Laboratory Phonology
Journal of the Association for Laboratory Phonology

Jin, S.-J., Wang, S.-F., & Lu, Y.-A. (2023). Identifying generalizable knowledge from the distribution of tonotactic accidental gaps in Mandarin. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 14(1), pp. 1–42. DOI: https://doi.org/10.16995/labphon.6455

Open Library of Humanities

Identifying generalizable knowledge from the distribution of tonotactic accidental gaps in Mandarin

Shao-Jie Jin, Department of Foreign Languages and Literatures, National Yang Ming Chiao Tung University, Hsinchu, Taiwan, shaojiejin.c@nycu.edu.tw

Sheng-Fu Wang, Institute of Linguistics, Academia Sinica, Taipei, Taiwan, sftwang@gate.sinica.edu.tw

Yu-An Lu*, Department of Foreign Languages and Literatures, National Yang Ming Chiao Tung University, Hsinchu, Taiwan, yuanlu@nycu.edu.tw

*****Corresponding author.

This study investigates tonotactic accidental gaps (unattested syllable-tone combinations) in Mandarin Chinese. In a corpus study, we found that, independent of syllable type, T2 (rising) and T3 (falling-rising) gaps were over-represented, whereas T1 (high level) and T4 (falling) gaps were under-represented. We also observed fewer T1 gaps with voiceless onsets and more T2 and T3 gaps with voiceless onsets, a pattern that is consistent with cross-linguistic observations. While these trends were generally reflected in a wordlikeness rating experiment by Mandarin listeners, their judgements of these gaps, similar to those of real words, were also guided by neighborhood density. Furthermore, T2 gaps with real-word T3 counterparts were rated as more wordlike, a result attributed to the T3 sandhi in Mandarin Chinese. Finally, we used harmonic scores generated from the UCLA Phonotactic Learner to explicitly test the role of lexical knowledge and markedness constraints in modeling speakers' tonotactic knowledge reflected in the wordlikeness ratings. We found that grammars induced from lexical data were the most successful at predicting wordlikeness ratings of gaps and lexical syllables combined. However, when focused on the ratings of tonotactic gaps, grammars with markedness constraints informed by cross-linguistic observations were more successful even without the constraints being weighted on lexical data. The results show how lexical knowledge and universal markedness, which is not entirely learnable from the lexicon, may account for some tonotactic generalizations.

1. Introduction

One of the central goals of phonology is to describe what structures are and are not possible in a given language (Fischer-Jørgensen, 1952; Halle, 1962). However, relatively less research has considered what seems to be possible yet does not exist. These unattested "accidental gaps" have traditionally been dismissed, considered possible and thus left unexplained, as opposed to systematic gaps, which violate systematic phonotactic constraints and thus are deemed impossible. Previous studies have shown that speakers' acceptance of these possible yet unattested forms is generally gradient and based on grammatical principles, such as *markedness* (e.g., Frisch, Pierrehumbert, & Broe, 2004; Zuraw, 2000, 2002), or *lexical statistics*, such as neighborhood density and the probability or frequency of attested forms (e.g., Albright & Hayes, 2003; Coleman & Pierrehumbert, 1997; Frisch et al., 2004; Gong & Zhang, 2021; Myers & Tsay, 2004). These studies, however, mainly focus on unattested segmental combinations. This study, on the other hand, explores unattested forms involving syllable-tone combinations, or "tonotactic accidental gaps" (Gong & Zhang, 2021; Lai, 2003; Wang, 1998), in Mandarin Chinese.[1] Since the processing of tone is distinct from that of segmental information (e.g., Cutler & Chen, 1997; Lee, 2007; Wiener & Turnbull, 2016), we investigate if the aforementioned generalizations (i.e., cross-linguistic grammatical principles and lexical statistics) drawn from unattested segmental combinations can also be applied to tonotactic accidental gaps. Using a corpus study and a wordlikeness rating experiment, we investigate whether the patterns observed in the corpus for gaps are reflected in Mandarin speakers' judgments of wordlikeness. Furthermore, given the finding that the phonetic naturalness of onset-tone interactions and lexical statistics both predicted speakers' judgments, we used computational modeling analysis to investigate their relationship further. Specifically, we asked whether and to what extent relevant knowledge of tone and onset-tone markedness is learnable from the lexicon using constraint-based learning simulation.

Standard Mandarin is generally described as having five vowels (/i, y, u, ə, a/) and 25 consonants with the maximum syllable structure (C)(G)V(G)/(C) (Lin, 2007). It has four phonemic tones: High-level Tone 1 (55), rising Tone 2 (35), falling-rising Tone 3 (214), and falling Tone 4 (51) (Duanmu, 2007; Lin, 2007). Tone numbers here indicate relative pitch height—the higher the number, the higher the relative pitch. Though less mainstream, some phonologists do not consider Mandarin to be a four-toneme language, treating the neutral tone as lexically specified (Chen & Xu, 2006; K. Huang, 2012). Not all tones, however, can be combined with every possible syllable. For example, the syllable [ts^{h}u] can be combined with T1 ([ts^{h}u]^{55} "coarse"), T2 ([ts^{h}u]^{35} "die" in Classical or Literary Chinese), and T4 ([ts^{h}u]^{51} "vinegar"), but not with T3. The syllable-tone combination *[ts^{h}u]^{214} does not violate any obvious phonotactic constraints in Mandarin, yet it fails to exist

[1] By using "tone-syllable" combination, we did not make any claims on the status of "tone" being separable from the rest of the syllable.

in any dictionary—this is an example of a tonotactic accidental gap (Duanmu, 2011; Lai, 2003). Because their occurrence seems random, tonotactic accidental gaps have been all but ignored in the literature. Note that, unlike segmental gaps in the aforementioned studies in which the number of possible unattested forms are hard to define, the number of tonotactic gaps can be easily calculated since the allowable syllables in Mandarin Chinese is straightforward. Thus, the tonotactic gaps reported in this study can also be understood as the inverse of actual Mandarin Chinese syllables (e.g., more T2 gaps means fewer actual T2 syllables). This study focuses on tonotactic accidental gaps to reveal relevant grammatical properties of the linguistic system. Motivated by studies demonstrating the importance of a priori grammar states and analytic biases independent of lexical statistics (e.g., Berent, Wilson, Marcus, Bemis, 2012; Becker, Nevins, & Levine, 2012) in modeling native speakers' phonotactic knowledge, we aimed to investigate whether speakers' differential preferences for unattested forms require knowledge that is either supplied by some a priori state (markedness informed by cross-linguistic observations) or from the attested lexicon (lexical statistics).

Among a handful of studies that have examined this issue, Wang (1998) asked native Taiwan Mandarin speakers to rate the wordlikeness of target syllables on a scale from 0 to 10, 0 indicating that the target syllable was very close to a real Mandarin word and 10 indicating that the target syllable was completely unlike a real word. The target syllables included tonotactic accidental gaps, phonotactic accidental gaps (phonotactically legal syllables that fail to exist), systematic gaps (phonotactically illegal syllables), and existing words. The results showed a clear distinction between existing and non-existing words, suggesting that tonotactic accidental gaps generally pattern together with phonotactically illegal syllables. However, Wang also noted that, among the non-existing words, accidental gaps (both tonotactic and phonotactic) were more readily accepted by native speakers compared to systematic gaps.

On the other hand, Myers and Tsay (2004) showed that native Taiwan Mandarin speakers' judgments of tonotactic accidental gaps in a wordlikeness rating experiment differed from those of phonotactically legal syllables and were judged similarly to systematic gaps. They concluded that phonotactics affects the judgement of both real words and non-words while frequency and neighborhood density only affect words. In an investigation of gap distribution and native speakers' judgements, however, Lai (2003) showed that there was an effect of tone frequency on non-words. In this study, tonotactic gaps with T2 were shown to be more common than those with the other tones. Moreover, T2 combined with closed syllables and [p, t, k, tɕ, tʂ, ts] onsets and T1 with [m, n, l, ʐ̩] onsets accounted for a large proportion of gaps. Lai further conducted rating and preference experiments investigating native Taiwan Mandarin speakers' judgments of tonotactic accidental gaps and found that T4 gaps were more readily accepted as real Mandarin words compared with T2 gaps, and T2 gaps with [p, t, k, tɕ, tʂ, ts] onsets were generally disfavored. These results were attributed to tone frequency: There are more real words with T4 than T2, and there are more T2 gaps with those particular onsets.

In a more recent study, Gong and Zhang (2021) collected native Mandarin speakers' well-formedness judgments of five types of T1 syllables—real words, tonotactic gaps, allophonic gaps (gaps that only violate allophonic rules; e.g., vowel backness depending on the place of the nasal codas), phonotactic accidental gaps, and systematic gaps. They found that the five different types of stimuli were rated gradiently: Real words were considered more well-formed than tonotactic gaps, followed by allophonic gaps, phonotactic accidental gaps and finally systematic gaps. Furthermore, the judgments were positively correlated with neighborhood density, and this effect was found to be stronger for gaps than for real words.

The results from such behavioral experiments with gradient measurements of phonotactic probability or well-formedness can be computationally modeled. The UCLA Phonotactic Learner (Hayes & Wilson, 2008), which induces grammars consisting of weighted constraints based on the principle of Maximum entropy (Della Pietra, Della Pietra, & Lafferty, 1997; Goldwater & Johnson, 2003; Hayes & Wilson, 2008; Zuraw & Hayes, 2017), has been extensively used for this purpose (e.g., Berent et al., 2012; Daland, Hayes, White, Garellek, Davis, & Norrmann, 2011; Gallagher, Gouskova, & Camacho Rios, 2019; Goldwater & Johnson, 2003; Hayes & White, 2003; Wilson & Gallagher, 2018). Gong (2017), for example, used this method to model visual lexical decisions on segmental combinations in Mandarin Chinese. Gong and Zhang (2021) also used the learner to model the wordlikeness ratings of Mandarin word forms from different lexicality categories. Alternatives to the UCLA Phonotactic Learner in modeling lexical judgements in Mandarin include the probability of segmental strings (Myers & Tsay, 2005) and Bayesian probabilities (Do & Lai, 2020). In this study, we complement these works by modeling tonotactic generalizations with the UCLA Phonotactic Learner to compare grammars with different tonotactic constraints, namely *inductive constraints* with different levels of fit to the lexicon and *typologically-motivated markedness constraints* with or without access to the lexicon (i.e., if the weights are informed by learning simulations using the lexicon). These comparisons allow us to examine to what extent the effects observed in the behavioral data are learnable from the lexical data.

In the following sections, we give a comprehensive description of Mandarin tonotactic accidental gaps. First, we conduct a corpus study to examine the distribution of all lexical segmental syllables, that, when combined with the four lexical tones, yield non-lexical syllables. The results show that, independent of syllable type, T2 (rising) gaps are over-represented. Since T2 and T3 are intrinsically more marked than T1 and T4 in terms of contour complexity and aerodynamics (see Section 2), and the phonetic realization of T2 and T3 contours requires a longer duration (Zhang, 2001), we further investigate to what extent would the T2 overrepresentation and T2/T3 markedness reflect on speakers' wordlikeness rating. The results reveal that T2 is not disfavored while T3 is generally disfavored independent of syllable structures. Furthermore, native speakers' wordlikeness ratings of gaps are gradient and heavily guided by neighborhood density. While speakers' wordlikeness judgment does reflect the markedness of T3 with a falling-rising contour,

the over-representation of T2 gaps from the lexicon is not similarly evident. Motivated by the mismatches between the statistical properties of the lexicon and speakers' judgments, we use the UCLA Phonotactic Learner as a computational tool to incorporate and compare different degrees of lexical access in modeling wordlikeness ratings. We find that while tonotactic constraints induced from the lexical data can successfully model the results overall, typologically-motivated markedness constraints are better at predicting which gaps receive higher ratings, and their success could largely be achieved independent of the lexicon. The modeling results suggest that speakers' tonotactic knowledge may be disassociated from statistical patterns in the lexicon.

2. Corpus study of Mandarin tonotactic accidental gaps

To examine if the possible yet unattested Mandarin tonotactic accidental gaps follow any particular pattern, we first investigate the distribution of these gaps by compiling a corpus of gaps, which we named the 'Mandarin Accidental Gap Corpus'. The corpus included the 398 allowable Mandarin syllables (taken from Lin, 2007, p. 283). Two definitions of accidental gaps were employed: (1) A narrow view, in which syllable-tone combinations do not exist as a lexical syllable in the *Revised Mandarin Chinese Dictionary*, compiled by the Ministry of Education, Taiwan (https://dict.revised.moe.edu.tw/), and (2) a broad view, where syllable-tone combinations *do* form lexical syllables but have zero-frequency in the *Taiwan Mandarin Conversational Corpus* (TMC corpus) (Tseng, 2019).[2] For example, the syllable [ts^{h}u] in T2 ([ts^{h}u]^{35} "die" in Classical or Literary Chinese), historically exists as a lexical syllable and is known by Mandarin speakers through poetry but is not listed. As such, this lexical syllable might be considered a gap by native Mandarin speakers because it is rarely, if ever, used in spoken Mandarin. This lexical syllable was thus counted as a gap in the broad view but not in the narrow view. Note that we use "lexical syllable" here and throughout this work instead of "word" because while Mandarin morphemes are mostly monosyllabic, around 72% of the lexicon is made up of disyllabic words (Li, 2013).

Our investigation of the corpus data revealed that accidental gaps were not evenly distributed across the four tones, as shown in **Figure 1**. In the narrow view, a one-way chi-square test showed that T2 gaps were over-represented while T4 gaps were under-represented ($\chi^2(3) = 68.8$, $p < .001$). Another one-way chi-square test revealed that T2 gaps were over-represented in the broad view ($\chi^2(3) = 25.01$, $p < .001$). In the aforementioned study, Lai (2003) made a similar observation that T2 gaps outnumbered gaps with the other tones.

There are several possible explanations for the asymmetrical distribution of accidental gaps. First, it could be attributed to *Zhuó Shǎng Biàn Qù* (voiced *shǎng* tone entering *qù* tone), a historical tone merging process in which a number of voiced *shǎng* tones (i.e., T3) merged into *qù* tones

[2] We referred to a Taiwan-based conversational corpus in this study because Taiwan Mandarin participants were recruited for our wordlikeness rating experiment. We did not assume a substantial difference in the generalizations drawn here and those drawn from Putunghua.

