DRAFT: Perceptual effects of lexical competition on Cantonese tone categories

Rachel Soo and Molly Babel

Department of Linguistics, University of British Columbia

Abstract

In the face of substantial acoustic-phonetic variation in natural speech, listeners use lexical information along with the speech signal to categorize sounds and recognize words. In instances where systematic variation acts to neutralize lexical contrasts, as in diachronic mergers, the role of the lexicon becomes less clear. To this end, we examine how lexical competition structures phonetic variability in lexical tone categories in three experiments examining a series of ongoing tone mergers in Cantonese. Experiment 1 is a categorization task in which listeners were presented with tokens from lexical tone continua generated from minimal pairs, while Experiment 2 is a lexical decision task in which listeners categorized tokens from tone continua generated from word-nonword pairs. The presence of a lexical competitor at both continuum ends in Experiment 1 maintained more discrete categorization functions for non-merging tone pairs than in Experiment 2 where only one endpoint was a word. In the merging tone pairs however, overall categorization was less discrete and the effect of lexical competition was not observable. Preliminary data from a goodness rating task, Experiment 3, suggest that lexical competition affects internal category structure for merging tones, but not non-merging tones. Overall, these data provide evidence that tone mergers impact phonetic categories and the lexicon, in addition to showing that, for non-merging tones, the range of acceptable phonetic variation is constrained by the presence of a lexical competitor.

Keywords: Lexical competition, tone mergers, Cantonese, speech perception

DRAFT: Perceptual effects of lexical competition on Cantonese tone categories

1 Introduction

1.1 Lexical influence in phonetic encoding

Everyday speech perception requires listeners to process speech in acoustically adverse conditions. The lexicon appears to play a role in allowing listeners to achieve this task (Connine and Clifton, 1987; Marslen-Wilson, 1984; McClelland and Elman, 1986; Samuel, 1996; 1997; 2001). Early work by Ganong (1980) established that lexical context serves to disambiguate ambiguous speech sounds. That is, an ambiguous sound midway between a /t/ and /d/ was more likely to be identified as /t/ in a context like "?ask", and /d/ in a context like "?ash", because "task" is a real word (while *"dask" is not), and "dash" is a real word (while *"tash" is not). Similar Ganong-type lexical bias effects have been observed with different phoneme pairs (Connine and Clifton, 1987; Pitt, 1995), in different word positions (Pitt and Samuel, 1993; 1995) and with words of varying lengths (Pitt and Samuel, 2006). In the current study, we examine how the lexicon may contribute to the perceptual category structure of lexical tones in Cantonese in a series of experiments.

Although there is substantial debate regarding the mechanism that allows for lexical information to shape perceptual decisions like the Ganong effect and similar lexical bias behaviours (e.g., Norris, McQueen, and Cutler, 2000; Samuel, 2000), most models of speech perception integrate a lexical mechanism at some point during processing (Samuel, 2011). The influence of the lexicon on speech perception has been observed in a number of experimental paradigms. In phoneme restoration studies for example, listeners report hearing a word as intact despite the presence of noise obscuring particular sounds in the word (Warren, 1970). This "restoration effect" is stronger in words than in nonwords, suggesting that phonetic encoding is supported by the lexicon (Samuel, 1981; 1996). Likewise, in phoneme monitoring tasks, target phonemes are identified faster in words than in nonwords (Rubin, Turvey, and van Gelder, 1976). Furthermore, nonword processing has been shown to vary as a function of its similarity to real words (Connine, Titone, Deelman, and Blasko, 1997; Wurm and Samuel, 1997). Lexical knowledge has also been shown to guide the retuning of phonetic category boundaries in perceptual learning paradigms (e.g., Norris, McQueen, & Cutler, 2003). Finally, in studies examining bilingual populations, lexical support has been shown to vary as a function of L2 proficiency (Samuel and Frost, 2015; Soo et al., 2020). Specifically, since lexical support for phonetic encoding is only present insofar as listeners have developed fully functional lexical representations, a robust lexicon is a clear hallmark of successful speech perception in the language.

1.2 Lexical competition

On a broad level, the body of literature reviewed above shows that listeners are sensitive to the lexical items that do and do not exist in their lexicon as they are able to use this knowledge to guide the perception and learning of ambiguous speech sounds. At the same time, listeners show more subtle, fine-tuned knowledge of their lexicon and exploit it in production. For instance, native listeners are sensitive to the density of the phonological neighbours of specific words. Vowels in words from dense phonological neighborhoods exhibit more reduction in spontaneous speech (Gahl, Yao, & Johnson, 2012). This finding has been attributed to online retrieval mechanisms; however, in single word production, vowels in dense lexical neighborhoods exhibit expansion, suggesting that talkers have implicit knowledge of when and how to make sounds, and consequently, words, more distinct (Munson & Solomon, 2004; Wright, 1997, 2004).

Listeners also appear to have implicit knowledge of the specific lexical competitors for a word along a given acoustic dimension in production. Baese-Berk and Goldrick (2009) tested the degree to which lexical competition affects stop voicing contrasts in a production study using words with a voiced stop competitor (e.g., "pox", whose competitor is"box") and those without a voiced stop competitor (e.g., "posh", whose competitor is *"bosh"). Words with a voiced stop lexical competitor were produced with longer voice onset time

(VOT) than those without, suggesting that listeners are implicitly aware of the lexical competitors for a given word and exploit this in production for the purposes of contrast enhancement. These effects were replicated with word-initial alveolar /t/ and velar /k/ stops. [1]

The effects of lexical competition also extend to suprasegmental phenomena. In Yang et al. (2019), native Mandarin speakers performed a two-alternative forced choice task on CV syllables created from PSOLA (Charpentier, 1986) and TANDEM-STRAIGHT (Kawahara, 2008) synthesized lexical tone continua. The two tone contrasts that were tested included high-level tone (T) 1 versus falling T4, and T1 versus rising T2. For each tone contrast, three types of pairings/continua were created: (1) one end was a real word, while the other was an accidental gap, (2) both ends were real words, or (3) both ends were accidental gaps. Listeners were significantly biased towards the real word endpoints for the T1-T4 continua synthesized in TANDEM-STRAIGHT and marginally so for T1-T2 continua. No significant effects were observed for the continua synthesized using PSOLA. While this bias was also observed in pairs where both ends were words, and in pairs where both ends were gaps, the authors attribute this to the selected endpoints, which were imbalanced in terms of word and tone token frequency.

[1] Goldrick, Vaughn, and Murphy (2013) examined these effects on the English voiced stop series word-initially and word-finally, observing different results. In word-initial position, unlike in Baese-Berk and Goldrick (2009), there were no differences in the positive VOT between items with and without lexical competitors. In word-final position, there were clear differences in the vowel durations of voiced stops as a function of lexical competition, but not in the expected direction. The preceding vowels of words with a competitor (e.g., “bud”, whose competitor is “but”) were produced as shorter than those without minimal pairs (e.g., “thud”, whose competitor is *“thut”) acting to reduce the voicing contrast for stops in word-final position. Goldrick and colleagues speculate that these differences across phonological contrasts and word positions are a consequence of the lexicon interacting with phonetic and phonological restrictions that constrain the range of allowable variation.

1.3 Maintaining recognition in sound change

Apart from contrast enhancement (Baese-Berk and Goldrick (2009)~~)~~, lexical competition also appears to play a role in contrast maintenance for sound changes. Specifically, minimal pairs have been shown to inhibit diachronic mergers. ~~A.~~ Wedel, Kaplan, and Jackson (2013) carried out a corpus study on a series of phoneme mergers from a diverse range of languages. By calculating the functional load for the phonemes taking part in mergers across these languages, they found an inverse relationship between the probability of a phoneme merger and functional load. That is, the probability of a merger was less likely between phonemes that distinguish a greater number of words in the language (i.e., when there are more minimal pairs).

The notion that gradual phonetic change in a diachronic merger may be inhibited for the purpose of maintaining word recognition is also observed in cases of sound change where there is no neutralization of contrasts. In push chains, for example, the gradual phonetic change of one category pushes it towards another category. In turn, that category moves in an effort to maintain a contrast. Hay, Pierrehumbert, Walker, and LaShell (2015) examined frequency effects on a push chain shift in New Zealand English involving DRESS, TRAP, and KIT words. Using corpus data that span 136 years, they established that low frequency words changed faster than high frequency words and reason that lower frequency items move first as a way to resist ambiguity and promote comprehension. This is supported by a computational model in Todd, Pierrehumbert, and Hay (2019) who found that high and low frequency words changed at a similar rate in the context of change that does not threaten a contrast.

These results are also supported by computational models, which formalize contrast maintenance as a function of category competition (~~A.~~ Wedel, 2012; ~~A.~~ Wedel & Fatkullin, 2017; ~~A. B.~~ Wedel, 2006; Winter & Wedel, 2016). In ~~A.~~ Wedel and Fatkullin (2017), categories consist of a series of exemplars. These categories are updated as new percepts are assigned to the category that contains exemplars in a perceptual space similar to that of the incoming percept. In turn, this newly mapped exemplar allows new incoming percepts occupying a similar perceptual space to be even more likely to be mapped to that category. This produces a positive feedback loop between production and perception driven by lexical competition, which ultimately acts to sharpen category boundaries. Thus, mergers may be represented as cases where the distribution of two sound categories begin to approach one another and produce an area of overlap at category boundaries. In these cases, it is the presence of a lexical competitor that drives the distributions to compete for the incoming percept and ultimately maintain independent, non-overlapping portions at the opposing extreme ends of the distribution. In the absence of a lexical competitor however, listeners are able to map the percept to the appropriate category simply by knowing if the resulting sound would produce a real word in the language. As such, the absence of a lexical competitor produces less competition between categories, and less category contrast.

The lexicality of a given token is often clear in languages with complex syllable templates. For example, although English permits complex onsets and codas, and thus has rather large number of logically possible monosyllabic words, the proportion of lexical items that actually occupy this phonotactic space is small (Kirby and Yu, 2007). As such, it is fairly straightforward to assess the lexicality of a particular token, and to that end, the lexical competitors that may exist for a given word. Indeed in Kirby and Yu (2007), wordlikeliness ratings were found to be positively correlated with measures of phonotactic probability. Languages with comparatively more restricted templates may utilize lexical tones to broaden the range of possible lexical items in the language. In Cantonese, for instance, syllables are maximally (C)V(N), but the language has six lexical tones that are distinguished along three dimensions: pitch height, contour and magnitude of change (Bauer and Benedict, 1997; Fok-Chan and Yuen, 1974; Gandour, 1981; Khouw and Ciocca, 2007). Since a larger proportion of all possible Cantonese monosyllables are real words compared to English, it may be more difficult for listeners to ascertain whether a particular word bears a tonal lexical competitor or not. From this perspective, the contribution of lexical competition in maintaining sound contrasts may be smaller in cases of phonetic variation involving lexical tones, as compared to segments. What's more, several of the tones in Cantonese are undergoing a merger in progress. Sound changes involving lexical tones add an additional layer of complexity as the contrasts being neutralized may involve a number of acoustic cues across a comparatively longer period of time. For instance, the phonetic differences in VOT for stop voicing contrasts or spectral properties for fricative contrasts are localized to a relatively small stretch of time. On the other hand, lexical tone contrasts may differ along a number of phonetic cues across the entire tone bearing syllable (Gandour, 1981; 1983; Khouw and Ciocca, 2007).

