_{lp,j}+\alpha_{20i}+\gamma_{2j}\right), \phi(1 - logit^{-1}(\beta_{20}+(\beta_{21}+\alpha_{21i})x_{lp,j}+\alpha_{20i}+\gamma_{1j})\right)$$

In the above formula, the means of the two Bernoulli distributions (0s and 1s) and the beta distribution (gradient judgments) depend on the same set of predictors, in this case the log-probability ($x_{lp,j}$). There are two population-level coefficients ('fixed effects' in frequentist terms) for each of the three parts of the model, namely the population-level intercept $\beta_{00}, \beta_{10}, \beta_{20}$ and the population-level slopes $\beta_{01}, \beta_{11}, \beta_{21}$. There are also participant-level predictors ('random effects' in frequentist terms) that allow for variability across participants, including the three random intercepts α_{00i}, α_{10i} and α_{20i}, and the three random slopes α_{01i}, α_{11i} and α_{21i}. Finally, there is an item-level intercept.

The means of the two Bernoulli distributions are related to the probability of through a logit link, as is the case for standard logistic regression. For the beta regression, the formula shown here is derived from a reparametrisation of the beta regression in terms of the mean and a precision parameter ϕ.

We will now look at the distributions of the model parameters in detail. Firstly, the group-level effects for each component come from bivariate normal distributions. The covariance matrix allows for correlations. There is an LKJ prior with one degree of freedom (Lewandowski *et al*., 2009) on the lower Cholesky decomposition of the correlation matrix, and half-*t* priors (Gelman, 2006) on the standard deviations:

$$(\alpha_{c0i}, \alpha_{c1i}) \sim N(0, \Sigma_{ac})\ for\ c \in \{0,1,2\}, i \in \{1,2,\dots I\}$$
$$\text{where } \Sigma_{ac} = D_{ac}R_{ac}D_{ac}, R_{ac} = L_{ac}L_{ac}^{T}, D_{ac} = diag(\sigma_{ac1}, \sigma_{ac2}),$$
$$L_{ac} \sim LKJ(1), \sigma_{ac1}, \sigma_{ac2} \sim half-t(3,0,2.5)$$

The item-level intercept simply follows a univariate normal distribution, again with a half-*t* prior on its standard deviation:

$$\gamma_{0j} \sim N(0, \sigma_{\gamma c})\ for\ c \in \{0,1,2\}, i \in \{1,2,\dots I\}, \sigma_{\gamma c} \sim half-t(3,0,2.5)$$

There is a default standard normal prior on the 'fixed-effect' slopes, a *t*-distributed prior on the population-level intercept for the beta component, and a logistic-distributed prior on the population-level intercept for the logistic components:

$$\beta_{c1} \sim N(0,1)\ for\ c \in \{0,1,2\}$$
$$\beta_{01}, \beta_{02} \sim Logistic(0,1)$$
$$\beta_{00} \sim t(3,0,2.5)$$

Finally, there is a gamma prior on the precision parameter of the beta distribution:

$$\phi \sim \Gamma(0.01, 0.01)$$

Supplementary Materials E

Confidence intervals for the multiverse analysis

	T|O	T|N	T|C	T|S	T	NoT
NN	**(0.03, 0.08)**	**(0.05, 0.11)**	**(0.03, 0.08)**	**(0.02, 0.07)**	**(0.06, 0.12)**	**(0.03, 0.08)**
GCM	**(0.05, 0.08)**	**(0.06, 0.09)**	**(0.04, 0.08)**	**(0.04, 0.07)**	**(0.06, 0.10)**	**(0.06, 0.09)**
None	**(0.04, 0.08)**	**(0.05, 0.08)**	**(0.05, 0.08)**	**(0.06, 0.10)**	**(0.06, 0.10)**	**(0.06, 0.10)**

Table 8a. Multiverse results for the 95% CI of the effect of log-probabilities on gradient judgements.

	T|O	T|N	T|C	T|S	T	NoT
NN	(−0.03, 0.04)	(−0.03, 0.04)	**(0.00, 0.07)**	(−0.02, 0.04)	(−0.03, 0.05)	**(0.00, 0.07)**
GCM	(−0.02, 0.03)	(−0.02, 0.03)	(−0.01, 0.04)	(−0.01, 0.03)	(−0.02, 0.03)	(−0.02, 0.03)
None	(−0.01, 0.03)	(−0.02, 0.03)	(−0.00, 0.04)	(−0.02, 0.03)	(−0.02, 0.03)	(−0.02, 0.03)

Table 8b. Multiverse results for the 95% CI of the effect of log-probabilities on the judgments of 0s.

	T|O	T|N	T|C	T|S	T	NoT
NN	**(0.13, 0.44)**	**(0.30, 0.64)**	**(0.13, 0.45)**	**(0.06, 0.35)**	**(0.37, 0.73)**	**(0.13, 0.45)**
GCM	**(0.27, 0.50)**	**(0.35, 0.58)**	**(0.25, 0.47)**	**(0.21, 0.43)**	**(0.37, 0.61)**	**(0.36, 0.59)**
None	**(0.28, 0.51)**	**(0.28, 0.51)**	**(0.27, 0.48)**	**(0.37, 0.60)**	**(0.38, 0.62)**	**(0.38, 0.62)**

Table 8c. Multiverse results for the 95% CI of the effect of log-probabilities on the judgments of 1s.

	T|O	T|N	T|C	T|S	T	NoT	None
NN	(−0.00,0.01)	(−0.01, 0.00)	(−0.00, 0.01)	(−0.00, 0.01)	(−0.01, 0.00)	(−0.00, 0.01)	(−0.01, 0.00)
GCM	**(0.01**, **0.10)**	**(0.01**, **0.10)**	(−0.00, 0.01)	**(0.00**, **0.02)**	**(0.00**, **0.02)**	**(0.00**, **0.02)**	**(0.00**, **0.02)**

Table 9a. Multiverse results for the 95% CI of the effect of neighborhood density on gradient judgements

	T|O	T|N	T|C	T|S	T	NoT	None
NN	(−0.01,0.01)	(−0.01, 0.01)	(−0.01, 0.00)	(−0.01, 0.01)	(−0.01, 0.01)	(−0.01, 0.00)	(−0.01, 0.01)
GCM	(−0.01,0.02)	(−0.01, 0.02)	(−0.01, 0.01)	(−0.01, 0.01)	(−0.01,0.02)	(−0.01, 0.02)	(−0.01, 0.01)

Table 9b. Multiverse results for the 95% CI of the effect of neighborhood density on the judgments of 0s.

	T|O	T|N	T|C	T|S	T	NoT	None
NN	**(0.00, 0.07)**	(−0.04, 0.04)	(−0.01, 0.07)	**(0.01, 0.08)**	(−0.05,0.02)	(−0.01, 0.07)	(−0.05, 0.02)
GCM	**(0.02, 0.11)**	**(0.02, 0.11)**	(−0.01, 0.09)	**(0.01, 0.11)**	**(0.01, 0.10)**	**(0.02, 0.11)**	**(0.02, 0.13)**

Table 9c. Multiverse results for the 95% CI of the effect of neighborhood density on the judgments of 1s.