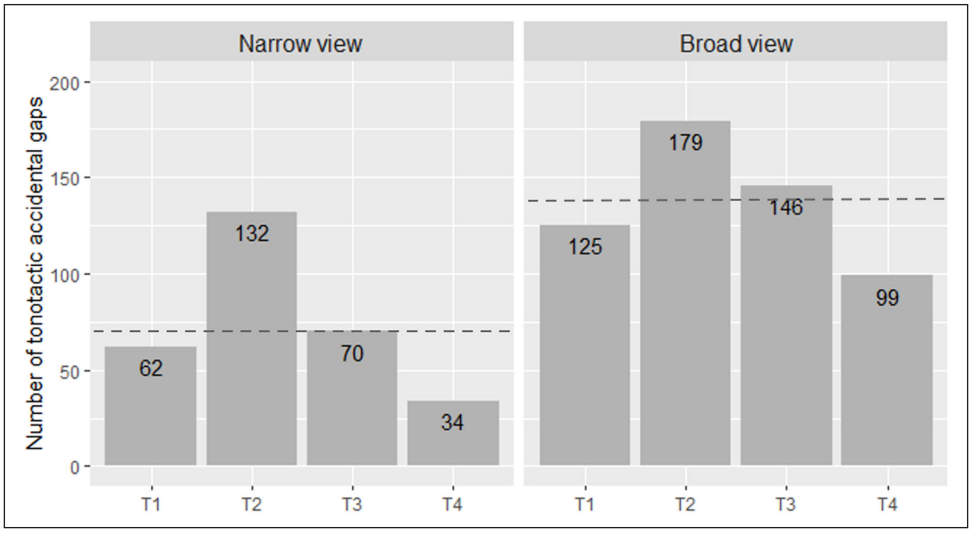

Figure 1: Numbers of Mandarin accidental gaps as a function of tone in the narrow and broad views. The horizontal lines indicate the predicted numbers.

(i.e., T4) in Middle Chinese (Mei, 1970, 1977; Wang, 1972). This may account for the lower number of T4 gaps. Second, the large percentage of T2 gaps could be attributed to the markedness of rising (i.e., T2) in comparison with level and falling tones (i.e., T1 and T4). Specifically, among simple contour tones, falls are much more common than rises, presumably due to the physiological difficulty associated with the production of rising contours against natural airflow dynamics (Zhang, 2001). Despite the fact that T3 (falling-rising) involves the most complex contour, we did not observe any obvious overrepresentation of T3 gaps except for the greater number of T3 gaps observed in the broad view compared with those in the narrow view. This may be attributed to the marked status of the complex tonal contour or to its greatly confusable nature with T2 due to phonetic similarity and a morphophonemic alternation involving T3 sandhi (Hao, 2012; T. Huang, 2001; Huang & Johnson, 2010; Hume & Johnson, 2003; Mei, 1977). This speculation is not without grounding as T3 sandhi has indeed emerged within the past few centuries (Mei, 1977).

It should be noted that, cross-linguistically, contour tones are generally preferred in longer rimes, presumably because they provide a duration long enough to realize the complex tone targets (Zhang, 2000, 2001). Studies have shown that Mandarin CGVN syllables are indeed longer than other syllable types (i.e., CV, CVN, CGV) (Wu & Kenstowicz, 2015) and that T2 and T3 are longer than T1 and T4 (Lu & Lee-Kim, 2021; Wu & Kenstowicz, 2015). We thus divided the syllables into different types (CV/CGV, CVG/CGVG, CVN/CGVN) to examine if there were any tone-syllable type dependencies. Here we follow a conventional definition of syllable structure in Mandarin, assuming that on-glides are grouped with the onset and thus do not contribute to syllable weight (Duanmu, 2007). However, as the results in **Figure 2** show, we did not observe any tendencies in this direction (narrow view: $\chi^2(6) = 8.05$, $p = .24$; broad view: $\chi^2(6) = 4.11$, $p = .66$).

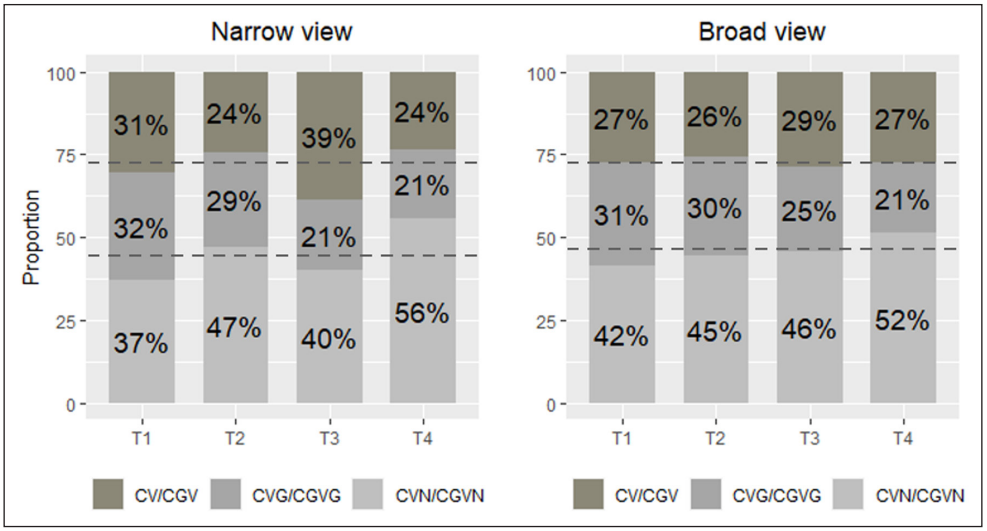

Figure 2: Proportions of Mandarin tonotactic accidental gaps as a function of syllable types in the narrow and broad views. The horizontal lines indicate the predicted proportions.

Cross-linguistic and diachronic studies have observed that high-*f0* tones are more compatible with voiceless onsets while low-*f0* tones are more compatible with voiced onsets (e.g., Hsieh & Kenstowicz, 2008; Kenstowicz & Suchato, 2006; Ohala, 1978; Sagart, 1999; Yip, 2002). This can be explained by the aerodynamics involved in articulation in that voiceless consonants exert a pitch-raising effect on the following tone (Hombert, Ohala, & Ewan, 1979; Ohala, 1978). It is generally agreed upon that some Chinese tones originated via a similar mechanism. Sagart (1999) reports a clear correspondence between onset voicing in Middle Chinese and tones in Modern Chinese—voiced onsets, both obstruents and sonorants, induced a tone lowering of the following vowel resulting in high and low allotones that eventually phonologized into different tonal contrasts. We thus examined if T1 and T4, tones with an initially high pitch, were more likely to appear with voiceless onsets (i.e., having fewer gaps), and if T2 and T3, tones with initially low pitch, were more likely to appear with voiced onsets. In other words, there should be fewer T1 and T4 gaps coupled with voiceless onsets and T2 and T3 gaps with voiced onsets. The contrast between voiced and voiceless consonants is essentially obstruent vs. sonorant since Mandarin obstruents lack a voicing contrast.[3] The results (**Figure 3**) showed that the distribution of gaps mostly conformed to this trend: More gaps with voiced onsets were observed in T1 than in T2 and T3, in which gaps with voiceless onsets dominated (narrow view: $\chi^2(3) = 80.37$, $p < .001$; broad view: $\chi^2(3) = 64.52$, $p < .001$). However, T4 gaps did not pattern as predicted. Although the pitch of T4 is initially high, there were still more T4 gaps with voiceless onsets. This may be attributed to the historical tone merging process mentioned earlier whereby T3, presumably with more voiced onsets, merged into T4, disrupting the connection between onset voicing and tone.

[3] Note that the Mandarin rhotic is variably treated as a voiced obstruent [ʐ] or an approximant [ɹ] (Duanmu, 2007; Lin, 2007).

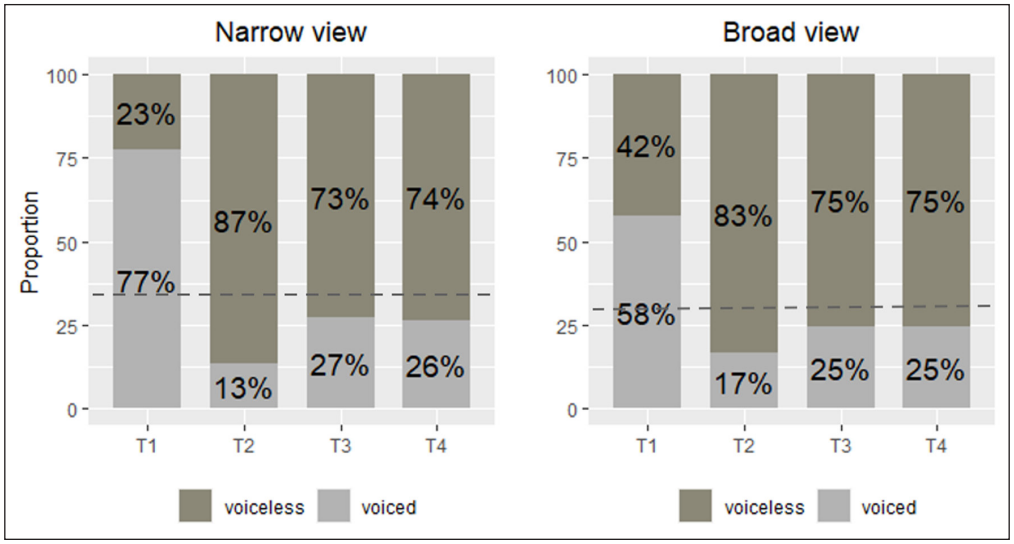

Figure 3: Proportions of Mandarin tonotactic accidental gaps as a function of onset voicing in the narrow and broad views. The horizontal lines indicate the predicted proportions.

Recall that Lai (2003) reported that T2 closed syllables with [p, t, k, tɕ, tʂ, ts] onsets and T1 with [m, n, l, ʐ] onsets accounted for a large proportion of gaps. When we focused our analysis on individual onsets, we found that onset voicing, again, is a better indicator of the general pattern, as shown in **Figure 3**.

The analyses of our corpus data were aimed at determining if the accidental gaps in Mandarin follow any particular pattern. Our findings suggest that the occurrence of gaps is not completely random. We found that T2 gaps are over-represented while the most marked T3 gaps are not. We also found more gaps with voiced onsets in T1 than in all other tones, in which gaps with voiceless onsets dominated, a pattern that is partially observed cross-linguistically and diachronically.

In the next section, we investigate Mandarin speakers' judgments of these accidental gaps. Specifically, we conducted a wordlikeness judgement experiment to explore whether their judgements of accidental gaps would follow the same tendencies that we observed in our corpus study and/or by grammatical principles that were absent from the corpus.

3. Wordlikeness judgment experiment

We conducted a wordlikeness judgement experiment to investigate Mandarin listeners' perception of tonotactic accidental gaps and to determine if their perceptual tendencies reflect what has been observed both cross-linguistically and in our corpus study. The factors being examined along with our predictions are summarized in **Table 1**.

Observation	Possible explanations
a. More T2 gaps than T1/T4 gaps observed in the corpus	T2/T3 are more marked than T1/T4.
→ Are T2 gaps judged as less wordlike? → Is the most marked T3 judged as less wordlike?	
b. T2/T3: More gaps with voiceless onset T1: Fewer gaps with voiceless onset Note: T4 does not conform to this pattern due to a historical tone merging process.	High-*f0* is more compatible with voiceless segments while low-*f0* is more compatible with voiced segments.
→ Are T2/T3 gaps with voiceless onset and T1 gaps with voiced onset judged as less wordlike?	
c. Lexical statistics: Neighborhood density, frequency, and phonotactic probability	Previous studies have found effects of frequency, neighborhood density and phonotactic probability effects on unattested forms (Albright, 2003; Coleman & Pierrehumbert, 1997; Frisch, Large, & Pisoni, 2000; Gong & Zhang, 2021; Lai, 2003; Myers & Tsay, 2005).
→ Do lexical statistics have a gradient effect on the wordlikeness judgment of gaps?	

Table 1: Factors that may affect the wordlikeness judgments of Mandarin tonotactic accidental gaps.

3.1. Methodology

3.1.1. Participants

Thirty-seven Taiwan Mandarin native speakers (10 male, 27 female; aged 20–37, $M = 21.68$) were recruited from National Yang Ming Chiao Tung University. These participants were all Mandarin-dominant speakers with some exposure to other dialects of Chinese spoken in Taiwan (Taiwanese Southern Min and Hakka). None of the participants reported hearing or speaking deficiencies. The study was conducted in accordance with the ethical guidelines approved by the Research Ethics Committee for Human Subject Protection, National Yang Ming Chiao Tung University. All participants were compensated monetarily for their time.

3.1.2. Materials

To examine whether the patterns observed in the corpus and cross-linguistically (**Table 1**) are reflected in native Mandarin speakers' judgments of accidental gaps, 96 Mandarin accidental gaps (as defined by both the narrow and broad views) were selected. The gaps were counterbalanced across the four Mandarin tones and different syllable types (open: CV, CGV; closed: CVN, CGVN). Another 48 Mandarin lexical syllables, referred to as "words" in the figures for the sake of brevity, fulfilling the same criteria were also selected. These stimuli were selected such that they represented the distribution in the corpus in terms of onset voicing and tone combinations

(e.g., more voiced-onset T1 gaps, more voiceless-onset T2 gaps).[4] The 144 stimuli (see Appendix I) were produced by a male native speaker of Taiwan Mandarin. Though previous studies have shown that including real words can de-sensitize participants' ratings of non-words and is more likely to activate lexical neighbors than when all stimuli are non-words (Albright, 2009), we included lexical syllables to enable a comparison between the gaps and lexical syllables.

The realization of the phonetic tonal contours of these naturally produced stimuli were checked to ensure that the gap and lexical tokens were comparable. Time normalized *f0* contours of these tokens (excluding obstruent onsets, if any) were obtained using ProsodyPro (Xu, 2013). As seen in **Figure 4**, the tonal contours were comparable between the gap and lexical tokens. Note that the final rise of T3 (falling-rising) contours fell short of the final rise target, a well-known characteristic of T3 production in Taiwan Mandarin (Fon & Chiang, 1999; Kubler, 1985). Despite the final-rise undershoot of T3, Lu and Lee-Kim (2021) showed that this tone is still perceived as having a complex fall-rise contour. All else being equal, Taiwan Mandarin speakers perceived T3 tokens without final rise as the longest among the four lexical tones. Furthermore, when asked to imitate T3 with a final-rise undershoot, these speakers implemented a final rise, similar to that in T3 with a full concave contour.