In the current study, we take up the question of how lexical competition shapes perceptual category structure of Cantonese lexical tones. Since tone minimal pairs occupy a large proportion of the lexical space in Cantonese, the relative contribution of lexical competition in structuring sound changes may be different from that of previous studies investigating segmental sound changes. To this end, we assess listener categorization of merging and non-merging tone pairs with and without lexical competitors in two experiments. Experiment 1 was a word identification task utilizing tone pairs with lexical competitors (i.e., tone minimal pairs) and Experiment 2 was a lexical decision task utilizing tone pairs without lexical competitors (i.e., word-nonword pairs). Overall, the tone categories in non-merging tone pairs were expected to be more perceptually distinct, showing steeper categorization functions than merging tone pairs. Moreover, in line with the literature on sound change (Todd et al., 2019; ~~A.~~ Wedel, Jackson, & Kaplan, 2013), we predicted that lexical competition would result in the maintenance of more distinct tone categories, manifested with steeper categorization functions in items with lexical competitors than those without. This effect of lexical competition was also predicted to be stronger in non-merging tone pairs than in merging tone pairs, since the distinction between words is already in the process of being neutralized in merging tone pairs.

1.4 Cantonese tone mergers

Cantonese is a Sino-Tibetan language spoken primarily in Hong Kong, Macau, and Guangzhou (Eberhard et al., 2020). Hong Kong Cantonese (the focus of the current study) contains six lexical tones on open syllables with optional nasal codas (Bauer and Benedict, 1997; Matthews and Yip, 2013). These six lexical tones consist of three level tones (high-level: 1, mid-level: 3, low-level: 6), two rising tones (high-rising: 2, mid-rising: 5), and one falling tone (low-falling: 4). This tone inventory has been provided in Table 1 and visually depicted in Figure 1.

Contour	Tone	Description	Example word
Level	1 [55]	High-level	衣 ji1 clothes
	3 [33]	Mid-level	意 ji3 idea
	6 [22]	Low-level	二 ji6 two
Rising	2 [25]	High-rising	椅 ji2 chair
	5 [23]	Mid-rising	耳 ji5 ear
Falling	4 [21]	Low-falling	疑 ji4 suspicious

Table 1

Cantonese phonemic tone inventory. Tone numerals following Bauer and Benedict (1997) are given in square brackets to characterize the pitch contour. In the Jyutping transcription tones are represented with numbers.

Several of these lexical tones been reported to be merging. The literature documents ongoing mergers between T2 and T5, T3 and T6, T4 and T6, and T3 and T5, in both production and perception (Bauer et al., 2003; Fung and Lee, 2019; Mok et al., 2013; Lam, 2018; Lee et al., 2015; Tsui, 2012; Vance 1976, Wong 2008).

The T2-T5 merger is said to be nearly complete (Fung and Lee, 2019). Indeed, the merger between T2-T5 was the first one to be observed among the three attested mergers in the literature thus far. In a production study by Kei et al. in 2002, six out of fifteen participants in their sample were unable to distinguish T2-T5. The authors interpreted these results as production errors, though, as more work was carried out, this non-distinction of T2-T5 was corroborated by other production (Bauer et al., 2003; Yiu, 2009) and perception studies (Lee et al., 2015; Mok et al., 2003; Yiu, 2009). In addition, this merger seems to be progressing in the direction of T2. Data from confusion matrices of a word identification paradigm in Lam (2018) show that Cantonese listeners are more likely to hear T5 as T2, than T2 as T5.

The merger between T3 and T6 in particular has been described as "partial" with significant variation across individuals (Fung and Lee, 2019; Lam, 2018; Mok et al., 2013). Confusion matrices in Lam (2018) revealed no consistent patterns in the confusion of T3 as T6 or T6 as T3. More generally, many of these studies report conflicting evidence about how this merger is progressing in production versus perception. For instance, Mok et al. (2013) report that while the acoustic distances between T3 and T6 is smaller in native speaker productions, they are nonetheless able to maintain distinct tone categories. On the other hand, Fung and Lee (2019) observed that native Cantonese speakers were no longer distinguishing T3 and T6 in production, though they had high accuracy scores when they were asked to discriminate these tones in an AX task. In contrast, Soo and Monahan (2017) also carried out an AX task on Cantonese tones. Using d' scores to quantify accuracy, very low discriminability for T3 and T6 was observed. Finally, Peng and Wang (2015) report that T3 and T6 were most likely to be confused in perception by both native speakers and a support vector machine trained to recognize tones.

The T4-T6 merger patterns more consistently than the T3-T6 merger, with behavioural and ERP studies finding that the contrast between these tones has been neutralized in perception, though maintained in production, making it a near-merger (Fung et al., 2012; Fung and Lee, 2019; Mok et al., 2013). As mentioned above, acoustic similarity across the f0 contour of a tone may be a driving force for the merger. Since T4 is the only tone in Cantonese produced with creakiness, many have speculated that the voice quality is what supports its continued distinction from T6 in production.

Since T2, T3, T4, T5, and T6 are all part of one or more mergers-in-progress, T1 represents the only tone that is not merging in the language. As a result, some have speculated that the crowded tone space of Cantonese is responsible for the observed tone mergers (Mok et al., 2013; Varley and So, 1995). Four tones start at a pitch level of [2] (T6, T2, T5, T4) and four tones end at a low to mid pitch range [1-3] (T3, T6, T5, T4; Bauer and Benedict, 1997). In addition to start and end pitch heights, many of these tones share a contour shape. T2 and T5 are both rising tones, and T3 and T6 are both level tones. Thus, their similarity in shape across time might also be a driving force for these mergers (Soo and Monahan, 2017; Varley and So, 1995). Tsui (2012) likewise suggests that acoustic similarity plays a role in the mergers, but additionally provides evidence for the role of functional load in these mergers. This was particularly evident for T3-T5 and T2-T5, which are antepenultimate and penultimate in the ranking of tone pairs by functional load. Mok et al. (2013) also proposed that type frequency may play a role in the directionality of the mergers. In their production study, tones of lower type frequency demonstrating greater variability and merging in the direction of the corresponding tone of the merging pair.

2 Methodology

2.1 Materials

Items for Experiment 1 consisted of four tone minimal pairs selected for merging tone pairs 2-5, 3-6 and 4-6, and 11 tone minimal pairs for non-merging tone pairs 2-3 and 5-6 each.[2]. See Table 2 for a full list of stimuli. This produced a total of 34 unique pairs in Experiment 1. Words in each pair were chosen to be familiar, and approximately matched for lexical frequency according to counts from the Hong Kong Cantonese Corpus Luke and

[2] The imbalanced numbers are due to the fact this experiment served double duty: testing the hypothesis described here about lexical competition and perceptual category structure and pre-testing for a future experiment on perceptual learning of tones in Cantonese

Wong (2015). Items in Experiment 2 consisted of eight word-nonword tone pairs selected for the merging pairs (2-5, 3-6 and 4-6) and 22 such tone pairs for the non-merging tone pairs each (2-3 and 5-6; see table 5). Since each tone in a pair had to be represented as a word and a nonword, this produced twice as many tone pairs for each tone pair compared to Experiment 1, however, half of the items in Experiment 2 were nonwords. This produced a total of 68 unique tone pairs in Experiment 2. Altogether, there were 102 unique pairs; tone minimal pairs were intended for Experiment 1 and word-nonword pairs were intended for Experiment 2.

All stimuli were recorded by a linguistically trained female native speaker of Cantonese originally from Hong Kong (35 years old), who is also a Cantonese language teacher and self-reports not producing the tone mergers. Items were recorded in a sound-attenuated cabin in Audacity through a Samson C03U USB microphone at a sampling rate of 44.1 kHz and 24-bit depth (Team, 2010). Following recording, 11-step continua were created for each unique tone pair. We used TANDEM-STRAIGHT to synthesize tone continua (Kawahara et al., 2008), as opposed to PSOLA since Yang et al. (2019) only observed significant tonal categorization effects for TANDEM-STRAIGHT continua. Unlike PSOLA, where only the pitch of a token is manipulated, TANDEM-STRAIGHT considers several natural acoustic cues of the CV syllable during resynthesis by proportionally manipulating the duration and voice quality of the entire tone-bearing syllable as well.

2.2 Acoustic-Auditory Analysis

The purpose of the Experiments 1 and 2 is to quantify the role that lexical competition plays in the perception of merging and non-merging tones. While several tone mergers have been documented in the literature (see Section 1.4), not all speakers exhibit these tone mergers to the same degree, and some maintain distinct categories. In order to conduct these experiments without confounds of, for example, the inherent ambiguity of merging tones or the reduction of contrast in the stimuli, we first quantify the distinctiveness of our stimuli.

F0 was estimated from the tone-bearing units in Voicesauce (Shue, Keating, Vicenik, & Yu, 2011) using the STRAIGHT algorithm (Kawahara, Masuda-Katsuse, & De Cheveigne, 1999), which generated estimates every millisecond. These f0 estimates were transformed into equivalent rectangular bandwidths (ERBs, Moore & Glasberg, 1983). These ERB estimates were then smoothed using the loess() function in R with a span of 0.25 to allow for the maintenance of natural variation. Figure 1 presents the smoothed ERB estimates by tone pair, with the two non-merging tones presented on the leftmost part of the graph, and the three merging tones on the right.

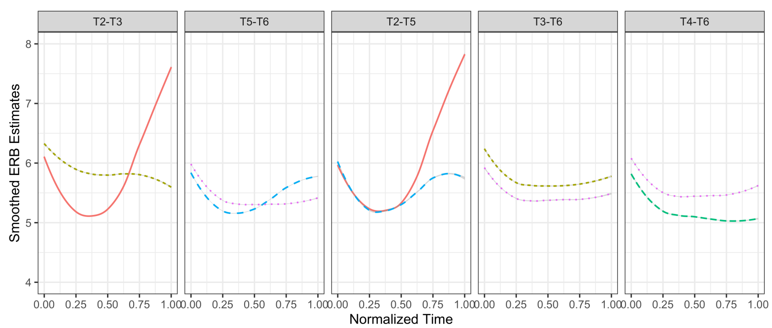

Figure 1. Smoothed ERB estimates for the items in each tone pair.