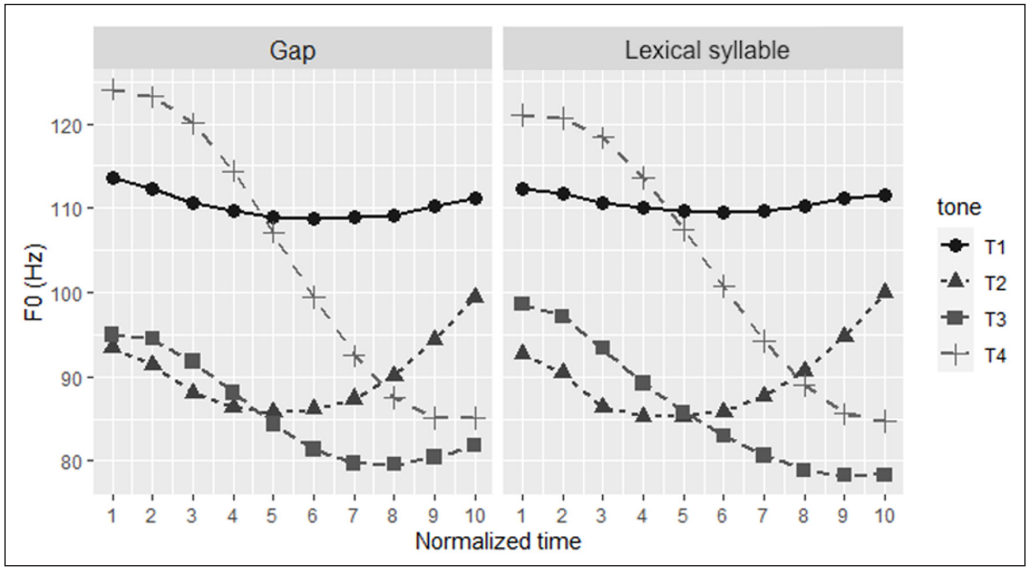

Figure 4: Time normalized f0 as a function of tone paneled by lexicality.

The durations of gap syllables were longer (M = 470.33 ms, SD = 111.15 ms) than those of lexical syllables (M = 441.47 ms, SD = 106.89 ms). Since this difference in the naturally produced stimuli could have confounded the wordlikeness ratings, the stimuli were further resynthesized

[4] The distribution of onset voicing according to tone is listed here: T1 voiced = 13, voiceless = 11; T2 voiced = 6, voiceless = 18; T3 voiced = 6, voiceless = 18; T4 voiced = 6, voiceless = 18.

into two durations, 300 ms and 500 ms, reflecting the range of Mandarin syllable duration (Lu & Lee-Kim, 2021; Wu & Kenstowicz, 2015), using the Pitch Synchronous Overlap and Add (PSOLA) algorithm in Praat (Boersma & Weenink, 2017). The manipulation ensured that any differences in the gap and word ratings would be unlikely to be due to any acoustic artifacts of the stimuli.

3.1.3. Procedure

The 288 stimuli ([96 accidental gaps + 48 lexical syllables] × 2 durations) were randomized for each participant and presented auditorily in three blocks using E-Prime software (Schneider, Eschman, & Zuccolotto, 2012). The participants were instructed with written instructions on a computer screen (*Qǐngwèn nín tīngdào de zì yǒu duō xiàng zhōngwén?* "How Mandarin-like is the word you just heard?") to rate each word on a 7-point scale, with 7 being the most wordlike and 1 the least wordlike.[5] Nine practice trials were presented before the experiment to familiarize participants with the task. Participants were tested individually in a sound-attenuated booth using AKG K240 headphones and their responses were recorded using E-Prime. The total duration of the experiment was around 15 minutes.

3.2. Results

Linear mixed-effects regression models were fitted in R using the *lme4* package (Bates, Maechler, Bolker, & Walker, 2015) and *p*-values were obtained using the *lmerTest* package (Kuznetsova et al., 2016). The visualizations were plotted using the *ggplot2* package (Wickham, 2009). Models were fitted with the participants' wordlikeness ratings on the 7-point scale converted into *z*-scores for each speaker as the dependent variable. For our analyses, the experimental variables of interest included *Tone* (4 levels), *SyllableType* (open vs. closed), *OnsetVoicing* (voiced vs. voiceless), and *Lexicality* (tonotactic gap vs. lexical syllables). A set of variables on lexical statistics was also included. For a balanced comparison of the tonotactic gaps and lexical syllables, we used *SyllableFrequency* and *SyllableGapFrequency* as the indices to calculate the effect of frequency, if any.

The calculation of *SyllableFrequency* was straightforward; we calculated the overall token frequency of each syllable regardless of tone and morphemes using the TMC corpus (Tseng, 2019). We grouped homophonic morphemes together and only considered syllable token frequency since an auditory experiment had been employed. *SyllableGapFrequency* is the inverse of tonal neighborhood density, calculated by the number of lexical syllables differing from the test item only in tone. That is, with only one gap in a certain syllable, "1" would be considered more frequent while "3", with three gaps, would be considered less frequent. Note that there is no "4" because in this case all tone-syllable combinations would be impossible.

[5] One might question the possibility that the reading of "zhōngwén", literally meaning "Chinese", could refer to a Chinese dialect other than Mandarin. This reading is unlikely due to the fact that these participants were Mandarin-dominant and that in Taiwan, Taiwanese Southern Min is referred to as "tái yǔ" and Hakka as "kè yǔ". The term "zhōngwén" almost exclusively refers to Mandarin in Taiwan.

NeighborhoodDensity and *PhonotacticProbability* were included to provide additional quantification of possible lexical influence. *NeighborhoodDensity* was calculated by the summed frequency of the words generated by adding, deleting, or substituting a single phoneme. In this calculation, we treated diphthong vowels as sequences of two phonemes (e.g., [a], [i], and [ei] as neighbors of [ai]). Note that we used a tone-blind *NeighborhoodDensity* since the previously mentioned *SyllableGapFrequency* variable already reflected the number of syllables differing in tones. Finally, *PhonotacticProbability* was defined by onset-rime transitional probability (Tseng, 2019).

In addition to the lexical statistics variables (*SyllableFrequency*, *NeighborhoodDensity*, *SyllableGapFrequency* and *PhonotacticProbability*), the model also included the *Tone*OnsetVoicing* interaction to examine if there was any correlation between *Tone* and the two factors (**Table 1(a, b)**) as well as the *Tone*Lexicality* interaction to determine if gaps, like lexical syllables, were rated based on lexical statistics (**Table 1(c)**). The model also included the random intercepts for *Participant* and *Item* as well as by-participant random slopes for *Tone*, *OnsetVoicing*, and *NeighborhoodDensity*. Models including other by-participant random slopes failed to converge.

Descriptions of each variable and how they were coded are listed in **Table 2**. T2, which yielded intermediate wordlikeness ratings, was set as the reference level to facilitate the interpretation of the results. The binary variables *SyllableType*, *OnsetType*, and *Lexicality* were contrast coded so the sum of the weight of each level would be 0 so we could interpret the results as main effects (Davis, 2010).

Variable	Description	Coding
WordlikenessRating	1–7 rating scale transformed into z-score	Numerical
Tone	T1, T2, T3, T4	4 levels: T2 as reference
SyllableType	Open vs. closed	2 levels: –1 vs. 1
OnsetVoicing	Voiceless vs. voiced	2 levels: –1 vs. 1
Lexicality	Gap vs. lexical syllable	2 levels: –1 vs. 1
SyllableFrequency	Z-scored log transformed token frequency of each syllable regardless of tone	Numerical
SyllableGapFrequency	The inverse of tonal neighborhood density	Numerical
NeighborhoodDensity	Z-scored summed frequency of the words generated by adding, deleting, or substituting of a single phoneme	Numerical
PhonotacticProbability	Z-scored onset-rime transitional probability	Numerical
Participant	Participant ID	Factorial
Item	Test item	Factorial

Table 2: Variables considered for analysis in wordlikeness rating experiment.

The statistical model is summarized in **Table 3**. As would be expected, Mandarin speakers generally rated lexical syllables as more wordlike than gaps (*Lexicality*: $p < .0001$). In the following, we discuss each of the factors and interactions that were relevant to patterns reported in the corpus study and previous cross-linguistic observations.

	R^2 = .55			
	B	**SE**	**t**	**p**
(Intercept)	0.29	0.11	2.75	.007
T1	0.08	0.08	1.02	.311
T3	**–0.22**	**0.08**	**–2.74**	**.007**
T4	0.06	0.08	0.77	.445
OnsetVoicing	–0.03	0.06	–0.57	.567
Lexicality	**0.44**	**0.09**	**4.69**	**<.0001**
SyllableFrequency	0.06	0.05	1.34	.184
SyllableGapFrequency	–0.03	0.04	–0.81	.418
SyllableType	0.02	0.02	0.90	.371
NeighborhoodDensity	**0.11**	**0.05**	**2.24**	**.027**
PhonProbability	0.02	0.07	0.25	.802
T1:OnsetVoicing	**0.18**	**0.08**	**2.27**	**.025**
T3:OnsetVoicing	0.03	0.08	0.45	.657
T4:OnsetVoicing	0.04	0.07	0.59	.558
T1:Lexicality	–0.10	0.07	–1.37	.174
T3:Lexicality	–0.04	0.07	–0.53	.594
T4:Lexicality	–0.07	0.07	–1.03	.306
Lexicality:SyllableFrequency	0.00	0.05	–0.04	.967
Lexicality:SyllableGapFrequency	**0.10**	**0.04**	**2.67**	**.008**
Lexicality:NeighborhoodDensity	0.05	0.05	1.10	.273

Table 3: Summary of the statistical model for the wordlikeness judgment experiment.
Model: Wordlikeness rating ~ Tone * OnsetVoicing + Tone * Lexicality + Lexicality * SyllableFrequency + Lexicality * SyllableGapFrequency + Lexicality * NeighborhoodDensity + PhonotacticProbability + SyllableType + (1 + Tone + OnsetVoicing + NeighborhoodDensity | Participant) + (1 | Item).

3.2.1. Corpus observation: More T2 and T3 gaps than T1 and T4 gaps

One of the main observations from our corpus study was that there were more T2 gaps than gaps of other tones (see **Figure 1**). One goal of this experiment was to determine if this pattern would

also be observed in Mandarin native speakers' wordlikeness ratings. That is, would Mandarin speakers rate T2 gaps as less wordlike than the T1, T2, and T4 gaps? The results are graphed in **Figure 5**, which shows that T3 syllables, instead of T2 syllables, among gaps and real words were rated as the least wordlike, as indicated by the significant T3 effect ($p = .007$; **Table 3**). Post-hoc tests using the *emmeans* package (Lenth, Singmann, Love, Buerkner, & Herve, 2019) showed that, though T2 was rated as less wordlike than T1 and T4, the ratings for T2, T1, and T4 did not significantly differ. These patterns held true in both words and gaps, as indicated by the lack of a *Tone*Lexicality* interaction (all $p > .05$).

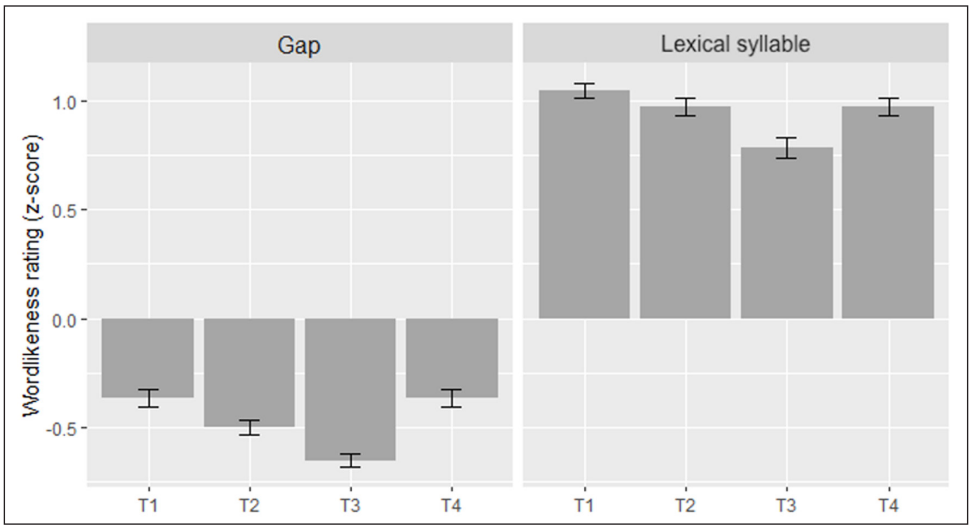

Figure 5: Standardized wordlikeness ratings as a function of tone and lexicality.

These findings diverge from the patterns observed in the corpus in the following ways. First, T2 gaps, rather than T3 gaps, were found to be over-represented. If the wordlikeness ratings strictly followed the pattern observed in our corpus study, we should have seen T2 gaps judged as the least wordlike. Instead, T3 gaps were rated as the least wordlike. Second, Mandarin speakers also rated T3 *lexical syllables* as less wordlike than lexical syllables with other tones, a pattern that also diverged from the Mandarin speakers' linguistic experience, as there were fewer T2 lexical syllables in the corpus (**Figure 6**). Furthermore, T2 is the least frequent tone for lexical syllables, as indicated by a calculation of token and type frequency (again, based on syllables, not morphemes) of Mandarin tones using the TMC corpus (Tseng, 2019) (**Table 4**). As such, the aversion to T3 lexical syllables cannot be attributed to the linguistic experience of the Mandarin native speakers. One possible explanation is that T3, being a complex contour tone, is more marked than the other tones in Mandarin, which may lead to it seeming less wordlike.

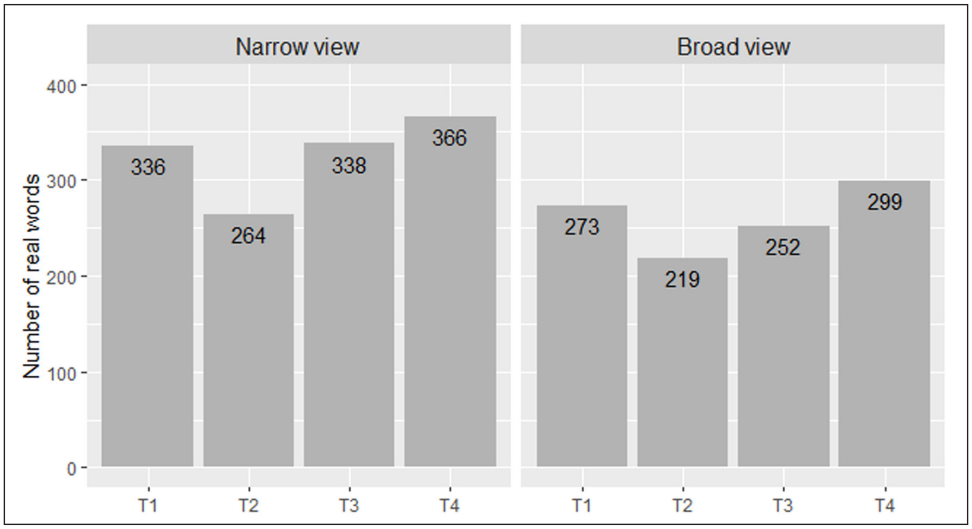

Figure 6: Number of existing syllables as a function of tone from the corpus study.

	Token frequency	Type frequency
T1	105168	272
T2	96586	220
T3	129505	250
T4	228182	301

Table 4: Tone frequency in real words from TMC corpus (Tseng, 2019).

In our corpus study, we also found more T3 gaps as the definition of gaps was shifted from the broad to the narrow view relative to the other tones. We speculated that the large number of T3 gaps might have arisen from avoiding the confusability between T2 and T3, as a T3 becomes a T2 before another T3 in a tone sandhi process. To explore this idea in terms of wordlikeness judgments, we compiled a subset of T2 and T3 gaps from the data comparing how Mandarin speakers rated T2 and T3 gaps whose T3 or T2 counterparts were *not* gaps as opposed to the T3 and T2 gaps with counterparts that were also gaps. For example, T3 [tsuŋ]^{214} 'always' is a lexical syllable, but T2 *[tsuŋ]^{35} is a gap; however, Mandarin speakers would still have experience with T2 *[tsuŋ]^{35} as a sandhi form of T3 [tsuŋ]^{214}. In contrast, both T2 *[ɻɻ̩]^{35} and T3 *[ɻɻ̩]^{214} are gaps, so Mandarin speakers would not have been exposed to either form. **Figure 7** shows the results of this analysis, which indicates that there was indeed a general tendency for T2 gaps with T3 lexical syllable counterparts to be given higher wordlikeness ratings. This suggests that T2 gaps may have been interpreted as a sandhi-ed T3, thereby improving their wordlikeness ratings. This trend, however, was not observed with T3 gaps with T2 lexical syllable counterparts, since

surface T3 cannot be derived from T2 by any sandhi process in Mandarin. Chien et al. (2017) observed a similar asymmetrical pattern: Presenting a T3 prime facilitated a lexical decision for a T2 (underlyingT3)-T3 disyllabic word, while presenting a surface T2 prime did not facilitate a T2 (underlyingT3)-T3 disyllabic word. Our analysis of these subset data suggests that there is a close relationship between T2 and T3 in a direction that can be predicted by the T3 sandhi in Mandarin Chinese. The higher wordlikeness ratings of T2 gaps may partially explain why T3 words and gaps were rated as less wordlike overall (cf. Section 3.2.1).

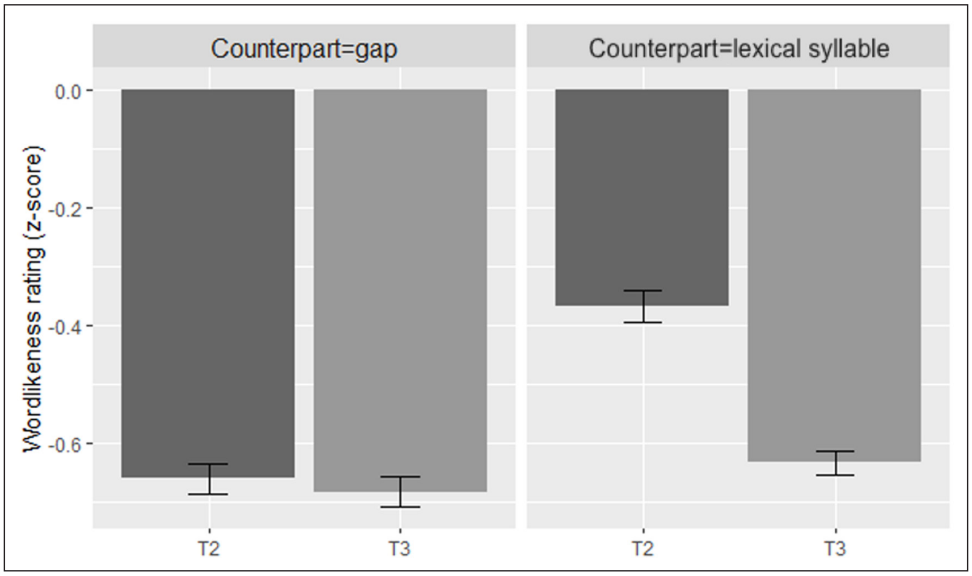

Figure 7: Standardized wordlikeness ratings on T2/T3 gaps in which the T3/T2 counterparts are either gaps or real words.

3.2.2. Corpus observation: More T2 and T3 gaps with voiceless onsets but fewer T1 gaps with voiceless onsets

We found in our corpus study, and others have observed diachronically and cross-linguistically, that T2 and T3 syllables, with initially low *f0*, are less compatible with voiceless onsets giving rise to more T2 and T3 gaps with voiceless onsets. In contrast, T1, with an initially high *f0*, is more compatible with a voiceless onset and thus there are fewer T1 gaps with voiceless onsets. In our wordlikeness rating experiment, we found a significant *Tone*OnsetVoicing* interaction driven by the higher wordlikeness ratings of T1 gaps with voiceless onsets (*T1*OnsetVoicing*, $p = .025$; **Table 3**), as shown in **Figure 8**. Post-hoc tests using the *emmeans* package (Lenth et al., 2019) confirmed no other *Tone*OnsetVoicing* interactions. However, the same was not observed for T2 and T3 gaps with voiced onsets, despite the compatibility between these tones with voiced onsets. This finding could be attributed to the general disfavoring of T2 and T3.

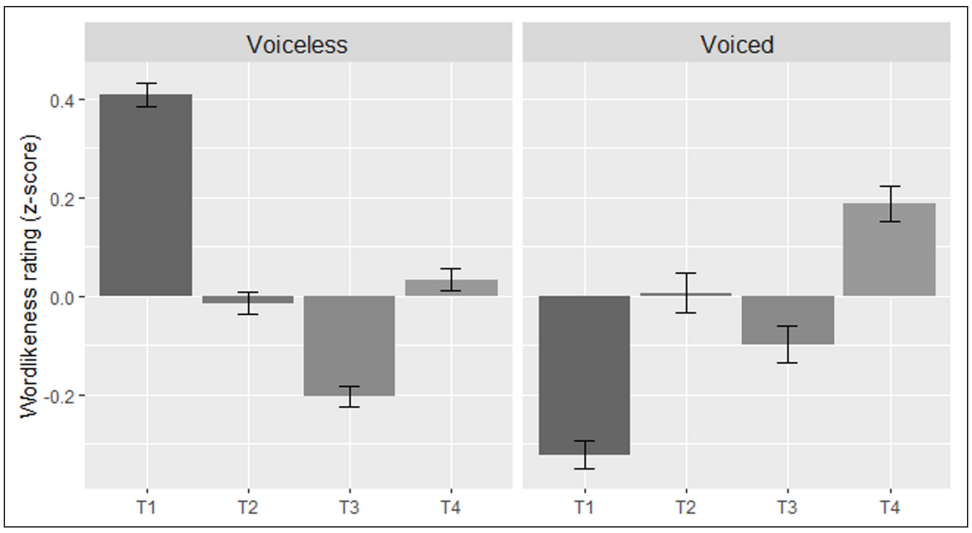

Figure 8: Standardized wordlikeness ratings as a function of onset voicing.

3.2.3. The effect of lexical statistics

Previous studies have demonstrated effects of lexical statistics for real words and unattested forms. Here, we aimed to determine if such effects would be present in the wordlikeness judgements of both gaps and lexical syllables. We found a significant *NeighborhoodDensity* effect ($p = .027$; **Figure 9**), indicating that the more neighbors the syllable had, the more wordlike it was judged. These effects on lexical syllables and gaps were comparable, as suggested by the lack of interactions with *Lexicality*. These findings are in line with those in Lai (2003) and Gong and Zhang (2021). A *SyllableGapFrequency*Lexicality* interaction ($p = .008$; **Figure 10**) was found, since only gap syllables were judged as more wordlike the more tonal neighbors they had. No *SyllableFrequency* or *PhonotacticProbability* effect was found.

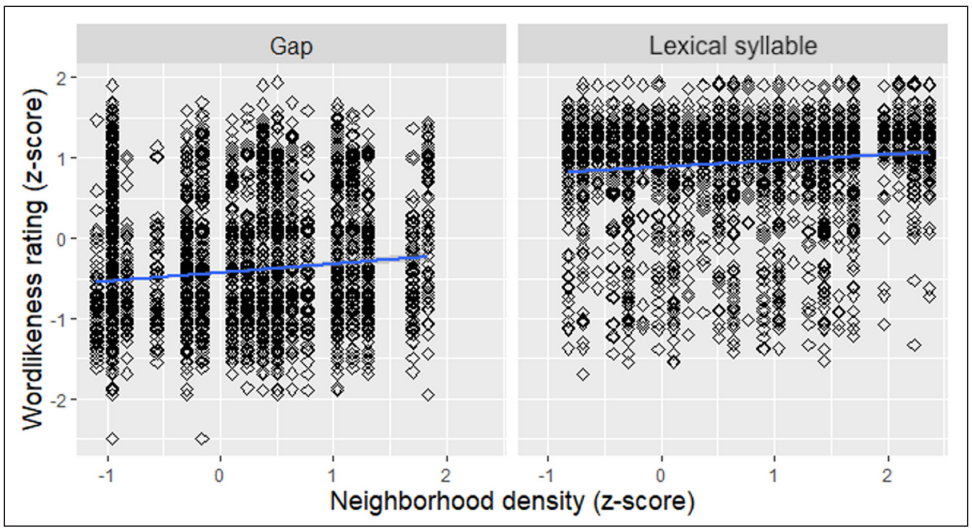

Figure 9: Standardized wordlikeness ratings as a function of standardized neighborhood density.

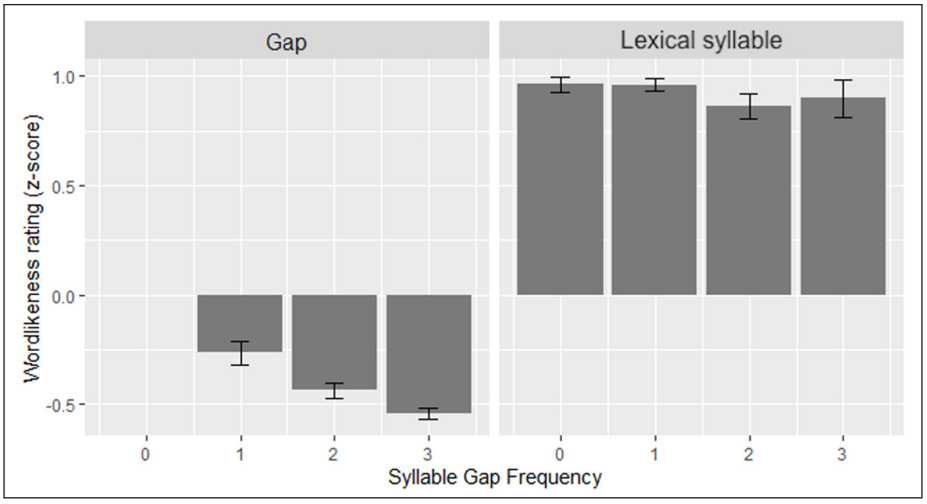

Figure 10: Standardized wordlikeness ratings as a function of Syllable Gap Frequency.

3.3. Summary

Based on the results of the wordlikeness judgement task, the patterns found in Mandarin speakers' perception of accidental gaps did not entirely match those found in our corpus study or other cross-linguistic studies. The Mandarin speakers rated T3 as less wordlike than T2, T1, and T4, both for gaps and lexical syllables, a result that was not predicted based on the patterns observed in our corpus study alone. We attributed this to the marked complex contour of T3. Furthermore, T2 and T3 are confusable due to the aforementioned tone sandhi process. Mandarin listeners' experience with T3 sandhi may have caused T2 gaps to be considered more acceptable to some degree, particularly when their T3 counterparts were real words. Mandarin listeners' judgements were not affected by syllables type. In fact, gaps with different syllable types were rated comparably. We did, however, find that T1 gaps with voiceless onset were judged as more acceptable, a pattern that was also observed in our corpus study and cross-linguistically.

We also found that, similar to lexical syllables, gap syllables were affected by neighborhood density—the more neighbors the syllable had, the higher the ratings.

4. Modeling wordlikeness with phonotactic grammars

In this section, we model the results of the wordlikeness experiment with constraint-based grammars using the UCLA Phonotactic Learner (Hayes & Wilson, 2008) to explore the extent to which Mandarin lexicon is useful in establishing a phonotactic grammar that models the speakers' tonotactic knowledge. To answer this question, we generated tonotactic constraints in three settings according to the degree of lexical access. In the first setting, we built the tonotactic grammar with the learner *entirely* from the inductive process, reflecting the view that the lexicon

itself is sufficient as the source of phonotactic knowledge. In the second setting, we used the lexicon and the inductive process to select and weight a small set of typologically motivated constraints, reflecting the view that phonotactic knowledge is built based on a smaller hypothesis space in which inductive learning also plays a role. In the third setting, the tonotactic grammar also contains typologically motivated constraints but *makes no reference* to the lexicon and the inductive process, reflecting the view that innate constraints are separate from lexical knowledge. These settings enable us to more accurately evaluate the role of lexical statistics in shaping native speakers' tonotactic knowledge as revealed through the wordlikeness ratings.

The UCLA Phonotactic Learner is an inductive learning tool that takes a wordlist as input and yields a constraint-based phonotactic grammar based on the principle of Maximum Entropy (Della Pietra et al., 1997; Goldwater & Johnson, 2003; Hayes & Wilson, 2008; Zuraw & Hayes, 2017). The induced grammar contains surface constraints that are weighted as in the framework of Harmonic Grammar (Legendre, Miyata, & Smolensky, 1990; Smolensky & Legendre, 2006). The constraints refer to sequences of under-attested natural classes in the wordlist. The natural classes are defined by a feature matrix provided by the user prior to a learning simulation. The learner assumes a probability space shared by all possible word forms based on the segments that are provided. Unattested and under-attested forms in the provided lexicon are penalized, which can be translated into lower probabilities of those forms; meanwhile, the learner increases the probabilities that it assigns to the attested forms, especially over-attested ones, in the lexicon. Maximum Entropy here thus refers to the fact that the weights in the grammar are induced in a way that maximizes the probabilities of the possible word forms that the training lexicon is drawn from, not just the lexicon itself. The induced constraints target sequences of natural classes. For example, a constraint that penalizes consonant clusters bears the form of *[+consonantal][+consonantal].

A few parameters control how the learner induces a grammar. The adjusted observed-over-expected (O/E) threshold restricts the learner to only induce constraints that refer to co-occurrences of natural groups whose adjusted O/E ratio is below a specified number. The O/E ratio describes the actual number of occurrences of certain combinations (O) divided by the expected number (E) based on random and unrestricted combinations. For example, given a strictly CV language with only five vowels /a, i, y, o, u/ and three onset consonants /p, b, t/, we would expect six occurrences of [+labial][+round] syllables (i.e., /po/, /bo/, /pu/, /bu/, /py/, /by/). If we only see /po/ in the actual lexicon, the O/E of [+labial][+round] would be $1/6 \approx 0.167$. A smaller number indicates that a particular sequence is under-attested. The UCLA Learner uses the statistical "upper confidence limit" (Mikheev, 1997; Albright & Hayes, 2002, 2003) to adjust O/E. This adjustment method has the effect of treating generalizations with a larger E as stronger. For example, under this method, the difference between an O/E of 0/10 and 0/1000 would be adjusted to 0.22 and 0.002, respectively.