2.3 Analysis and Results

To quantify acoustic-auditory distance between minimal pair items, an area between two curves algorithm was used (Jekel, Venter, Venter, Stander, & Haftka, 2019). This measure is computed by summing the area of the quadrilaterals that are constructed by joining consecutive pairs of points from the two signals. In the case that the signals do not have equal numbers of points, as is the case with tokens varying in duration, new points are created at the bisection of the segment with the largest Euclidean distance. Adding points in this manner does not change the area but allows for efficient computation. Computing similarity this way makes no assumption of the shape of the trajectories, in either local fluctuations or global patterns of rising or falling. A higher degree of similarity is thus quantified as a smaller area between trajectories; a curve would have no area between itself. In contrast to (Pearson R) correlation, computing area between curves has no upper limit. Figure 2 presents the output of this analysis as boxplots.

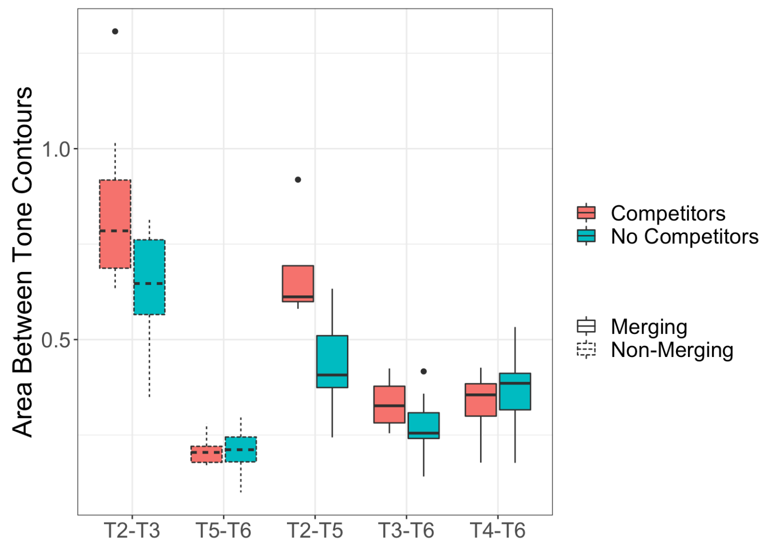

Figure 2. Acoustic distance between tones in non-merging (dashed) and merging pairs (solid) across items.

To assess whether there was a difference in acoustic-auditory distance between merging and non-merging tone pairs and between pairs with real word competitors (i.e., minimal pairs) and those without real word competitors (i.e., word-nonword pairs), the area between the smoothed tone curves was used as the dependent measure in a linear mixed effects regression model. The model included fixed effects of Tone Type (Non-Merging, Merging) and Condition (Competitor, No competitor) with Non-merging and Competitor as reference levels. Tone Pair was as a random effect with random slopes by Condition.[3].

The model returned a significant intercept $[B = 4.475e-01, SE = 1.712e-01, t = 2.613]$, but none of the effects were significant [Tone Type $B = 7.292e-02, SE = 2.688e-01, t = 0.271$; Condition

[3] Model Syntax: AreaBetween ToneType * Condition + (1 + Condition | TonePair)

$B = -8.985e - 02, SE = 6.900e - 02, t = -1.302$; Tone Type * Condition $B = -5.196e - 05, SE = 1.016e - 01, t = -0.001$]. These results suggest that, overall, there were no significant differences between Merging and Non-Merging tone pairs in the magnitude of their acoustic-auditory distance, and that items with lexical competitors were not overall produced with greater contrast than those without lexical competitors.

3 Experiment 1: Word identification task

To examine whether tones that are undergoing mergers in progress are perceived as less categorical than non-merging tones, we carried out a word identification task using pictures with the selected tone minimal pairs.

3.1 Participants

Thirty individuals completed the task. Six individuals who did not learn Cantonese before the age of 5 were removed prior to analysis. One individual who reported learning Cantonese from age 1, but self-reported Cantonese understanding ability as 0 was also removed. The remaining 23 early Cantonese-English bilinguals' demographic summaries and their dominance scores as computed through the Bilingual Language Profile Questionnaire (BLP) are reported in Table 3. The BLP questionnaire asks a series of questions about participants' language history, background, and use. These responses are used to compute a quantified measure of language dominance called a "dominance score" on a scale of +/- 218, where positive scores indicate greater English dominance and negative scores indicate greater Cantonese dominance. As shown in Table 3, dominance scores span a wide range, and this heterogeneity is typical of the Cantonese-English speech community at our university. All participants provided verbal informed consent and were compensated with partial course credit for their time.

3.2 Procedure

Participants first heard a Cantonese version of “The Northwind and the Sun” story from the Aesop Language Bank produced by the same speaker who produced the materials so that participants could be familiarized with her tone range.

Participants completed a word identification task, which was elicited through pictures, as Cantonese literacy cannot be assumed with the participant population. Listeners were presented with a full randomization of the continuum steps from the merging (2-5, 3-6, 4-6) and non-merging (2-3 and 5-6) tone pairs, and asked to categorize each item using the pictures. An example screen shot of a trial is shown in Figure 3. All items were repeated 3 times and the order of the images was not counter-balanced across listeners. Following the task, participants completed the BLP questionnaire. The entire experimental session took approximately 45 minutes.

Figure 3. Trial schematic of Experiment 1: Word identification task. Continua endpoints were real Cantonese words (e.g., /loŋ4/ “wolf” - /loŋ6/ “ocean wave”) matched for frequency and pictured on either side of the screen.

3.3 Analysis and Results

Responses latencies of less than 250 ms and greater than three standard deviations from each individual subjects’ mean were removed from the data set, removing approximately 4% of the responses. These results are visualized in Figure 4, where the proportion of responses for the word at step 1 is plotted as a function of continuum step for non-merging and merging pairs. Step 1 corresponds to the lower numbered tone (e.g., T2 in a T2-T5 pair), while Step 11 corresponds to the higher numbered tone in a pair (e.g., T5 in a T2-T5 pair). The data were entered into a mixed effects generalized logistic model with a maximal random effects structure that converged without error. Tone type (Merging, Non-Merging; sum coded) and Continuum Step (centered, scaled) were fixed effects. There were by-listener, by-item, and by-tone pair random intercepts, with by-listener random slopes for the effect of Step and Tone Type, and by-item random slopes for the effect of Step.[4] The dependent variable was the proportion of responses for the tone at step 1 of each tone continuum. A model that also included lexical frequency differences between the two lexical endpoints did not substantially improve the model $[\chi(4) = 2.7, p = 0.62]$. The model returned a non-significant intercept $[\beta = 0.4696, SE = 0.3188, z = 1.473, p = 0.1408]$. There was a main effect of Step $(\beta = -1.3177, SE = 0.2023 z = 6.513, p < 0.0016.79e - 11)$. While there was no main effect of Tone Type $(\beta = -0.5824, SE = 0.3177, z = -1.833, p = 0.0668)$, there was a significant interaction between Step x Tone Type $(\beta = 0.6943, SE = 0.1737, z = 3.997, p < 0001)$.

Figure 4, illustrates the effect of Step and the interaction of Step by Tone Type. For both merging and non-merging tone pairs, listeners show sensitivity to tonal changes at each continuum step. However, this sensitivity is greater for the non-merging pairs, as shown by the fact that listeners are more extreme in the proportion of their responses at the endpoints, producing a steeper categorization slope. Both of these descriptions suggest more categorical perception for non-merging pairs compared to merging pairs (Schertz & Clare, 2020).

Figure 5 shows these data separated by tone pair, which was included as a random

[4] Model syntax: glmer(Response Step (centered, scaled) * Tone Type (Merging, Non-Merging; contrast-coded) + (1 + Step + Tone Type|Subject) + (1+Step|Item) + (1|TonePair), family = "binomial").

effect in the model. The patterns presented in this visualization indicate that listeners are responding to the two non-merging tone pairs (i.e., T2-T3 and T5-T6) similarly. In contrast, the categorization functions and the relative positions of the categorization functions for the merging tone pairs show more differences. We comment on these differences below, as we discuss the results of Experiment 1.

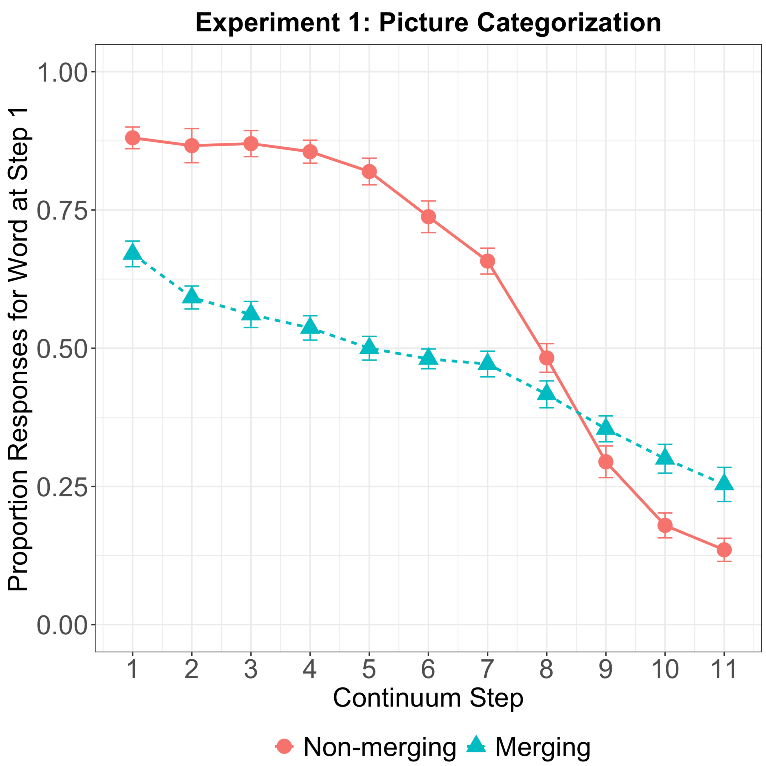

Figure 4. Proportion of trials categorized as the word at Step 1 for Merging and Non-merging tone sets in Experiment 1: Word identification task.

3.4 Discussion

When presented with tone continua where each endpoint is a real word, listeners perceived non-merging tone pairs more categorically than merging tone pairs. This suggests that merging tones have less discrete tone boundaries. This is expected since the phonetic distribution of tones undergoing a merger show more overlap as speakers' productions of one tone category approximate another.