The user can also provide a number of constraints that the learner should aim to induce. Beyond affecting the size of the induced grammar, varying the targeted number of constraints also alters the nature of the learned constraints. The learner prioritizes inducing constraints with a lower adjusted O/E value. This is referred to as the "accuracy" heuristic. Given the same level of accuracy, the learner prioritizes constraints that describe smaller n-grams (e.g., bigrams preferred over trigrams) and constraints with natural classes that cover more segments. This is referred to as the "generality" heuristic. With these two heuristics, constraints that are induced earlier or are induced when a simulation aims for fewer constraints would be more general and accurate (i.e., exceptionless or with a larger E).

Finally, the user can specify the maximum n-grams a constraint should try to capture. A larger *n* number increases the number of possible constraints for the learner to consider, especially given a larger number of natural classes. For example, with 400 natural classes, there would be 160,000 possible bigram constraints, 64 million possible trigram constraints, 26 billion possible 4-gram constraints, and 10 trillion possible 5-gram constraints; thus, bigrams and trigrams are preferred for segmental constraints (Hayes & Wilson, 2008).

The grammar induced by the learner can then be used to assign harmonic scores to word forms, and the scores can be compared with results of behavioral experiments, such as wordlikeness and lexical decision tasks (Berent et al., 2012; Daland et al., 2011; Gallagher et al., 2019; Goldwater & Johnson, 2003; Hayes & White, 2013; Wilson & Gallagher, 2018). The inductive process can also start with a set of pre-written constraints. In such cases, the learner can be used to add more constraints to the grammar or simply to determine the weights of the pre-written constraints based on a provided list of word forms.

In the current study, we extend this methodology to the modeling of wordlikeness ratings of syllable-tone gaps. This provides an opportunity to employ a computational learner to model possible syllable-tone phonotactics, which has not yet been fully explored in the literature. A few recent studies have employed similar approaches in using phonotactic well-formedness to measure experimental results in tone languages. Gong (2017) also used the UCLA Phonotactic Learner, but with two crucial differences. First, his data came from a visual lexical decision experiment, where stimuli were represented with *Bopomofo*, a phonetic alphabet used in Taiwan (Myers & Tsay, 2005). Second, and more importantly, he focused solely on segmental patterns without modeling syllable-tone combinations and found that unattested segmental combinations with higher harmonic scores elicited a higher proportion of wordlike responses and shorter response times. In a study similar to ours, Gong and Zhang (2021) compared wordlikeness ratings and harmonic scores from the UCLA Phonotactic Learner. Their approach also compared handwritten and induced grammars; however, the focus of their modeling was on attested and unattested forms in Mandarin. Moreover, their stimuli only included syllables with a high tone, without a specific focus on tonotactic gaps as in this study.

In a similar line of research, Do and Lai (2020) modeled nonce syllables in Cantonese, including both unattested segmental combinations and syllable-tone gaps. To estimate the probability of a syllable-tone combination, they used the probability of tone given the entire or part of the segmental strings. For example, the probability of /pit/^{55} was estimated by the likelihood of /pit/ having the high level tone (i.e., P(X^{55} | pit)), the likelihood of a syllable with the vowel /i/ having the high level tone (i.e., P(X^{55} | _i_), the likelihood of a syllable with the onset /p/ having the high level tone (i.e., P(X^{55} | p_)), the likelihood of a syllable with the coda /t/ having the high level tone (i.e., P(X^{55} | _t)), and the likelihood of a syllable with the rime /it/ having the high level tone (i.e., P(X^{55} | _it)). These probabilities were estimated by multinomial logistic regression analyses. Their Bayesian statistical analysis showed that phonotactic probabilities calculated in this way affected how wordlike a nonword syllable was judged to be, but in cases where the stimuli were judged to be absolutely unwordlike, there was no effect of phonotactic probability.

This study complements previous studies by testing a range of phonotactic grammars with varying degrees of access to different types of lexical data on how well they model wordlikeness ratings. More importantly, by testing tonotactic constraints and weights that reflect statistics gleaned from the lexicon as well as those that do not, we aim to tease apart statistical patterns of gaps and universal markedness and how they may account for native speakers' tonotactic knowledge.

The rest of this section is organized as follows: Section 4.1 describes the different settings for building the constraint-based phonotactic grammars. Section 4.2 discusses the induced phonotactic constraints, particularly whether they capture similar generalizations by typologically-motivated markedness constraints. Section 4.3 examines the correlation between the phonotactic grammars' well-formedness scores (MaxEnt scores) and the wordlikeness ratings from our behavioral experiment, with a focus on whether the inductive learning process from the lexicon is necessary for building a grammar that predicts the behavioral results. Finally, Section 4.4 summarizes our analysis on tonotactic grammars.

4.1. Building the phonotactic grammars

The set of phonetic symbols used in our analysis is shown in **Table 5**, with a breakdown of prosodic positions in which these sounds may occur. The symbol set was similar to that in Gong and Zhang (2021), treating [e] and [o] as separate sounds from [ə] despite a possible phonemic analysis that treats them as the same underlying phoneme. The learner was thus expected to induce phonotactic constraints that describe the complementary distribution of [e] and [o]. In the other two instances of complementary distribution, our symbol set took an allophonic analysis. First, the apical vowels behind sibilants were transcribed as /i/ (similar to Gong and Zhang). Second, unlike Gong and Zhang who separated /a/ and /ɑ/, we transcribed the low vowel as /a/ regardless of where it occurred. We also added the vowel /ɚ/, which was not

included in Gong and Zhang's study. All segments were annotated with the distinctive features required by the learner. Following Gong and Zhang, we used binary place features.

C	p p^h m f t t^h n l tɕ $tɕ^h$ ɕ ts ts^h s tʂ $tʂ^h$ ʂ ʐ k k^h x
G	j w ɥ
V	a ə e o i u y ɚ
X	i u n ŋ

Table 5: Segmental inventories in different prosodic positions in the learning simulation.

We applied Hayes and Wilson's (2008) treatment of lexical stress to the four lexical tones. Specifically, vowels with different tones were treated as separate vowel types (Kirby & Yu, 2007). For example, /a/ with different tones was transcribed as /a/^{55}, /a/^{35}, /a/^{214}, and /a/^{51} in the lexicon. Following Gong and Zhang (2021), we used privative tonal features to refer to these four tones. This treatment represents a null hypothesis concerning how the tones may be grouped into natural classes; that is, we did not force tones to be grouped into [+high tones] or [+rising tones] and bias the learners towards adopting such generalizations. Under this treatment, a constraint penalizing voiced onsets in T1 syllables would bear the form of *[+voice, +consonantal][T1]. Note that this approach can also handle the interaction between tones and vowels, penalizing certain tone-vowel combinations. For example, *[+low, T3] penalizes low vowels in T3. This way, under-attested combinations of classes of consonants, vowels, and tones could all be formulated and induced as constraints by the learner.

In addition to the provided segmental and tonal features, the learner adds a [syllable boundary] feature (abbreviated as [sb] henceforth). The feature describes the contrast between the syllable boundary ([+sb]) and non-boundary tokens ([−sb]), which essentially refers to all segments. This is how the learner expresses constraints that refer to syllable boundaries. For instance, *[+consonant][+sb] is a constraint that penalizes codas. This also allows the learner to describe a constraint that refers to "all segments" in a particular position. For example, *[+consonant][−sb][+syllabic] penalizes any segment between a consonant and a vowel.

The learner's inductive process required a list of word forms as the training data. We used a word list with 16,684 distinct lexical items from the TMC corpus. During training, these lexical items were broken down into 36,623 monosyllabic units (i.e., individual tokens of syllable-tone combinations). Similar types of training data have been used for inducing phonotactic constraints (e.g., Gong & Zhang, 2021; Gouskova & Gallagher, 2020; Hayes & Wilson, 2008). Since a syllable-tone combination may appear as homophonous, the word list did include some information about type frequency of distinct syllable-tone combinations in the lexicon. For example, the training data were given the information that the syllable-tone combination /t^ha^{55}/ occurs in

24 distinct words in the TMC corpus. Note that this type of frequency count only refers to how a particular tone-syllable combination occurs in unique lexical items and makes no reference to lexical token frequency in the TMC corpus, which was used as the Frequency variable in our experimental analysis.

Since our goal was to see the different extents to which the lexicon is needed for inducing tonotactic constraints that best model speakers' wordlikeness ratings, we ran the learning simulation in three different settings: The strong induction setting, the weak induction setting, and the no induction setting, each of which are described in greater detail in the following.

The **Strong Induction** setting: We ran learning simulations that aimed to induce 300 constraints. The maximum constraint length was set to trigrams. Out of the first 75, 150, and all 300 constraints[6], we selected constraints referring to segment-tone interactions, resulting in increasing sets of 25, 78, and 210 inductive tonotactic constraints (hereafter, the Small, Medium, and Large Strong Induction grammars). As mentioned earlier, constraints that are learned later in the simulation tend to be less accurate and less general. Grammars with more constraints learned later in the simulation and grammars aiming to learn more constraints are more likely to overfit to the lexicon by capturing statistical patterns that describe accidental gaps instead of describing more general phonotactic knowledge. On the other hand, grammars with only a few constraints induced in an early stage might not have captured relevant phonotactic knowledge, as only strong generalizations that target larger natural classes would be included. By varying the number of tonotactic constraints in this setting, we tested the extent to which the levels of statistical fit needed to be leveraged to induce a grammar that best models speakers' wordlikeness ratings. Finally, since we are only interested in tonotactic constraints, we only took the induced constraints that refer to segment-tone interactions from the simulations.

The **Weak Induction** setting: In this setting, we did not ask the learner to induce novel tonotactic constraints. Instead, we made it reweight a smaller set of ten typologically-motivated constraints, shown in (1) below. The constraints *[T3] and *[T2] are motivated by the typological markedness of contour and rising tones (e.g., Yip, 2002; Zhang, 2001). The other eight constraints refer to the incompatibility of high tones with voiced onsets, and low tones with voiceless onsets (e.g., Hsieh & Kenstowicz, 2008; Kenstowicz & Suchato, 2006; Ohala, 1978; Sagart, 1999; Yip, 2002). It is worth noting that since tones are carried by vowels in our learning simulation, the onset-tone interaction in syllables with an onglide has to be captured in a separate series of constraints (e.g., *[+voice][−consonantal][T1] and

[6] Targeting between 100 and 200 inductive constraints is common in recent works (e.g., Gouskova & Gallagher, 2020; Gallagher et al., 2019). We took the midpoint of this range, as well as the half and the doubled numbers to explore the potential effects of underfit and overfit to the lexicon.

*[−voice][−consonantal][T2]). This Weak Induction setting represents a scenario where the learner has a smaller hypothesis space concerning what the relevant and important tonotactic constraints are in this language.

(1) Typologically-motivated tonotactic constraints:
Markedness of contour and rising tones: *[T3], *[T2]
Markedness of voiced onsets with high tones: *[+voice][T1], *[+voice][T4],
*[+voice][−consonantal][T1], *[+voice][−consonantal][T4]
Markedness of voiceless onsets with low tones: *[−voice][T2], *[−voice][T3],
*[−voice][−consonantal][T2], *[−voice][−consonantal][T3]

The **No Induction** setting: Phonotactic grammars in this setting also contain typologically-motivated tonotactic constraints. The difference from the Weak Induction setting is that the weights of these constraints were decided independent of the lexicon. This represents a hypothesis where tonotactic constraints do not need to be informed by lexical knowledge, which we also call the baseline setting.

Since we are interested in tonotactic constraints, in all three settings, we used Gong and Zhang's (2021) 38 handwritten segmental constraints (Appendix II) along with the induced and typologically-motivated tonotactic constraints. These handwritten constraints refer to systematic, allophonic, and accidental segmental gaps in the Mandarin lexicon, and have shown to be much more effective in modeling native speakers' behavioral results than inductive segmental constraints. These segmental constraints served as the baseline grammar for accounting for variances in wordlikeness ratings that are not related to segment-tone interactions.

As part of this process, in the Strong Induction setting, the induced tonotactic constraints were added to the handwritten segmental constraints before undergoing another round of reweighting. In the Weak Induction setting, the typologically-motivated tonotactic constraints were reweighted along with the handwritten segmental constraints. In the No Induction setting, the handwritten segmental constraints were themselves weighted by the lexicon and used alongside the typologically-motivated tonotactic constraints. In other words, in all three induction conditions, the handwritten segmental constraints were weighted by the lexicon. This was a methodological choice made to simplify the non-tonal part of the grammar induction process. However, we acknowledge the possibility that, similar to tonotactic constraints, there may be other methods for assigning weights to segmental constraints that better account for native speakers' knowledge of segmental phonotactics. The procedure for generating the phonotactic grammars is summarized in **Table 6**.

Setting	Procedure	Number of resulting grammars
Strong Induction	Pick the first 75, 150, and all 300 constraints from a simulation → pick tonotactic constraints (25, 78, 210) → reweight these tonotactic constraints along with 38 handwritten segmental constraints	3
Weak Induction	Weight 10 typologically motivated constraints along with 38 handwritten segmental constraints	1
No Induction (baseline)	Weight 38 handwritten segmental constraints → use the weighted constraints along with 10 typologically motivated constraints with baseline weights (a weight of 3 for all 10 constraints)	1

Table 6: Generating phonotactic grammars with different tonotactic constraints and weights.

4.2. The induced tonotactic constraints

In this section, we discuss the content of the inductive tonotactic constraints, with a focus on whether they refer to the markedness of T3 and T2, or to the interaction between onset voicing and tone.

From the Small Strong Induction grammar with 25 induced tonotactic constraints, six constraints referred to the interaction between onset voicing and tone in the same direction as the typologically-motivated constraints, as shown in **Table 7**. Two of these penalized sequences of voiced onsets before a vowel with T1, even though they targeted smaller natural classes (e.g., non-/a/ vowel and non-labial nasals) instead of all voiced onsets and all vowels with T1. Other constraints consistent with the hypothesized direction targeted much smaller natural classes: Non-labial nasals before /o, ə, ɚ/ with T4; /s, ɕ/ before non-high vowels with T2; and the aspirated /t^h, ts^h, $tɕ^h$/ before mid vowels with T3. No constraints could be interpreted as a general

Constraint	Weight	Penalized sequences
*[+cons, +voice][−low,T1]	2.164	Voiced onset and non-/a/ vowels in T1
*[+nasal, −labial][T1]	2,753	Onset /n/ in T1
*[+nasal, −labial][−high, −low, −front,T4]	3.772	/n/ before /o, ə, ɚ/ in T4
*[−voice, +cont, +ant][−high,T2]	1.512	/s, ɕ/ before non-high vowels in T2
*[+aspirated, +anterior][−high, −low,T3]	1.673	/t^h, ts^h, $tɕ^h$/ before /e, o, ə, ɚ/ in T3
*[−voice, +coronal][−high, +front,T3]	2.296	Voiceless coronals before /e/ in T3

Table 7: Constraints from the Small Strong Induction grammar that targeted the interaction between onset voicing and tones in the expected directions given typological observations.

restriction against T2 or T3 syllables, even though there were slightly more constraints referring to sequences with T3 and T2 (seven and eight constraints, respectively) than with T1 and T4 (four and six constraints, respectively). See Appendix III for the full list of inductive tonotactic constraints in the Small Strong Induction grammar.