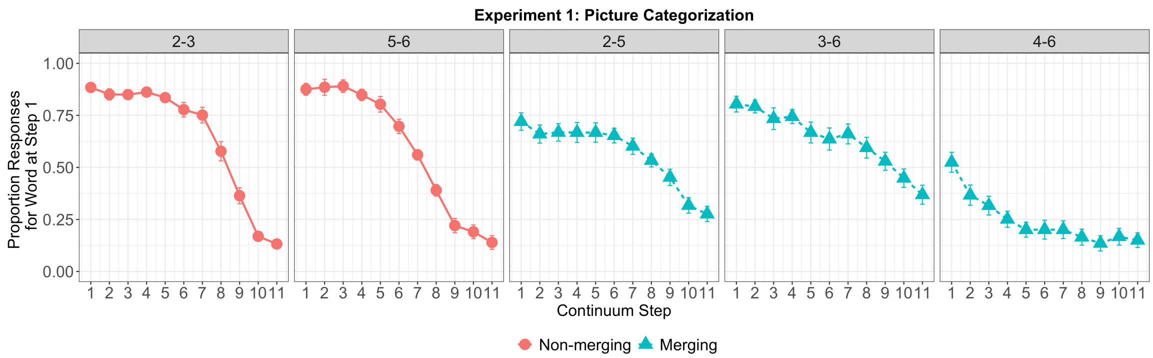

Figure 5. Proportion of trials categorized as the word at Step 1, separated by non-merging and merging tone sets in Experiment 1: Word identification task.

Returning to the observation that the merging tone pairs show pair-specific differences, we note that the relative position of the categorization function for T4-T6 is relatively low, with subjects' mean response for every step except step 1 below 0.5, indicating that listeners were more likely to categorize nearly the entire continuum as tone 6. This is likely because tone 4 is typically accompanied by creaky voice quality, which the talker in this experiment intentionally avoided in order to make speech synthesis less challenging. This likely led to an overall bias to identify this continuum as T6.

These data also provide some support for the claim by Mok et al. (2013) that the direction of the tone mergers is biased by tone type frequency. Specifically, Mok et al., (2013) claim that tones of low type frequency show more variability in production and are thus, more likely to merge in the direction of the other tone category. The type frequency calculations by Fok-Chan (1974) and Leung et al. (2004) show that T5 has a lower type frequency than T2, T4 has a lower type frequency than T6, and T3 and T6 are similar in type frequency. Therefore, T5 is expected to be merging towards T2, T4 is expected to be merging toward T6, and T3 and T6 are expected to be merging toward each other. The predictions for merging pairs T2-T5 and T4-T6 are consistent with our data, as there was an overall higher proportion of T2 and T6 responses in each corresponding tone pair. However, the patterns for T3-T6 do not support these type frequency predictions as our data show an overall bias towards T3. Given we only used one speaker in this study, we are unable to adjudicate whether our results are because of the specific realization of T3 and T6 by our speaker or whether this is indeed representative of these categories as produced by the Cantonese speech community at large.

4 Experiment 2: Lexical Decision

In Experiment 2, we test how the absence of a lexical competitor affects tonal category boundaries, as the absence of an abutting lexical item may lead to category boundaries that are less well-specified for both merging and non-merging tone pairs.

4.1 Participants

The same participants from Experiment 1 also took part in Experiment 2. These experiments were separated by approximately 1.5 weeks.

4.2 Procedure

After providing informed consent, participants again heard a Cantonese version of "The Northwind and the Sun" story produced by the same speaker who produced the materials so that participants could be (re)familiarized with her tone range. Listeners performed a lexical decision task elicited through thumbs up and thumbs down images, since continua endpoints were not always (imageable) words. An example screen shot of a trial is shown in Figure . Listeners were presented with a full randomization of the continuum steps from the merging and non-merging tone pairs, and asked to respond as to whether they heard a real or nonword, by selecting the thumbs up or thumbs down image, respectively. Items were not repeated so as to keep the experiment at a reasonable length. The order of the images was counter-balanced across listeners. Following the task, participants completed the BLP questionnaire. The entire experimental session took approximately 45 minutes.

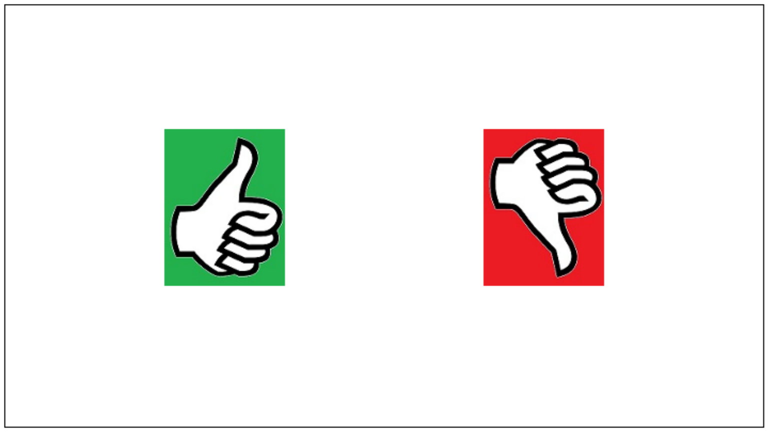

Figure 6. Trial schematic of Experiment 2: Lexical decision task. Continua endpoints were either real Cantonese words or nonwords and pictured on either side of the screen with a thumbs up or down image, respectively. These images were swapped and counterbalanced across participants such that the presentation of images did not favour one side of the screen over the other.

4.3 Analysis and Results

Responses latencies of less than 250 ms and greater than three standard deviations from subjects' individual means were removed from the data set, removing approximately 4% of the responses. Continuum step was flipped such that the real word endpoints of the continua were consistently positioned at continuum step 1. Figure 7 presents these results. These data were entered into a mixed effects generalized logistic model with a maximally warranted structure. Tone Type (Merging, Non-Merging; sum coded) and Continuum Step (centered, scaled) were fixed effects. There were by-listener, by-item and by-tone pair random intercepts, with the interaction of Step and Tone Type as by-listener random slopes, and Step as a by-item random slope.[5] The dependent variable was the proportion of responses for the tone at step 1 of each tone continuum, or the proportion of word endorsements, since the continua were flipped to position real words at step 1.

[5] Model syntax: Response Step (centered, scaled) * Tone Type (Merging, Non-Merging; sum coded) + (1 + Step * Tone Type | Subject) + (1 + Step | Item) + (1|Tone Pair).

The model returned a significant intercept $[\beta = 0.75868, SE = 0.21203, z = 3.578, p < 0.001]$. There were effects of Step $[\beta = -0.64831, SE = 0.09396, z = -6.900, p < 0.001]$ and Tone Type $[\beta = 0.35589, SE = 0.17031, z = 2.090, p = 0.037]$, as well as an interaction between the two $[\beta = 0.20709, SE = 0.06795, z = 3.047, p = 0.0023]$.

As seen in 7, the effect of Step indicates that listeners were sensitive to the acoustic-auditory differences across continuum steps, such that word endorsements were higher for items at the lower steps of the continua. The effect of Tone Type suggests that overall, items from merging pairs were more likely to be categorized as words compared to non-merging pairs. As can be seen in Figure 7, listeners' mean responses to merging tone pairs are above 0.5 at all continuum steps, indicating that these words received word endorsements across the entire continuum. The interaction of Tone Type and Step indicates that at higher continuum steps, items from non-merging tones were more likely to receive non-word responses. Thus, non-merging tones were perceived more categorically than merging tones in these word-nonword pairs where there is no lexical competitor. Figure 8 shows listener responses by tone pair. These by-tone pair results are consistent with the overall responses. Non-merging pairs 2-3 and 5-6 are perceived more categorically than merging pairs 2-5, 3-6, and 4-6, with mean responses to merging tone pairs above 0.5 at nearly all steps of the continuum.

4.4 Discussion

The results of Experiment 2 are similar to that of Experiment 1. That is, the significant interaction between Step and Tone Type suggests that the absence of a lexical competitor affects merging and non-merging tones differently. Specifically, listeners perceived non-merging tone pairs more categorically than merging tone pairs even in the absence of a tonal lexical competitor reifying the tone boundary. As the continua were flipped to reflect word responses at lower steps of the continua, these data do not inform on

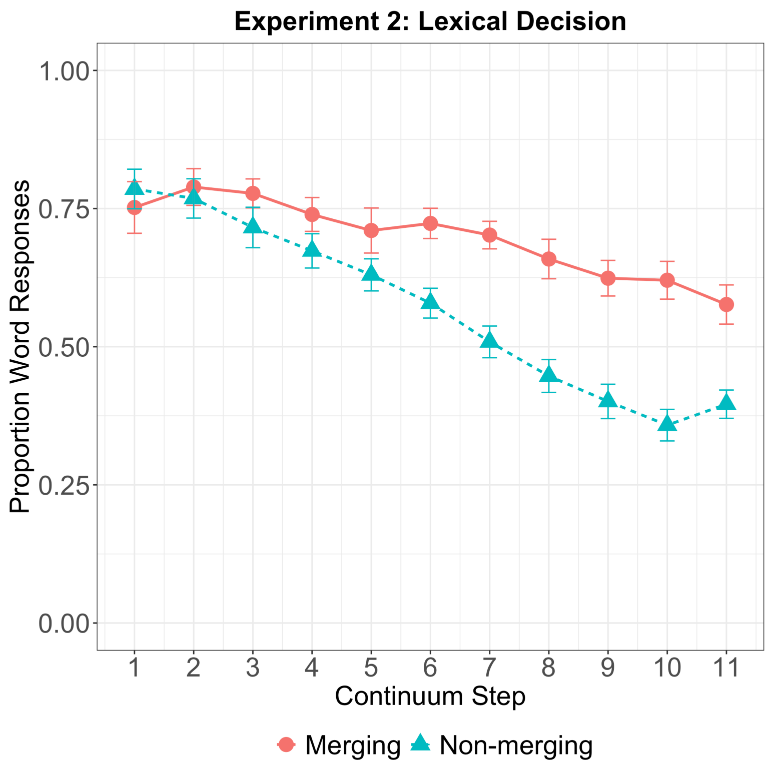

Figure 7. Proportion of trials with word responses for merging and non-merging tones. The scales have been arranged so that all word responses are displayed as continuum step 1.

the directionality of the mergers like the data from Experiment 1 do. However, the finding that listeners' mean responses to all merging tone pairs were above 0.5 at all steps of the continua suggests that listeners were overall more accepting of phonetic variation when there is no lexical competitor in the abutting tone category that would act to constrain it. As such, these data provide further evidence for the tone mergers in perception; not only are the category boundaries less discrete for merging tone pairs, listeners also accept non-canonical pronunciations of tokens as the real word member of a merging tone pair.

The items used in Experiments 1 and 2 differed in whether they had lexical competitors or not, which required different tasks to identify the category boundary for each word. Words with lexical competitors were amenable to a task that queried the tone category boundary as a function of word identification, and words without lexical competitors were assessed by a lexical decision task. While these tasks may invite the use

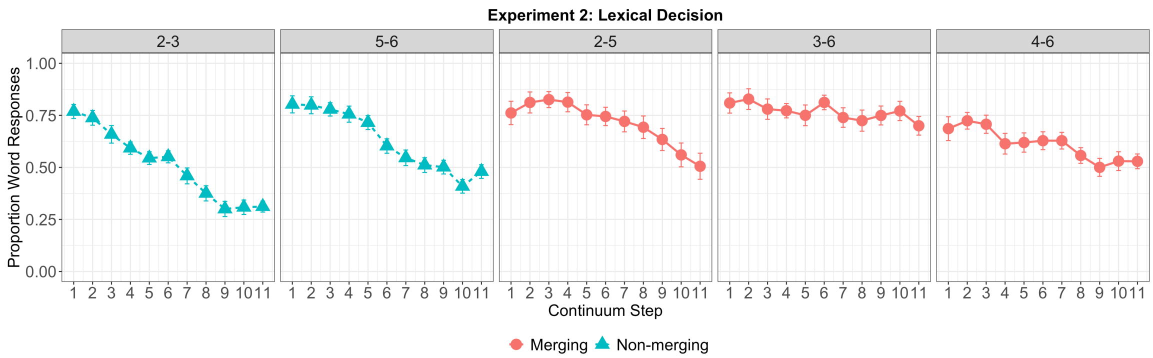

Figure 8. Proportion of trials with word responses for merging and non-merging tones, separated by tone contrast. The scales have been arranged so that all word responses are displayed as continuum step 1.

of different thresholds in assessing the tone category boundaries, in the following section, we ~~nevertheless~~ compare performance across these tasks so that merging and non-merging tones can be directly compared within a single statistical model. This is necessary to directly address the question of whether lexical competition moderates category boundaries.

5 Combined Analysis Across Experiments 1 and 2

As outlined in the introduction, if lexical competition affects how tone is treated at the word level, we expect that listeners will be less accepting of phonetic variation in Experiment 1 where both continua endpoints are words, such that listeners are more categorical in their response functions. In Experiment 2, where one endpoint is a word and the other endpoint is a nonword, we predict that listeners will be more accepting of variation, showing reduced categoricity in their response functions at tonal boundaries. These predictions reflect the Ganong effect (Ganong, 1980) for the non-merging tone pairs, and follow from the idea that lexical competition inhibits mergers. For the merging tone pairs, however, this is ultimately a question of whether the merger has obfuscated the effects of lexical competition. Have mergers-in-progress already blurred the tonal category boundaries, and consequently also neutralized the lexical distinctions? As such, the items involved in a merger-in-progress are not predicted to show differences in categorization functions as a function of lexical competition.

5.1 Analysis and Results

To this end, we carried out two hierarchical generalized logistic models with maximally warranted structure, one for non-merging tones[6] and another for merging tones[7]. Both models had fixed effects of Step (centered, scaled) and Experiment (Word Identification vs. Lexical Decision; sum coded). The dependent variable was the proportion of items categorized as step 1. This refers to the proportion of responses for the lower numbered tone in a pair for Experiment 1, and the proportion of word responses for Experiment 2.

The model for the non-merging tones returned a significant intercept $[\beta = 0.7883, SE = 0.1561, z = 5.050, p < 0.001]$. There was an effect of Step $[\beta = -1.5312, SE = 0.1741, z = -8.793, p < 0.001]$, Experiment $[\beta = 0.3767, SE = 0.1325, z = 2.844, p = 0.004]$, and an interaction between Step and Experiment $[\beta = -0.6641, SE = 0.1236, z = -5.372, p < 0.001]$. This interaction captures that listeners had a steeper categorization function in the word identification task where both continuum endpoints were real words. The categorization function in the lexical decision task was less steep, suggesting that listeners were more accepting of phonetic variation when making decisions where only one endpoint was a real word. These results are visualized in the left panel of Figure 9.

The model for the merging tones returned a non-significant intercept $[\beta = 0.48900, SE = 0.31061, z = 1.574, p = 0.12]$. There was an effect of Step

[6] Model Syntax: Response Step (centered, scaled) * Procedure (Exp 1, Exp 2; sum coded) + (1 + Step * Procedure | Subject) + (1 + Step | Item) + (1|Tone Pair)).

[7] Model Syntax: Response Step (centered, scaled) * Procedure (Exp 1, Exp 2; sum coded) + (1 + Step * Procedure | Subject) + (1 + Step | Item) + (1|Tone Pair)).

$[\beta = -0.52543, SE = 0.08708, z = -6.034, p < 0.001]$ and Experiment $[\beta = -0.61601, SE = 0.17333, z = -3.554, p = 0.0004]$, but the interaction was not significant $[\beta = -0.09222, SE = 0.06299, z = -1.464, p = 0.14]$. The effect of Step indicates that listeners' responses varied as a function of each incremental continuum step, with responses for an item at continuum step 1 becoming less likely at higher continuum steps regardless of whether the item had a lexical competitor. The effect of Experiment indicates that overall, listeners had a higher proportion of responses at all continua steps when there was no lexical competitor. However, since there was no interaction between Experiment and Step, the presence or absence of a lexical competitor did not affect the shape of the categorization function. These data can be seen in the right panel of Figure 9.

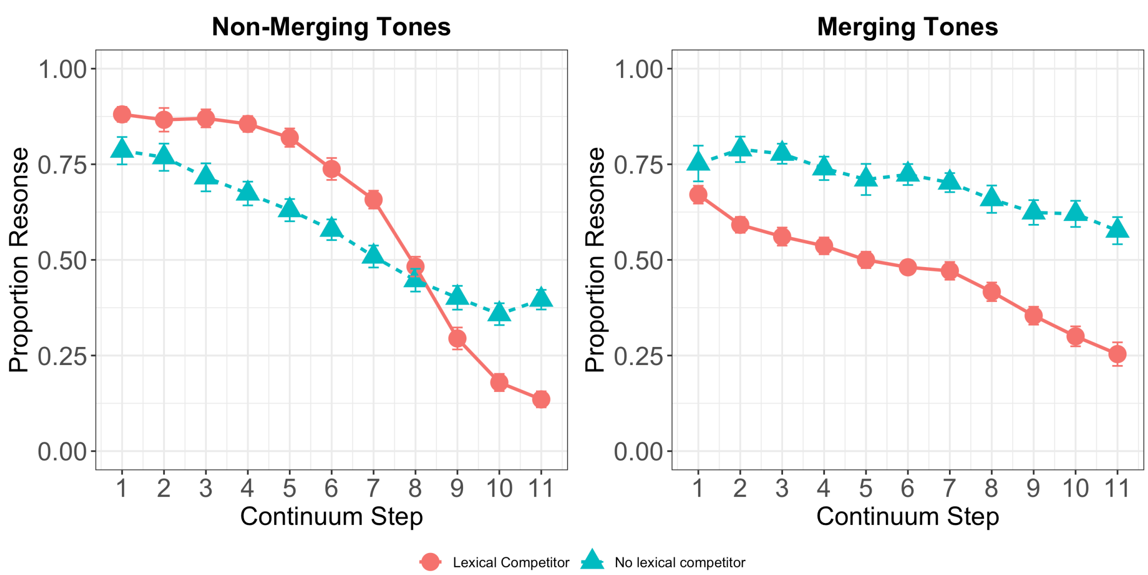

Figure 9. Proportion of responses for the tone at Step 1 in Experiment 1 (Word identification task, red solid line) and Word in Experiment 2 (Lexical decision task, blue dashed line). The scales have been arranged so that all word responses are displayed as continuum step 1. The left panel shows these patterns for non-merging tone sets and the right panel presents data for merging tone sets.

5.2 Discussion

The results of the combined analysis indicate that there is an effect of lexical competition for non-merging tone pairs, but this is less robust for merging tone pairs. Specifically, for non-merging tone pairs, there was an interaction between Step and Experiment, while this interaction was not significant for merging tone pairs. This is evident in the visualization of these data in Figure 9. The left panel shows that the response functions for non-merging pairs with and without lexical competitors cross over, while the right panel shows that the response functions for the merging pairs with and without lexical competitors largely run parallel to each other. Although the interaction between Step and Experiment was not significant for merging tone pairs, responses in the lexical decision task for items without lexical competitors were above 0.5 at all continuum steps. These results demonstrate the Ganong effect, as listeners were more likely to perceive these items as words than not across the entire continuum from the clear word endpoint to that of a nonword production. The response function for the merging tones in the word identification task crosses the 0.5 threshold, but the lower rate of identification at step 1 suggests that listeners may be identify step 1 as the word at step 11 with some frequency. Thus, it may be the case that the category threshold for merging tones is subject to the lexical status of the word in which it is part.

Probabilistic frameworks of speech perception predict that phonetic variation should affect both the boundary between categories and the perceived goodness within categories (Feldman, Griffiths, Goldwater, and Morgan, 2013; Kleinschmidt and Jaeger, 2015) As such, we carried out a goodness rating task to examine whether these findings are echoed in an experiment that examines the effect of lexical competition on internal category structure.

6 Experiment 3: Goodness Rating

Experiments 1 and 2 examined how lexical competition affects merging versus non-merging Cantonese tones at category boundaries. We found that lexical competition maintains more discrete categorization functions for non-merging tones than for merging tones, suggesting that, unsurprisingly, the boundaries for merging tones are less well-defined and the presence of an anchoring lexical item in the corresponding tones does little to help sharpen category boundaries, though the results of Experiment 2 suggests that merging tones without lexical competitors exhibit a global word-bias. As the tasks in Experiment 1 and 2 examine listener categorization functions, the internal structure of tonal categories cannot be elucidated with these data. To understand how lexical competition shapes the internal category structure of merging and non-merging tone pairs, we carried out a goodness rating task in which listeners heard the same tone continua from Experiments 1 and 2.

6.1 Materials

The stimuli for Experiment 3 consisted of a subset of the stimuli from Experiments 1 and 2. Specifically, for each of the non-merging and merging tone pairs, four minimal pairs from Experiment 1 and eight word-nonword tone pairs from Experiment 2 were selected. Items used in Experiment 3 are bolded in Table 2 and Table 5. This produced a total of 60 unique pairs in Experiment 3.