In the Medium Strong Induction grammar, there were a few constraints that penalized different subsets of the interaction between voiced onsets and T1. These are shown in **Table 8**. Other than the constraint *[+voice][−sb][T1], which penalized voiced onsets in a T1 syllable with an onglide, the other four constraints again targeted smaller natural classes than the typologically-motivated constraints.[7] It is also worth noting that there were no such constraints for T4.

Constraint	Weight	Penalized sequences
*[+voice][−front][−back,T1]	1.919	Voiced onsets before non-back vowels in a T1 syllable with the onglide /w/
*[+cons,+voice][+low,T1][−labial]	1.98	Voiced onsets before /a/ in a T1 syllable with nasal coda
*[+voice,+delayed][T1]	2.269	The onset /ʐ/ in a T1 syllable
*[+voice][−sb][T1]	4.124	Voiced onsets in a T1 syllable with an onglide ([−sb] refers to the natural class of all "non-boundary" segments.)

Table 8: Induced constraints in the Medium Strong Induction grammar that refer to incompatibility between voiced onsets and T1/T4.

Similar to the Small Strong Induction grammar, constraints that penalized voiceless onsets and T2/T3 mostly targeted smaller natural classes down to two consonants and one vowel. Some of these are shown in **Table 9**. Again, there were no constraints that targeted T2 and T3 in general, though slightly more constraints targeted T2 and T3 (26 and 21 constraints) than T1 and T4 (18 and 13 constraints).

Constraint	Weight	Penalized sequences
*[−aspirated][T2][+consonantal]	3.982	Unaspirated onsets in T2 syllables with codas
*[−voice,+cont,−labial,−dorsal][+high,+round,T2]	2.024	/s, ʂ/ before /u, y/ in T2
*[−aspirated,−anterior][−high,+back,T3]	1.954	Unaspirated onsets before /o/ in T3 syllables

Table 9: Induced constraints in the Medium Strong Induction grammar that referred to incompatibility between voiceless onsets and T2/T3.

[7] The learner used the binary feature [sb] to label syllable boundaries ([+sb]) and all non-boundary symbols (i.e., all segments ([−sb])).

A similar trend was observed when we examined the additional constraints in the Large Strong Induction grammar: There were constraints that penalized voiced onsets and T1/T4 in the same direction (e.g., *[+voice, −coronal][−sb][−back,T1], *[+nasal][−front][−back,T4]) and voiceless onsets and T2/T3 (e.g., *[+aspirated,labial][−front][+round,T2], *[−voice, −dorsal][−back][+round,T3]). As the additional constraints tended to target sequences of smaller natural classes, there were again no constraints specifically targeting T2 and T3 in general, but still more constraints targeting sequences bearing subsets of T3 and T2 (64 and 59 constraints) syllables than T1 and T4 (44 and 43 constraints).

Beyond these constraints that could be interpreted as being relevant for onset-tone interactions in the expected directions, many more constraints targeted other aspects of syllable-tone interactions such as dorsal onsets in certain vowel-tone configurations (e.g., *[+consonantal, −labial, −coronal][+front,T2]), certain groups of onsetless vowels in certain tones (*[+sb][−high, +back,T2]), or the /ɚ/ vowel with T1. They could also target specific tonotactic gaps; for example, *[+sb][+low,T3] describes the absence of onsetless /a/ in T3.

Overall, among all the inductive onset-tone interaction constraints, those involving voiced onsets and T1 tended to be more general, targeting broader natural classes). While constraints penalizing voiced-T4 and voiceless-T2/T3 combinations were also induced, they mostly targeted smaller natural classes. The completely bottom-up inductive method also failed to identify constraints that targeted a larger proportion of T2 and T3 syllables, although there were more constraints that targeted subsets of the interaction between T2/T3 syllables and the natural classes referring to onsets and codas.

4.3. Correlation between harmonic scores and wordlikeness ratings

Grammars generated in different settings were used to assign harmonic scores H(x) to each stimulus. The correlation between wordlikenss ratings and harmonic scores was then used to evaluate how well different grammars predict the experimental results. For each stimuli *x*, a grammar assigned a harmonic score based on the summed weights of the constraints violated by *x*, as shown in (3a). In this study, we follow Gong and Zhang (2021) in using MaxEnt scores calculated by raising *e* to the negative power of the harmonic score H(x), as shown in (3b) The MaxEnt scores ranged from 0 (least well-formed) to 1 (most well-formed).

(2) Harmonic score and MaxEnt score

a. $H(x) = \sum_i w_i C_i(x)$

b. $MaxEnt(x) = e^{-H(x)}$

Table 10 shows the correlation between the wordlikeness ratings and the MaxEnt scores assigned by the different grammars. Following Gong and Zhang (2021), we report the correlation between MaxEnt scores and wordlikeness ratings of all stimuli, including both the gaps and the lexical syllables ("all ratings"). Additionally, we report correlation results specifically for gaps, as we are particularly interested in how different phonotactic grammars account for participants' ratings of items outside of the lexicon. However, we have chosen not to report or discuss the correlation between ratings and MaxEnt scores for lexical syllables, as the high ratings for these syllables may create a strong ceiling effect that could reduce the informativeness of the phonotactic effect.

The correlations show that having more inductive tonotactic constraints allowed the grammars' harmonic scores to better predict the overall wordlikeness ratings. However, when we focus on the ratings for gaps, the grammars with typologically motivated tonotactic constraints (i.e., those with the Weak and No Induction settings) vastly outperformed the grammar with inductive tonotactic constraints (i.e., that with the Strong Induction settings). Having more inductive tonotactic constraints in the Strong Induction setting helped predict overall wordlikeness ratings but did not improve predictions on the ratings for gaps only, suggesting that the inductive tonotactic constraints are increasingly able to differentiate between lexical syllables and gaps. This contrast suggests that while learning more nuanced segment-tone interactions in the lexicon helps further modeling the difference between lexical syllables and gaps, it was not very helpful in modeling how native speakers judged certain gaps to be more wordlike. Another finding worth noting is that the baseline grammar with the No Induction setting outperformed the grammar with the Weak Induction setting in which the constraints were weighted by the lexicon.

Grammar	Number of tonotactic constraints	Correlation with all ratings	Correlation with gap ratings
No Induction baseline	9	0.229	0.372
Weak Induction	9	0.273	0.302
Strong Induction (small)	25	0.232	0.059
Strong Induction (medium)	78	0.468	0.117
Strong Induction (large)	210	0.558	0.116

Table 10: Comparison of correlation between wordlikeness ratings and MaxEnt scores from different grammars.

Figure 11 shows the correlations between the wordlikeness ratings and the MaxEnt scores from the grammars with inductive constraints (the Strong Induction condition). With only a small number of inductive tonotactic constraints, very few stimuli were marked as less well-formed. When more tonotactic constraints were induced, gradient differences in well-formedness started to emerge, and more gaps began receiving low MaxEnt scores, which explains why the

correlation between the MaxEnt scores and wordlikeness ratings increased as the number of inductive constraints grew. However, this trend did not strengthen the correlation between the MaxEnt scores and the ratings for gaps.

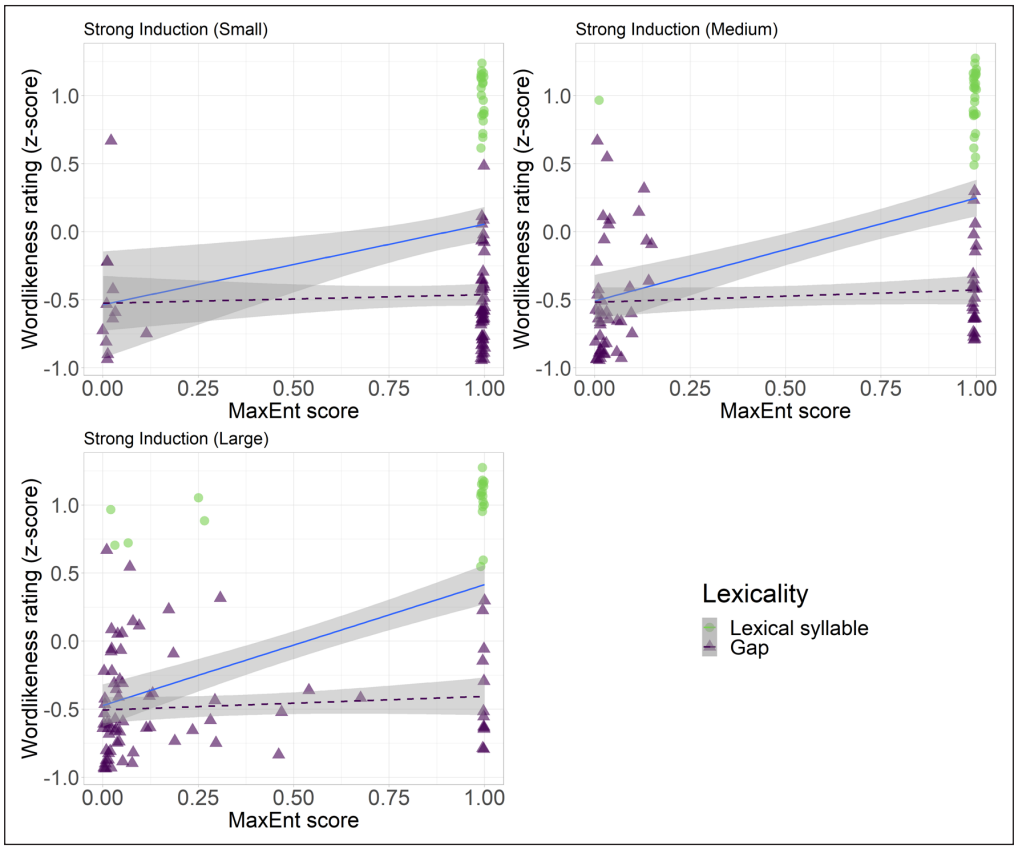

Figure 11: Correlation between the MaxEnt scores and wordlikeness ratings from the Strong Induction grammar. Light dots refer to lexical syllables, and dark dots refer to gaps. Horizontal jittering with a range of 0.02 was applied to the dots to reduce overlaps. The solid lines indicate slopes for MaxEnt scores predicting ratings for all items, and the dotted lines indicate slopes for predicting ratings for gaps only. The shading represents the 95% confidence interval for the slopes.

Figure 12 shows the correlation between the MaxEnt scores and wordlikeness ratings for the No Induction and Weak Induction grammars, both of which consisted of only typologically-motivated tonotactic constraints. For the No Induction grammar with baseline weights, the MaxEnt scores of a stimulus decreased if it was with T3 and T2, had a voiced onset with T1/T4, or had a voiceless onset with T2/T3, which explains the almost binary distribution horizontally. There were also a large number of lexical syllables with very low MaxEnt scores, which explains the relatively poor correlation between the MaxEnt scores and wordlikeness ratings in this condition. The fact that the No Induction baseline grammar failed to distinguish lexical syllables and gaps also suggests that it may simulate a cognitive module for tonotactics independent of the lexicon.

The Weak Induction grammar, on the other hand, only penalized items with T3 and T2 as well as items with a voiced onset and T1 (**Table 11**). In other words, based on the lexicon, the learner only found statistical support for the dispreference for T3, T2, and syllables with voiced onsets and T1. By incorporating these weights, the number of lexical syllables with low MaxEnt scores was greatly reduced, explaining the advantage of this grammar over the No Induction baseline grammar in terms of predicting ratings for all stimuli.

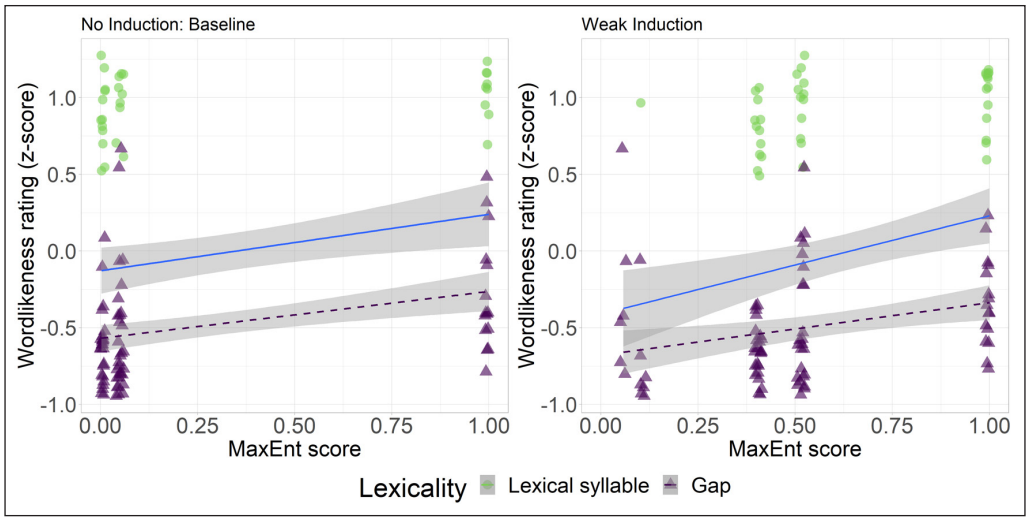

Figure 12: Correlation between the No Induction baseline (left) and Weak Induction grammars' MaxEnt scores and wordlikeness ratings. Light dots refer to lexical syllables, and dark dots refer to gaps. Horizontal jittering with a range of 0.02 was applied to the dots to reduce overlaps. The solid lines indicate slopes for MaxEnt scores predicting ratings for all items, and the dotted lines indicate slopes for predicting ratings for gaps only. The shading represents the 95% confidence interval for the slopes.

Constraint	Weights in Weak Induction grammar	Weights in No Induction baseline grammar
*[T3]	0.905	3
*[T2]	0.664	3
*[+voice][T1]	2.852/2.259	3
*[+voice][T4]	0	3
*[−voice][T2]	0	3
*[−voice][T3]	0	3
Correlation with overall ratings	0.273	0.229
Correlation with gap ratings	0.302	0.372

Table 11: Comparison of constraint weights in the No Induction baseline and the Weak Induction grammars.

However, when we focus on gaps only, it is the No Induction grammar that best predicts the ratings. In other words, blindly penalizing certain types of stimuli (i.e., those with T3, T2, voiced onset-T1/T4, voiceless onset-T2/T3) can better explain which gaps were rated as more wordlike by the native speakers.