6.2 Participants

Nine individuals took part in the goodness rating task. Two individuals who did not learn Cantonese before the age of 5 were removed prior to analysis. An additional two individuals were removed due to technical issues. The remaining 5 Cantonese-English bilinguals' demographic summaries are reported in Table 6. As with the participants in Experiments 1 and 2, these bilinguals had a wide range of scores on their BLP (Gertken et al., 2014). All participants provided verbal informed consent and were compensated with partial course credit for their time.

This experiment was implemented just prior to the quarantine imposed by the COVID-19 pandemic. As a result, the participant count is rather small. Since the goodness rating task was designed to complement Experiment 1 and 2, both of which were carried out in person, we ultimately decided against carrying out a web-based version of this task, which introduces audio quality as a dimension that does not exist in Experiments 1 and 2 where all participants used the same model of headphones. To that end, we present an exploratory analysis of these data in an effort to begin to understand how lexical competition affects internal category structure of lexical tones in Cantonese.

6.3 Procedure

Since the items for Experiment 3 were taken from Experiment 1 and 2, participants were again presented with the Cantonese version of "The Northwind and the Sun" story to be familiarized with the speaker's tone range. In the goodness rating task, participants were presented with a full randomization of the continuum steps from the selected minimal pairs and word-nonword pairs representing the merging (2-5, 3-6, 4-6) and non-merging (2-3, 5-6) tone pairs. They were instructed to rate the goodness of a token as the real word continua endpoint on a scale of 1 (poor) to 7 (good). The real word at continuum endpoint was presented on the screen to the participants with the Chinese character, English translation, and Jyutping, since endpoints were not always imageable words and Cantonese literacy cannot be assumed in this population of bilinguals.

Since the selected tone minimal pairs from Experiment 1 bore real words on both endpoints of the continuum, these continua were presented to participants twice where they rated the goodness of tokens according to the word on each end of the continuum. To prevent listeners from hearing the same continuum tokens twice, the stimuli were separated into two counterbalanced blocks which listeners completed in separate experimental sessions, separated by approximately 1.5 weeks. One block contained the words representing tones 2, 3, and 4 in the corresponding tone pairs (i.e., **2**-5, **2**-3, **3**-6, **4**-6, respectively), and the other block contained the words representing tones 5, and 6 in the corresponding tone pairs (i.e., 2-**5**, **5**-6, 4-**6**). Items were not repeated to keep each experimental sessions at a reasonable length.

Although this task does not query category boundaries, we nonetheless predicted similar sinusoidal response functions for non-merging tone pairs, such that high goodness ratings would be observed at continuum endpoints corresponding to the target tones being judged in each pair. Lexical competition should affect the perceived goodness of non-merging tone categories in a similar way by maintaining more discrete listener response functions. In merging tone pairs however, in line with the results from Experiment 1 and 2, the goodness ratings are not predicted to show such a discrete response functions. Instead, as the sounds are undergoing a merger, the perceived goodness of target tones as members of tonal categories are predicted to be relatively high at nearly all continuum steps. Likewise, the absence of a lexical competitor is predicted to affect the perceived goodness ratings for non-merging tone pairs by producing overall higher ratings at ll continuum steps when there is no abutting tone category.

6.4 Analysis and Results

Given the exploratory nature of this experiment, we examined the role of lexical competition for each tone pair separately. Specifically, for each tone pair, we carried out a full linear mixed effects model and a sub-model, both of which predicted rating responses. The full model included fixed effects of Lexicality (Competitor vs. No Competitor), Step (centered and scaled), and Target Tone. This full-model was compared to a sub-model in which Lexicality was removed as a fixed effect. Both models included by-item random slopes for the effect of Step, and by-subject random slopes for the effect of Step, Target, and Lexicality and their interactions. For tone pairs 3-6 and 5-6, only by-item and by-subject random intercepts were included in order for the model to converge. Likelihood ratio tests were carried out to compare the full model with the sub-model for each tone pair to examine the effect of Lexicality.

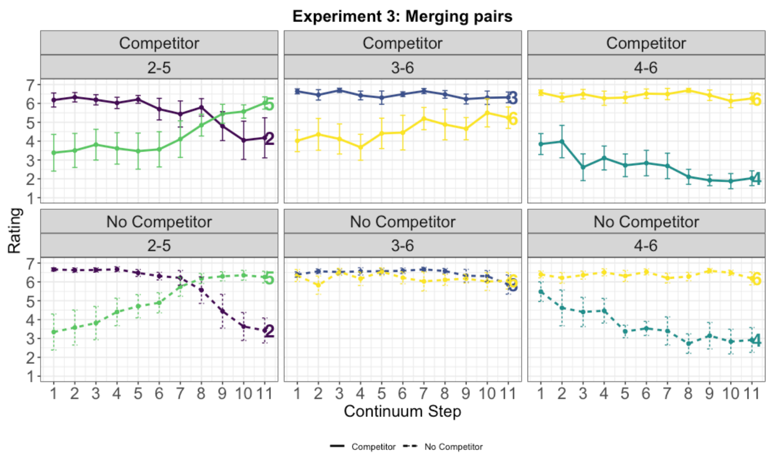

Figure 10. Rating results for Merging tone sets in Experiment 3: Goodness rating task. The tone at step 1 represents the first member of each tone set (e.g., for tone pair 2-5, step 1 is tone 2, while step 11 is tone 5.

The results of the model comparisons revealed that Lexicality significantly improved the fit of the model for all tone pairs except 2-3 ($\chi(30) = 26.168, p = 0.6665$). In other words, for non-merging pair 5-6 ($\chi(4) = 30.633, p < 0.001$) and merging pairs 2-5 ($\chi 30) = 55.428 p < 0.01$), 3-6 ($\chi(4) = 34.388, p < 0.001$), and 4-6 ($\chi(30) = 99.916, p < 0.001$), the rating results for each member of a tone pair were significantly different when both endpoints were real words.

6.5 Discussion

The results from the goodness rating task for merging tones and the non-merging tones are presented in Figures 10 and 11, respectively. The exploratory analysis carried out on the small set of goodness rating data indicate that lexical competition does not play a role in the perceived goodness of the tone categories in the non-merging tone pair 2-3. Indeed, in Figure 11, the response functions for target tones 2 and 3 are similarly shaped,

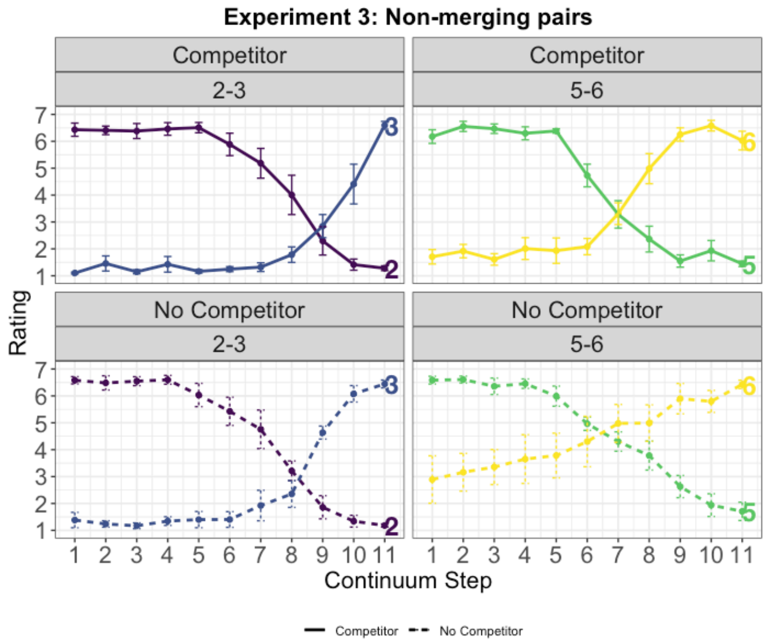

Figure 11. Rating results for Non-merging tone sets in Experiment 3: Goodness rating task. The tone at step 1 represents the first member of each tone set (e.g., for tone pair 2-3, step 1 is tone 2, while step 11 is tone 3.

irrespective of whether there was a competitor at both ends or not.

In all other tone pairs, the presence or absence of a real word in the abutting tone affected the perceived goodness of the target tone in question. For instance, the response functions for target tones 3 and 6 in pair 3-6 (middle panel of 10) are relatively high at all continuum steps, but the response functions are essentially overlapping in cases where there is no competitor in the abutting tone category. These data suggest that listeners are more accepting of phonetic variation in the absence of a competing lexical item that would anchor the tonal category. More broadly, while the sample size is small, these goodness rating results may provide additional evidence that lexical competition plays a role in the mergers between pairs 2-5, 3-6, and 4-6. Since the perceived goodness of the tonal categories appears to vary as a function of whether there is a real word to which the tone could be mapped, these data echo the sentiment in sound change literature that mergers may be inhibited for the purposes of maintaining recognition (A. Wedel, Jackson, &

Kaplan, 2013). However, there was also no effect of lexical competition pair 5-6, which is not reported to be merging. This lack of a lexical effect for tone pair 5-6 may be spurious, constrained by the small sample size, or be grounded in the fact that this tone pair is the most acoustically distinct of those examined here.

7 General Discussion

The current study was aimed at investigating the effect of lexical competition on the perception of Cantonese tone categories. In particular, since lexical competition has been shown to play a role in contrast maintenance for diachronic sound changes, we examined both merging tone pairs (T2-T5, T3-T6, T4-T6) and non-merging tone pairs (T2-T3, T5-T6) in the language (see Figure 1).

To do so, we created tone minimal pairs and word-nonword pairs for the set of merging and non-merging tone pairs. 11-step continua were synthesized for each of the minimal pairs and word-nonword pairs in TANDEM-STRAIGHT (Kawahara et al., 1999). As minimal pairs bore real words at both end points, they were utilized in a word identification task (Experiment 1) to examine how the presence of a lexical competitor may affect listener categorization functions. Likewise, since word-nonword pairs only had a real word on one endpoint (while the other endpoint was a nonword), they were included in a lexical decision task (Experiment 2) to examine how the absence of a lexical competitor would affect listener categorization functions.