To explore the optimal weights for the typologically-motivated constraints, we altered the weights of the constraints in the No Induction grammar from the fixed baseline with a grid search: The No Induction grammar with all combinations of the five weights (0, 1.5, 3, 4.5, 6) for the typologically-motivated constraints were tested. We simplified the process by giving the constraints with and without reference to glides (e.g., *[+voice][T1] and *[+voice][−consonantal][T1]) the same weight in each combination, yielding a total of 15,625 (5^6) combinations of weights. We view this as an effort to locate an optimal configuration for the cognitive module independent of lexical knowledge when it comes to modeling the behavioral results.

Table 12 shows the weights for the typologically-motivated constraints in two of the 15,625 grammars. The No Induction baseline and Weak Induction grammars (**Table 11**) are also listed for reference. The grammar that best predicted gap ratings weighted *[T3] heavily while placing smaller weights on *[+voice][T1] and *[+voice][T4] and no weight on other constraints. In other words, having a strong markedness constraint against the complex contour tone and modest constraints on voiced onsets and high tones helped predict the wordlikeness ratings for the gaps. **Table 12** also lists the weights for the grammar that best predict the wordlikeness ratings for all test items. This grammar weighted the [+voice][T1] constraint heavily while assigning a smaller weight to *[T3].

Constraint	No Induction: Optimized for gap prediction	No Induction: Optimized for overall prediction	No Induction: Baseline	Weak Induction
*[T3]	6	1.5	3	0.905
*[T2]	0	0	3	0.664
*[+voice][T1]	1.5	6	3	2.852/2.259
*[+voice][T4]	1.5	0	3	0
*[−voice][T2]	0	0	3	0
*[−voice][T3]	0	0	3	0
Correlation with all ratings	0.241	0.265	0.229	0.273
Correlation with gap ratings	0.428	0.369	0.372	0.302

Table 12: Comparisons of constraint weights for No Induction grammars that are best at predicting gaps. The No Induction baseline and Weak Induction grammars are listed for reference.

The correlation between the MaxEnt scores of these "optimized" grammars and the wordlikeness ratings is shown in **Figure 13**. Compared with the No Induction baseline grammar, the lower end of the MaxEnt scores shows greater gradience. The difference between the overall-optimized and the gap-optimized versions is that the former penalizes lexical syllables less than the latter.

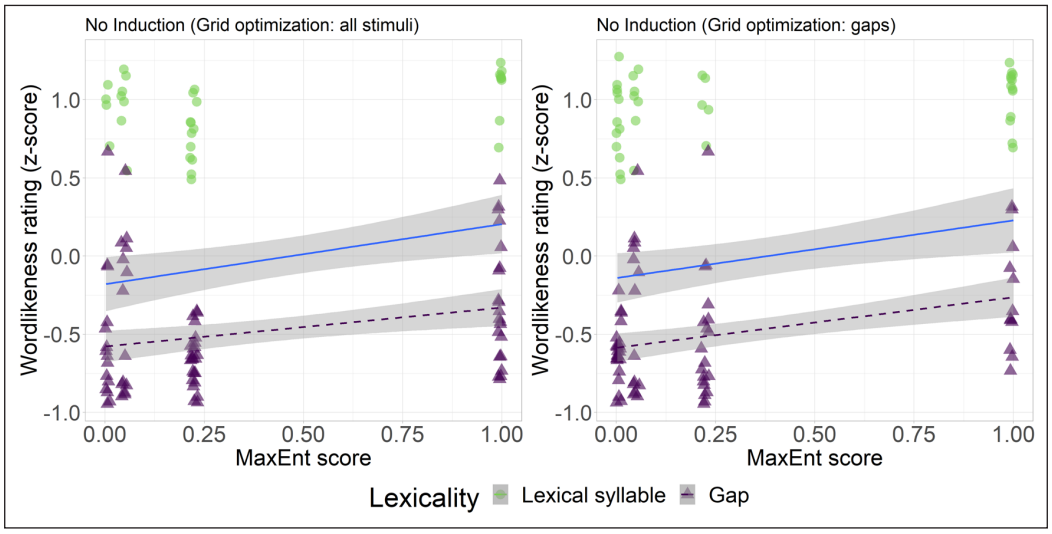

Figure 13: Correlation between the No Induction grammars' MaxEnt scores and wordlikeness ratings when the weights were the most successful at predicting all items (left) and gaps only (right) in the grid search. Light dots refer to lexical syllables, and dark dots refer to gaps. Horizontal jittering with a range of 0.02 was applied to the dots to reduce overlaps. The solid lines indicate slopes for MaxEnt scores predicting ratings for all items, and the dotted lines indicate slopes for predicting ratings for gaps only. The shading represents the 95% confidence interval for the slopes.

To further evaluate to what extent can we find the optimal weights for typologically-motivated tonotactic constraints, we ran 10,000 iterations of random weight assignments to *[+voice]([−cons])[T1], *[+voice]([−cons])[T4], *[−voice]([−cons])[T2], *[−voice]([−cons])[T3], *[T2], and *[T3]. That is, we assigned the same weights to the constraints with and without reference to glides. The random weights followed a uniform distribution between 0 and 6. **Figure 14** shows the distribution of grammars with typologically-motivated tonotactic constraints with random weights in terms of their correlation with the wordlikeness ratings for all stimuli (left) and gaps only (right).

Comparing the Weak Induction grammar to the No Induction grammar with random weights shows that weighting typologically-motivated tonotactic constraints based on the lexicon made the harmonic scores correlate almost as well as they possibly could given these constraints. On the other hand, when the focus was solely on the wordlikeness of gaps, weights based on the lexicon resulted in a very poor correlation between harmonic scores and the ratings relative to what

the random weights could do. These differences were confirmed by statistical tests. The MaxEnt scores from the No Induction grammars with randomized weights had a weaker correlation with wordlikeness ratings than those from the Weak Induction grammar, according to a one-sample t test ($t(9999) = -186$, $p < .0001$). On the other hand, the MaxEnt scores from the No Induction grammars with randomized weights had a stronger correlation with the wordlikeness ratings than those from the Weak Induction grammar, according to a one-sample t test ($t(9999) = 148.17$, $p < .0001$). In other words, while modeling the wordlikeness of all stimuli can benefit from the knowledge of statistical patterns in the lexicon, modeling the wordlikeness of tonotactic gaps does not need to be informed by the lexicon at all.

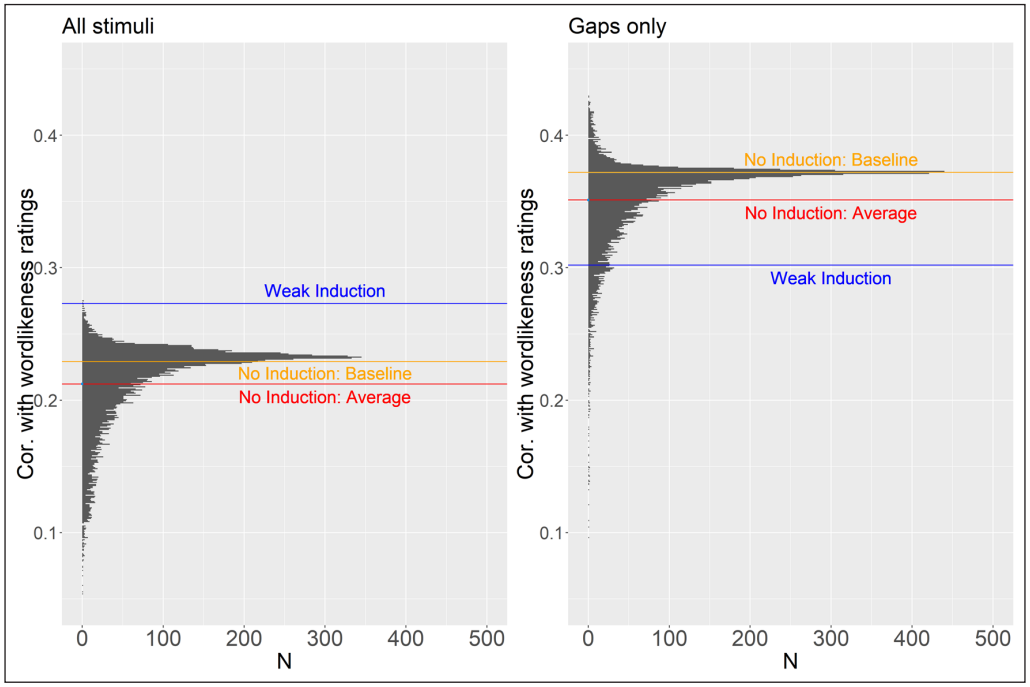

Figure 14: Distribution of the No Induction grammars with random weights for their correlation with the overall wordlikeness ratings (left) and ratings for gaps only (right). Horizontal lines indicate the mean correlation number (red), the correlation number for the Weak Induction (blue) and the No Induction baseline (yellow) settings.

4.4. Summary

In this section, we explored whether statistical information from the Mandarin lexicon is sufficient or necessary to build tonotactic constraints that account for the wordlikeness ratings in our behavioral experiment. This was done in two steps. We first examined the constraints induced from the lexicon. We then examined the correlation between the wordlikeness ratings and the MaxEnt scores of the different grammars. We observed some general constraints against voiced onsets in T1 syllables, which was consistent with one of our typologically-motivated constraints.

Even though we saw induced constraints that penalized voiceless onsets with T2/T3 and voiced onsets with T4, they mostly targeted much smaller natural classes and could be better viewed as constraints against specific accidental gaps. This potential "overfit" to accidental gaps in the constraint induction became more likely as the learner induced more constraints.

Our analysis of the correlation between the MaxEnt scores and wordlikeness ratings showed that speakers' different ratings for lexical syllables and tonotactic gaps could be successfully modeled by tonotactic constraints directly induced from the lexicon (the Strong Induction grammar) by the UCLA Phonotactic Learner. For modeling the ratings for tonotactic gaps, grammars with a smaller number of typologically-motivated constraints (i.e., *[T3], *[T2], *[+voice][T1/T4], *[−voice][T2/T3]) having arbitrary and random weights almost always outperformed the grammars with these constraints weighted from the lexicon and grammars with inductive tonotactic constraints. In other words, nearly all potential combinations of this limited set of typologically-motivated tonotactic constraints yielded better results than the lexicon-informed grammars when modeling the wordlikeness ratings of gaps. Among the possible configurations, we found that a large weight for *[T3] and a relatively smaller weight for *[+voice][T1/T4] could make the grammar assign harmonic scores that best correlated with the wordlikeness ratings of gaps. The finding on the importance of [+voice][T4] is particularly interesting since it was not supported by the lexicon at all (i.e., the inductive process assigned no weights to this constraint).

In short, results with the UCLA Phonotactic Learner suggested that learning constraints from the lexicon is neither sufficient nor necessary when it comes to predicting the wordlikeness ratings of tonotactic gaps, which is consistent with a view that phonological grammar for tonotactics is potentially a module independent of lexical memory. Replications using other phonotactic modeling tools (e.g., neural network-based models as reported in Mayer & Nelson, 2020) are needed to provide further support to this view, particularly in a similar setup where a priori and lexically-informed knowledge are compared.

5. General discussion

This study investigated Mandarin tonotactic accidental gaps by looking for patterns in corpus data and comparing the findings to wordlikeness ratings and to the harmonic scores generated by the UCLA Phonotactic Learner. Our corpus study revealed certain trends in the occurrence of accidental gaps in Mandarin. Specifically, we found that T2 gaps were over-represented, followed by T3 gaps, and both tended to occur with closed syllables. T4 gaps were the least common, a result that could be attributed to a historical tonal merging process. We also found fewer T1 gaps with voiced onsets than T2 and T3 gaps, which were more likely to occur with voiceless onsets, a pattern that has also been observed cross-linguistically. In listeners' wordlikeness ratings, however, T2, the tone with the most gaps, was not rated as the least wordlike. Instead, the

listeners rated T3 gaps as the least wordlike, a result that could be attributed to the markedness of the T3 contour. Furthermore, we found T1 gaps with voiced onsets were also rated as less wordlike, a pattern that was also observed in our corpus study. Although there was a significant difference in wordlikeness ratings between gaps and lexical syllables, they were both gradiently accepted as wordlike based on neighborhood density. Our findings across the corpus analysis, the wordlikeness rating experiment and modeling analyses are summarized in **Table 13**.

Corpus analysis	**Wordlikeness rating**	**Modeling with phonotactic grammars**	**What it means**
More T2 gaps	T3 least wordlike	*T2 is learnable from the lexical data but is not necessary for predicting wordlikeness ratings. The *T3 constraint is effective in modeling ratings especially for gaps. *T3 is learnable from the lexical data, but the induced weight was not as effective in predicting the ratings for gaps.	More T2 gaps is accidental, while more T3 gaps may reflect an effect from typological markedness.
More gaps with voiced onsets for T1 vs. T2/T3/T4	Voiced onset less wordlike for T1	*[+voice][T1] is learnable from the lexical data, but the induced weight is more useful for modeling all items than for modeling the gaps. *[+voice][T4] was not learnable from the lexical data, but it helped increase the correlation between the MaxEnt scores and ratings, especially for gaps.	There are potentially real phonotactic constraints that are not entirely learnable from the lexicon.
More gaps with closed syllables	No effect	Not explored.	Not real phonotactic constraints

Table 13: Summary of the findings across the corpus analysis, the wordlikeness rating experiment, and modeling analyses.

Taken together, not all patterns observed in the corpus were reflected in the wordlikeness ratings. For instance, contrary to the findings in the corpus, T2 gaps were not treated as the least wordlike by native speakers. Instead, T3 was rated the least wordlike. This pattern was not induced by the phonotactic learner, nor was the fact that T4, the tonal category with the fewest gaps, was rated as more wordlike. We attributed the speakers' general aversion to T3 to the universal markedness of its complex tonal contour.

Some patterns, however, were robustly observed throughout this study. More T1 gaps were found with voiced onsets in the corpus, while more T2, T3, and T4 gaps were found with

voiceless onsets. Among these patterns, we found T1 syllables with voiced onsets were rated as less wordlike, which suggests that T1 syllables with voiced onsets may not be as accidental as previously assumed; they also reflect psychologically real phonotactic restrictions regarding T1 gaps and voiced onsets. While the lack of a T2-*OnsetVoicing* effect in the wordlikeness experiment may suggest that the pattern was purely accidental, it could be the case that the overall low scores for T2 caused the interaction to fail to show.