As predicted, merging tones exhibited less steep categorization functions than non-merging tones in both experiments, indicating that merging tones were perceived less categorically than non-merging tones. These results provide additional evidence for how these mergers affect phonetic and phonological category structure in Cantonese. In an AX discrimination task testing all possible tonal combinations, Fung and Lee (2019) observed the lowest accuracy rates for pairs T2-T5 and T4-T6 by Hong Kong Cantonese speakers, which suggests that these tones are well advanced in terms of a merger. Mok et al. (2013), also observed that participants who are merged in production were likewise slower at discriminating merging tone pairs in an AX task. These results were paralleled in a heritage speaker population by Soo and Monahan (2017) who observed low sensitivity (*d'*) scores for T2-T5 as well as T3-T6 in an AX task. In a word-identification paradigm, Lam (2018) also observed overall higher confusion rates between tones in pairs T2-T5, T3-T6, and T4-T6 in heritage (raised in Canada) and homeland (raised in Hong Kong) listeners. Finally, children acquiring Cantonese as a heritage language in the United States studied in Kan and Schmid (2019) were also less accurate than children acquiring Cantonese in a Cantonese-dominant speech community when asked to discriminate T2-T5 contrasts in an ABX task.

Cross-experimental comparisons were carried out to examine the effect of lexical competition for non-merging and merging tone pairs. In non-merging tones, the range of acceptable phonetic variation was greater in the absence of a lexical competitor, suggesting that lexical competition maintained more discrete tonal category boundaries. These results align with those of Yang, Jin, and Lu (2019), who also observed more categorical perception of (non-merging) tones when a word was present at both ends of a tonal continuum. Overall, these data support the broader claim that lexical competitors restrict the range of acceptable phonetic variation (Baese-Berk & Goldrick, 2009; Goldrick et al., 2013), extending such conclusions to the realm of suprasegmentals. Conversely, merging tones were less affected by lexical competition. The boundary between merging tonal categories was less discrete irrespective of whether there was a real word at both or only one continuum endpoint. However, listeners globally shifted their tone categorization functions in the absence of a lexical competitor, such that at all steps of the continuum word-nonword pairs were above the 0.5 threshold. Thus, while listeners were more accepting of non-canonical tone productions in the absence of a lexical competitor, the categorization functions were not significantly different across Experiments 1 and 2.

Categorization tasks delineate and qualify the nature of the category boundaries between sounds, while goodness rating tasks query the internal category structure of a given sound. Experiment 3 serves to complement the first and second experiments with insight into the lexical competition effects on within-category structure. While the data in Experiment 3 come from a small sample size due to pandemic data collection restrictions, the perceived goodness of certain tones as members of the category to which it is merging was high overall, and indeed even higher in instances where only endpoint was a word. In contrast, non-merging tone categories were assigned more discrete ratings across continuum steps. This suggest that the within-category structure of merging tones identifies a wider range of phonetic variation as a good exemplar of the tone category, in comparison to non-merging tones, which exhibit a more limited distribution of good exemplars.

The previous body of work on tone mergers is united by its use of naturally produced stimuli. By probing listener categorization in the face of variation across synthesized continua, our results provide a more comprehensive picture of the nature of the category boundaries in merging and non-merging Cantonese tones. Merging tones have a less discrete category boundary compared to non-merging tones; however, merging tone continua still showed an effect of continuum step, indicating that the tones are not completely merged. That said, the distinctiveness of merging tones is no longer supported by lexical competition like non-merging tones are. The more speculative – due to the limited sample size – goodness rating data suggest that the internal structure of a category is generally affected by lexical competition, regardless of merger status. This suggests that while tone mergers may have muddied the category boundaries, the internal structure of a merging category still maintains a distinction between what is a good exemplar for a word with and without lexical competition. Again, given the small sample size for Experiment 3, this is speculative.

Overall, what do these results suggest about the nature of tone categories? Following A. Wedel and Fatkullin (2017) and Yu (2007), we propose that tone categories may be represented by a distribution of perceptual exemplars (Pierrehumbert, 2001; Yu, 2007) which, if merging, bear potentially substantial areas of overlap with other tonal distributions at category boundaries. Under the assumption that phonetic categories are distributions of experienced items (Bybee, 2001; Pierrehumbert 1990, 2001, 2002, Goldinger 1996, Johnson 1997, Ernestus and Baayen 2003), speech perception involves the categorization of new percepts to categories which occupy a similar perceptual space. In this sense, A. Wedel and Fatkullin (2017) formalize category competition as a means through which category contrasts can be maintained. In a one-dimensional space, the distribution of these categories are subject to "entrenchment", acting to tighten category distribution and "noise", acting to broaden the category distribution. The balance between these two forces can be disrupted in cases where two categories approach one another along a shared acoustic dimension. As the region of overlap at the category boundary begins to increase, usage frequency comes into play in promoting greater activation of one of these categories (Pierrehumbert, 2001). Since categories compete for new percepts, the category with increased activation will be assigned more percepts, gradually increasing the area of overlap between the two categories yet more (and eventually, theoretically, eradicating the distinction between the categories entirely in favour of one category). In this sense, the boundary between merging tone categories is less discrete than that of non-merging tones, as the distribution of exemplars bear an area of overlap at category boundaries. At the same time, listeners are still able to perceive the difference between merging tones overall, as the extreme ends of these distributions do not overlap.

As mentioned, while the combined analysis of Experiment 1 and 2 for merging tones showed a lack of an interaction between continuum step and the presence of a lexical competitor, there was a simple effect for lexical competition. Listeners' categorization functions were globally shifted such that while listeners were still sensitive to changes in step, all continua steps were classified as words. This suggests that for merging items without lexical competitors, the entire tone distribution is skewed in a way that tone realization may vary considerably while still being recognized as a word. This suggests that what is acceptable for a category boundary is influenced by its mapping to the lexicon, and not purely the acoustic-phonetic realization of a signal.

A. Wedel and Fatkullin (2017) also provide an account for the contribution of lexical factors. When two categories begin to approach one another, noise in the opposing extreme ends of each distribution can pull the category distributions away from the regions of overlap at category boundaries. This “variant trading” behaviour, which would normally act to sharpen category boundaries, is inhibited when ambiguous incoming percepts are disambiguated by the absence of a lexical competitor. In other words, when the percept can easily be categorized by using lexical knowledge about whether the resulting token would produce a real word in the language (e.g., Ganong, 1980; Norris et al., 2003), category competition, and ultimately, category contrast is uninhibited. In sum, the *absence* of a lexical competitor does not promote variant trading or category contrast, while the *presence* of a lexical competitor does.

From this perspective, the effect of lexical competition may be less robust for merging tone categories, as Cantonese phonological-lexical representations themselves may be stored with less phonological detail. This interpretation is consistent with the fuzzy lexicon hypothesis proposed for L2 speech processing (Gor, Cook, Jackson, 2010). Several studies have shown that categories that are confusable may produce lexical representations that lack phonological detail for L2 speakers (Darcy, Daidone and Kojima, 2013). While much of this literature situates confusable categories in the context of L2 learners who struggle to differentiate new L2 categories from similar L1 categories (Gor, Cook, and Jackson, 2010; Cook et al., 2016), the case of merging tones in early Cantonese-English bilinguals represents a unique departure from this body of work in two ways. Firstly, many of the bilinguals in the current study were more dominant in their L2, English (as evidenced by the distribution of BLP scores in Table 3). Secondly, the confusability of the sound categories in question (i.e., lexical tones) is due to their merging within the L1 (Cantonese), not due to their similarity to a category in the L2. In fact, for this population of bilinguals, these merging L1 tonal categories bear no direct analogue in the L2. In this respect, a number of factors conspire to produce lexical representations that lack phonological detail, as tone categories are merging in the less-dominant first language of these bilinguals, and no support can be found in English.

Finally, on a broader level, this study provides evidence for the notion that gradual phonetic change may be affected in cases where maintaining recognition in competing categories is paramount. Words without lexical competitors whose tones are merging are globally likely to be identified as words. The category threshold for merging tones is subject to the lexical status of the word in which it is part. Furthermore, a within-category examination of the tones undergoing mergers in Experiment 3 revealed that merging tone pairs were affected by the presence or absence of a lexical competitor in the abutting tone category. That is, goodness rating results were more discretized when target tones corresponded to real words, suggesting that listeners are implicitly aware of the presence or absence of a lexical competitor and that the internal structure of a merging category is improved when there is a competing lexical item that acts to constrain the range of acceptable phonetic variation. Thus, when listeners are asked to rate the goodness of a token as a member of a category, they are able to cull from the portions of the distributions that are not overlapping to make judgements about the goodness of a token.

8 Conclusion

We carried out three experiments to examine the effect of lexical competition on the categorization and internal category structure of merging and non-merging tone pairs in Cantonese. Experiments 1 and 2 showed that category boundaries between non-merging tone categories were discrete, and sharpened by the presence of a lexical competitor in the abutting tone category could, while merging tone category boundaries were less discrete and showed only a simple effect of lexical competition. On the other hand, the perceived goodness of merging tones in experiment 3 showed a stronger effect of lexical competition than non-merging tones. These data can be explained if we conceive of a tone category as a distribution of perceptual exemplars. Since tone categories that are merging bear areas of potential overlap with one another, the boundary between these categories will be less discrete while the distribution of the categories themselves will encapsulate a wider range of phonetic variation as a good exemplar of the corresponding merging tone category. These data provide an account of the lexicon's contribution to the ongoing tone mergers in Cantonese, and more broadly, demonstrate the effects lexical competition in structuring phonetic variation for suprasegmental sound categories.

Acknowledgments

This work has been supported by an NSERC grant to MB. Thanks to Martin Oberg for programming assistance on this project. We thank Zoe Lam for recording stimuli. We also thank Angelina Lloy, Fion Fung, Christina Sen, Ariana Zattera, and other members of the Speech in Context Laboratory for their assistance on this project.

Tone Pair	Jyutping	Character	English gloss	Jyutping	Character	English gloss
2-3	fan2	粉	powder	fan3	瞓	sleep
2-3	gwai2	鬼	ghost	gwai3	貴	expensive
2-3	geng2	頸	neck	geng3	鏡	mirror
2-3	deng2	頂	roof	deng3	掟	throw
2-3	fu2	虎	tiger	fu3	褲	pants
2-3	gwan2	滾	boil	gwan3	棍	rod
2-3	tong2	糖	candy	tong3	熨	iron
2-3	zin2	剪	to cut	zin3	箭	arrow
2-3	zaau2	爪	a claw	zaau3	罩	mouth mask
2-3	sai2	洗	wash	sai3	細	small
2-3	paa2	扒	steak	paa3	怕	fearful
5-6	hau5	厚	thick	hau6	後	behind
5-6	jim5	染	dye	jim6	驗	written test
5-6	jyun5	軟	soft	jyun6	願	wish
5-6	lai5	禮	a gift	lai6	荔	lychee
5-6	laan5	懶	lazy	laan6	爛	rotten
5-6	lei5	□ □ □ □ □ □	sail	lei6	脷	tongue
5-6	lou5	老	old	lou6	路	road
5-6	maa5	馬	horse	maa6	罵	scold
5-6	maan5	晚	night time	maan6	慢	slow
5-6	mun5	滿	full	mun6	悶	bored
5-6	ngo5	我	me	ngo6	餓	hungry
2-5	si2	屎	poop	si5	市	city
2-5	jyu2	瘀	a bruise	jyu5	雨	rain
2-5	sin2	癬	ringworm	sin5	鱔	eel
2-5	tou2	禱	prayer	tou5	肚	stomach
3-6	beng3	柄	a handle	beng6	病	sick
3-6	dung3	凍	cold	dung6	動	exercise
3-6	giu3	叫	call out	giu6	撬	pry open
3-6	zoeng3	醬	sauce	zoeng6	象	elephant
4-6	long4	狼	wolf	long6	浪	ocean wave
4-6	jau4	游	swim	jau6	右	right (direction)
4-6	se4	蛇	snake	se6	射	shoot
4-6	mou4	毛	fur	mou6	霧	fog

Table 2

Items used in Experiment 1. Bolded items were also used in Experiment 3.

Table 3

Participant information about Cantonese and English Age of Acquisition (AoA), Cantonese and English self-ratings for understanding (out of 7), and Bilingual Language Profile (BLP) scores.

	Age	Cantonese AoA	Cantonese Rating	English AoA	English Rating	BLP
s1	21	Birth	5	5	6	85
s2	20	Birth	6	Birth	4	-50
s4	18	Birth	6	1	6	-6
s6	19	Birth	5	4	6	79
s7	19	Birth	4	1	6	79
s9	20	Birth	6	6	6	-46
s10	21	Birth	6	4	5	-59
s11	22	2	5	8	5	34
s12	17	Birth	4	Birth	6	121
s13	21	Birth	5	3	6	56
s15	18	Birth	4	Birth	6	127
s16	19	Birth	3	3	5	97
s17	48	Birth	3	8	6	139
s18	20	Birth	6	2	5	-26
s19	20	Birth	6	2	6	-13
s20	18	Birth	5	Birth	6	52
s22	23	Birth	4	6	5	60
s24	21	Birth	6	3	6	25
s31	18	Birth	5	9	5	-28
s34	21	1	0	3	6	170
s37	18	Birth	5	3	6	55
s42	22	Birth	5	5	6	108
s46	18	4	5	7	5	51
s53	19	Birth	4	2	6	63

Tone Pair	Jyutping	Character	English gloss	Nonword Match in Jyutping
2-3	pui3	沛	abundant	pui2
2-3	ping3	聘	invite for service	ping2
2-3	paai3	派	distribute	paai2
2-3	leng3	靚	beautiful	leng2
2-3	kau3	釦	buckle	kau2
2-3	kaau3	靠	lean on for a favour	kaau2
2-3	hing3	㷫	to warm	hing2
2-3	gwaan3	慣	accustomed to	gwaan2
2-3	goe3	鋸	to saw	goe2
2-3	faai3	快	fast	faai2
2-3	caa3	岔	crossroads	caa2
2-3	wun2	碗	bowl	wun3
2-3	mung2	懵	ignorant	mung3
2-3	mo2	摸	touch	mo3
2-3	mang2	忟	to be irritated	mang3
2-3	loeng2	両	a unit of weight	loeng3
2-3	lem2	舐	lick	lem3
2-3	hoi2	海	ocean	hoi3
2-3	dai2	抵	worth the money	dai3
2-3	daa2	打	to fight	daa3
2-3	cin2	淺	shallow	cin3
2-3	ceng2	請	invite	ceng3
5-6	waai6	壞	spoilt	waai5
5-6	meng6	命	one’s life	meng5
5-6	gwai6	跪	kneel	gwai5
5-6	gui6	攰	tired	gui5
5-6	gau6	舊	worn, old	gau5
5-6	ding6	定	or	ding5
5-6	daam6	啖	a bite, mouthful	daam5
5-6	daai6	大	big	daai5
5-6	bou6	步	a step	bou5
5-6	bei6	備	ready	bei5
5-6	baan6	扮	dress up as someone else	baan5
5-6	tyun5	斷	break	tyun6
5-6	pui5	倍	double	pui6
5-6	pou5	抱	hug	pou6
5-6	pei5	被	blanket	pei6
5-6	leng5	領	shirt collar	leng6
5-6	keoi5	佢	him, her, she, he	keoi6
5-6	kei5	徛	to stand	kei6
5-6	kan5	近	near, close to	kan6
5-6	cou5	儲	save up	cou6
5-6	co5	坐	sit	co6
5-6	ci5	似	similar to	ci6

Table 4

Items used in Experiment 2. Bolded items were also used in Experiment 3.

Tone Pair	Jyutping	Character	English gloss	Nonword Match in Jyutping
2-5	wing5	永	forever	wing2
2-5	mei5	美	pretty	mei2
2-5	laang5	冷	chilly	laang2
2-5	je5	嘢	thing	je2
2-5	zai2	仔	son	zai5
2-5	sau2	手	hand	sau5
2-5	fo2	火	fire	fo5
2-5	beng2	餅	cookie	beng5
3-6	maai6	賣	sell	maai3
3-6	lyun6	亂	messy, disorganized	lyun3
3-6	lin6	練	practice	lin3
3-6	doi6	代	era	doi3
3-6	taam3	探	visit	taam6
3-6	hoeng3	向	towards	hoeng6
3-6	gwaa3	掛	hang up	gwaa6
3-6	gaai3	界	boundary	gaai6
4-6	daan6	彈	elasticity	daan4
4-6	jaa6	廿	twenty	jaa4
4-6	gam6	撳	to press	gam4
4-6	dei6	地	ground	dei4
4-6	tim4	甜	sweet	tim6
4-6	tau4	頭	head	tau6
4-6	kung4	窮	destitute, poor	kung6
4-6	haang4	行	walk	haang6

Table 5

More items used in Experiment 2. Bolded items were also used in Experiment 3.

Table 6

Cantonese and English Age of Acquisition (AoA), Cantonese and English self-ratings for understanding (out of 7), and Bilingual Language Profile (BLP) scores for participants in Experiment 3.

	Age	Cantonese AoA	Cantonese Rating	English AoA	English Rating	BLP
s3	20	Birth	6	2	6	-13.436
s6	22	Birth	5	5	6	154.6440
s7	20	Birth	6	3	6	22.616
s8	20	Birth	5	3	5	-53.664
s10	19	Birth	4	Birth	6	60.752

References

Baese-Berk, M., & Goldrick, M. (2009). Mechanisms of interaction in speech production. *Language and cognitive processes*, *24*(4), 527–554.

Gahl, S., Yao, Y., & Johnson, K. (2012). Why reduce? phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of memory and language*, *66*(4), 789–806.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of experimental psychology: Human perception and performance*, *6*(1), 110.

Goldrick, M., Vaughn, C., & Murphy, A. (2013). The effects of lexical neighbors on stop consonant articulation. *The Journal of the Acoustical Society of America*, *134*(2), EL172–EL177.

Hay, J. B., Pierrehumbert, J. B., Walker, A. J., & LaShell, P. (2015). Tracking word frequency effects through 130 years of sound change. *Cognition*, *139*, 83–91.

Jekel, C. F., Venter, G., Venter, M. P., Stander, N., & Haftka, R. T. (2019, may). Similarity measures for identifying material parameters from hysteresis loops using inverse analysis. *International Journal of Material Forming*. Retrieved from `https://doi.org/10.1007/s12289-018-1421-8` doi: 10.1007/s12289-018-1421-8

Kawahara, H., Masuda-Katsuse, I., & De Cheveigne, A. (1999). Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds. *Speech communication*, *27*(3-4), 187–207.

Luke, K.-K., & Wong, M. L. (2015). The hong kong cantonese corpus: design and uses. *Journal of Chinese Linguistics*, *25*(2015), 309–330.

Moore, B. C., & Glasberg, B. R. (1983). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. *The journal of the acoustical society of America*, *74*(3), 750–753.

Munson, B., & Solomon, N. P. (2004). The effect of phonological neighborhood density on vowel articulation. *Speech, Language, and Hearing Research*, 1048-1058.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive psychology*, *47*(2), 204–238.

Schertz, J., & Clare, E. J. (2020). Phonetic cue weighting in perception and production. *Wiley Interdisciplinary Reviews: Cognitive Science*, *11*(2), e1521.

Shue, Y.-L., Keating, P., Vicenik, C., & Yu, K. (2011). Voicesauce: A program for voice analysis. *Proceedings of the ICPhS XVII*, 1846-1849.

Todd, S., Pierrehumbert, J. B., & Hay, J. (2019). Word frequency effects in sound change as a consequence of perceptual asymmetries: An exemplar-based model. *Cognition*, *185*, 1–20.

Tsui, T.-H. (2012). Tonal variation in hong kong cantonese: Acoustic distance and functional load. *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, *48*, 579-588.

Wedel, A. (2012). Lexical contrast maintenance and the organization of sublexical contrast systems. *Language and Cognition*, *4*(4), 319–355.

Wedel, A., & Fatkullin, I. (2017). Category competition as a driver of category contrast. *Journal of Language Evolution*, *2*(1), 77–93.

Wedel, A., Jackson, S., & Kaplan, A. (2013). Functional load and the lexicon: Evidence that syntactic category and frequency relationships in minimal lemma pairs predict the loss of phoneme contrasts in language change. *Language and speech*, *56*(3), 395–417.

Wedel, A., Kaplan, A., & Jackson, S. (2013). High functional load inhibits phonological contrast loss: A corpus study. *Cognition*, *128*(2), 179–186.

Wedel, A. B. (2006). Exemplar models, evolution and language change. *The linguistic review*, *23*(3), 247–274.

Winter, B., & Wedel, A. (2016). The co-evolution of speech and the lexicon: The interaction of functional pressures, redundancy, and category variation. *Topics in* *cognitive science*, *8*(2), 503–513.

Wright, R. (1997). Lexical competition and reduction in speech: A preliminary report. *Indiana University Research on Spoken Language Processing Progress Report*, *21*, 471-485.

Wright, R. (2004). Factors of lexical competition in vowel articulation. In J. Local, R. Ogden, & R. Temple (Eds.), *Papers in laboratory phonology vi* (pp. 26–50). Cambridge University Press Cambridge, England.

Yang, T.-H., Jin, S.-J., & Lu, Y.-A. (2019). The effect of mandarin accidental gaps on perceptual categorization. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), (p. 2022-2026). Canberra, Australia: Australasian Speech Science and Technology Association Inc.