With the help of the UCLA Phonotactic Learner, we examined to what extent the lexical data could help induce constraints or weight handwritten constraints that best account for the wordlikeness rating results. The comparison showed the statistics from the lexicon were beneficial in inducing a grammar with tonotactic constraints that predicts the ratings for all stimuli. However, in predicting the ratings for tonotactic gaps only, typologically-motivated constraints without lexical access outperformed grammars with tonotactic constraints induced from the lexical data. Although important markedness constraints, such as *T3 and *[+voice][T1], were assigned weights when trained on the lexical data, grammars with these weights had a weaker correlation with wordlikeness ratings of gaps compared to grammars with arbitrary weights for these constraints. Crucially, iterations with random weight assignments revealed that lexically-informed weights were significantly worse than random weights in modeling gaps' ratings. It is important to note that, despite our exploration of various constraint induction and weighting setups, the results we have obtained may still be limited by the UCLA Phonotactic Learner, the training data, and the simulation setups we have used. Therefore, it is uncertain whether our findings are indicative of limitations in grammar induction from the lexicon alone. If this finding can be replicated using other simulation tools and settings that compare a priori and lexically-informed phonotactic knowledge, it may suggest that universal markedness is more relevant than patterns in the lexicon for modeling the wordlikeness of tonotactic gaps.

In their explanation of similar dissociations between phonotactic knowledge and statistical generalizations in the lexicon in modeling nonword perception and production, Becker et al. (2011) proposed that UG serves as a filter on possible generalizations that humans can make (see also Davidson, 2006; Moreton, 2002), which may, in turn, facilitate the (over)-learning of phonetically motivated patterns The lack of such filters explains why inductive statistical models fail to model behavioral results since these models are prone to learning accidental statistical patterns (the "surfeit-of-the-stimulus" effect). Our findings are consistent with the predictions made by the proposal that only a subset of generalizations are possible, as shown by the success of typologically-motivated tonotactic constraints over inductive ones in modeling nonword tonotactics. We further showed that there is little evidence for an association between statistical patterns in the lexicon and the exact configurations of the possible constraints (i.e., constraint weights) other than the fact that information from the lexicon is helpful in modeling

the separation between gaps and lexical syllables. Since *T3 and *[+voice][T1/T4] both have phonetic motivations, innateness and phonetic naturalness of these constraints are both potential sources of such tonotactic knowledge. The findings regarding the roles of *T3 and *[+voice][T4] are particularly interesting, as they are not as strongly supported by statistical patterns in the lexicon as *[+voice][T1]. This suggests that *T3 and *[+voice][T4] may have been potentially overlearned from the lexicon. On the other hand, despite *T2 being supported by the lexicon, it was not relevant in modeling nonword tonotactics. This indicates that the speakers may have underlearned this statistical pattern.

This study contributes to the general understanding of unattested forms, especially involving tone-segment combinations, and extends the modeling of phonotactic well-formedness that has been previously restricted to segmental combinations to tone-syllable combinations.

Additional files

The additional files for this article can be found as follows:

- **Appendix I.** Stimuli for Wordlikeness Rating Experiment. DOI: https://doi.org/10.16995/labphon.6455.s1
- **Appendix II.** Handwritten Segmental Constraints adopted from Gong & Zhang (2021). DOI: https://doi.org/10.16995/labphon.6455.s2
- **Appendix III.** Tonotactic constraints in the Small Strong Inductive grammar. DOI: https://doi.org/10.16995/labphon.6455.s3
- **Corpus-data.** The 'Mandarin Accidental Gap Corpus' (Section 2) includes a calculation of 398 allowable Mandarin syllables with all possible tonal combinations, whether they exist or not. DOI: https://doi.org/10.16995/labphon.6455.s4
- **Wordlikeness rating data.** Wordlikeness ratings (Section 3) were obtained from thirty-seven Taiwan Mandarin native speakers for 288 stimuli. DOI: https://doi.org/10.16995/labphon.6455.s5

Acknowledgements

We would like to thank Sang-Im Lee-Kim, Tsung-Ying Chen, the editors and reviewers of *Laboratory Phonology*, and the participants in ICPhS 2019 for their insights and suggestions. We especially thank Dr. Bruce Hayes for directing us to include a computational modeling component in this study. Any remaining errors are ours.

Competing interests

The authors have no competing interests to declare.

References

Albright, A. (2003). A quantitative study of Spanish paradigm gaps. Paper presented at the 22nd West Coast Conference on Formal Linguistics, University of California, San Diego.

Albright, A. (2009). Feature-based generalisation as a source of gradient acceptability. *Phonology*, *26*(1), 9–41. DOI: https://doi.org/10.1017/S0952675709001705

Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, *90*(2), 119–161. DOI: https://doi.org/10.1016/S0010-0277(03)00146-X

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. DOI: https://doi.org/10.18637/jss.v067.i01

Becker, M., Nevins, A., & Levine, J. (2012). Asymmetries in generalizing alternations to and from initial syllables. *Language*, *88*(2), 231–268. DOI: https://doi.org/10.1353/lan.2012.0049

Berent, I., Wilson, C., Marcus, G. F., & Bemis, D. K. (2012). On the role of variables in phonology: Remarks on Hayes and Wilson 2008. *Linguistic inquiry*, *43*(1), 97–119. DOI: https://doi.org/10.1162/LING_a_00075

Boersma, P., & Weenink, D. (2017). Praat: Doing phonetics by computer (Version 6.0.26). Retrieved from www.praat.org.

Chen, Y., & Xu, Y. (2006). Production of weak elements in speech–evidence from f_0 patterns of neutral tone in Standard Chinese. *Phonetica*, *63*(1), 47–75. DOI: https://doi.org/10.1159/000091406

Chien, Y.-F., Sereno, J. A., & Zhang, J. (2017). What's in a word: Observing the contribution of underlying and surface representations. *Language and Speech*, *60*(4), 643–657. DOI: https://doi.org/10.1177/0023830917690419

Coleman, J., & Pierrehumbert, J. (1997). Stochastic phonological grammars and acceptability. arXiv preprint cmp-lg/9707017.

Cutler, A., & Chen, H.-C. (1997). Lexical tone in Cantonese spoken-word processing. *Perception and Psychophysics*, *59*(2), 165–179. DOI: https://doi.org/10.3758/BF03211886

Daland, R., Hayes, B., White, J., Garellek, M., Davis, A., & Norrmann, I. (2011). Explaining sonority projection effects. *Phonology*, *28*, 197–234. DOI: https://doi.org/10.1017/S0952675711000145

Davis, M. J. (2010). Contrast coding in multiple regression analysis: Strengths, weaknesses, and utility of popular coding structures. *Journal of Data Science*, *8*(1), 61–73. DOI: https://doi.org/10.6339/JDS.2010.08(1).563

Della Pietra, S., Della Pietra, V., & Lafferty, J. (1997). Inducing features of random fields. *IEEE transactions on pattern analysis machine intelligence*, *19*(4), 380–393. DOI: https://doi.org/10.1109/34.588021

Do, Y., & Lai, R. K. Y. (2020). Incorporating tone in the modelling of wordlikeness judgements. *Phonology*, *37*(4), 577–615. DOI: https://doi.org/10.1017/S0952675720000287

Duanmu, S. (2007). *The Phonology of Standard Chinese*. New York: Oxford University Press.

Duanmu, S. (2011). Chinese syllable structure. *The Blackwell Companion to Phonology*, 1–24. DOI: https://doi.org/10.1002/9781444335262.wbctp0115

Fischer-Jørgensen, E. (1952). On the definition of phoneme categories on a distributional basis. *Acta linguistica*, *7*(1–2), 8–39. DOI: https://doi.org/10.1080/03740463.1952.10415400

Fon, Janice, & Chiang, Wen-Yu. (1999). What does Chao have to say about tones? A case study of Taiwan Mandarin/赵氏声调系统与声学之联结及量化–以台湾地区国语为例. *Journal of Chinese Linguistics*, *27*(1), 13–37.

Frisch, S. A., Large, N. R., & Pisoni, D. B. (2000). Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. *Journal of Memory and Language*, *42*(4), 481–496. DOI: https://doi.org/10.1006/jmla.1999.2692

Frisch, S. A., Pierrehumbert, J. B., & Broe, M. B. (2004). Similarity avoidance and the OCP. *Natural Language & Linguistic Theory*, *22*(1), 179–228. DOI: https://doi.org/10.1023/B:NALA.0000005557.78535.3c

Gallagher, G., Gouskova, M., & Camacho Rios, G. (2019). Phonotactic restrictions and morphology in Aymara. *Glossa: A Journal of General Linguistics*, *4*(1), 1–48. DOI: https://doi.org/10.5334/gjgl.826

Goldwater, S., & Johnson, M. (2003). Learning OT constraint rankings using a maximum entropy model. Paper presented at the Proceedings of the Stockholm workshop on variation within Optimality Theory.

Gong, S. (2017). Grammaticality and lexical statistics in Chinese unnatural phonotactics. *UCL Working Papers in Linguistics*, *29*, 1–23.

Gong, S., & Zhang, J. (2021). Modelling Mandarin speakers' phonotactic knowledge. *Phonology*, *38*(2), 241–275. DOI: https://doi.org/10.1017/S0952675721000166

Gouskova, M., & Gallagher, G. (2020). Inducing nonlocal constraints from baseline phonotactics. *Natural Language Linguistic Theory*, *38*(1), 77–116. DOI: https://doi.org/10.1007/s11049-019-09446-x

Halle, M. (1962). Phonology in generative grammar. *Word*, *18*(1–3), 54–72. DOI: https://doi.org/10.1080/00437956.1962.11659765

Hao, Y.-C. (2012). Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers. *Journal of Phonetics*, *40*(2), 269–279. DOI: https://doi.org/10.1016/j.wocn.2011.11.001

Hayes, B., & White, J. (2013). Phonological naturalness and phonotactic learning. *Linguistic Inquiry*, *44*(1), 45–75. DOI: https://doi.org/10.1162/LING_a_00119

Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, *39*(3), 379–440. DOI: https://doi.org/10.1162/ling.2008.39.3.379

Hombert, J.-M., Ohala, J. J., & Ewan, W. G. (1979). Phonetic explanations for the development of tones. *Language*, *55*(1), 37–58. DOI: https://doi.org/10.2307/412518

Hsieh, F.-F., & Kenstowicz, M. J. (2008). Phonetic knowledge in tonal adaptation: Mandarn and English loanwords in Lhasa Tibetan. *Journal of East Asian Linguistics*, *17*, 279–297. DOI: https://doi.org/10.1007/s10831-008-9027-7

Huang, K. (2012). *A study of neutral-tone syllables in Taiwan Mandarin*. Honolulu: University of Hawaii at Manoa.

Huang, T. (2001). The interplay of perception and phonology in tone 3 sandhi in Chinese Putonghua. In Hume, E., & Johnson, K., (Eds.), *Studies on the Interplay of Speech Perception and Phonology*, *55*, 23–42. Ohio State University.

Huang, T., & Johnson, K. (2010). Language specificity in speech perception: Perception of Mandarin tones by native and nonnative listeners. *Phonetica*, *67*(4), 243–267. DOI: https://doi.org/10.1159/000327392

Hume, E., & Johnson, K. (2003). The impact of partial phonological contrast on speech perception. Paper presented at the Proceedings of the fifteenth international congress of phonetic sciences.

Kenstowicz, M., & Suchato, A. (2006). Issues in loanword adaptations: A case study from Thai. *Lingua*, *116*(7), 921–949. DOI: https://doi.org/10.1016/j.lingua.2005.05.006

Kirby, J. P., & Yu, A. C. L. (2007). Lexical and phonotactic effects on wordlikeness judgments in Cantonese. Paper presented at the Proceedings of the International Congress of the Phonetic Sciences XVI.

Kubler, C. C. (1985). The influence of Southern Min on the Mandarin of Taiwan. *Anthropological Linguistics*, *27*(2), 156–176.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2016). lmerTest: Test in linear mixed effects model: R package version 2.0-33.

Lai, Y. C. (2003). *A Perceptual Investigation on Mandarin Tonotactic Gaps*. (M.A.), National Tsing Hua University, Hsinchu, Taiwan.

Lee, C.-Y. (2007). Does Horse Activate Mother? Processing lexical tone in form priming. *Language and Speech*, *50*(1), 101–123. DOI: https://doi.org/10.1177/00238309070500010501

Legendre, G., Miyata, Y., & Smolensky, P. (1990). Harmonic Grammar—A formal multi-level connectionist theory of linguistic well-formedness: Theoretical foundations. Paper presented at the Proceedings of the twelfth annual conference of the Cognitive Science Society, Cambridge, MA: Erlbaum.

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2019). emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.3.2.

Li, Jian. (2013). *The rise of disyllables in Old Chinese: The role of lianmian words*. City University of New York.

Lin, Y.-H. (2007). *The Sounds of Chinese*. Cambridge, UK: Cambridge University Press.

Lu, Y.-A., & Lee-Kim, S.-I. (2021). The effect of linguistic experience on perceived vowel duration: Evidence from Taiwan Mandarin speakers. *Journal of Phonetics*, *86*, 101049. DOI: https://doi.org/10.1016/j.wocn.2021.101049

Mayer, C., & Nelson, M. (2020). Phonotactic learning with neural language models. *Proceedings of the Society for Computation in Linguistics*, *3*(1), 149–159.

Mei, T.-L. (1970). Tones and prosody in Middle Chinese and the origin of the rising tone. *Harvard Journal of Asiatic Studies*, *30*, 86–110. DOI: https://doi.org/10.2307/2718766

Mei, T.-L. (1977). Tones and tone sandhi in 16th century Mandarin. *Journal of Chinese Linguistics*, *5*(2), 237–260.

Mikheev, A. 1997. Automatic rule induction for unknown word guessing. *Computational Linguistics*, *23*, 405–423.

Myers, J., & Tsay, J. (2004). Exploring performance-based predictors of phonological judgments in Mandarin. *Poster presented at Laboratory Phonology*, *9*.

Myers, J., & Tsay, J. (2005). The processing of phonological acceptability judgments. Paper presented at the Proceedings of symposium on 90–92 NSC projects.

Ohala, J. J. (1978). Production of tone. In Tone: A Linguistic Survey (pp. 5–39): Elsevier. DOI: https://doi.org/10.1016/B978-0-12-267350-4.50006-6

Sagart, L. (1999). The origin of Chinese tones. Paper presented at the Proceedings of the Symposium/Cross-Linguistic Studies of Tonal Phenomena/Tonogenesis, Typology and Related Topics.

Schneider, W., Eschman, A., & Zuccolotto, A. (2012). *E-Prime User's Guide*. Pittsburgh: Psychology Software Tools Inc.

Smolensky, P., & Legendre, G. (2006). *The harmonic mind: From neural computation to optimality-theoretic grammar (Cognitive architecture)*, Vol. 1. MIT press.

Tseng, S.-C. (2019). ILAS Chinese Spoken Language Resources. Paper presented at the Proceedings of LPSS 2019–the third International Symposium on Linguistic Patterns in Spontaneous Speech, Taipei.

Wang, H. S. (1998). An experimental study on the phonotactic constraints of Mandarin Chinese. *Studia Linguistica Serica*, 259–268.

Wang, L. (1972). *Chinese Phonology [Hànyǔ Yīnyùn Xué]*. Hong Kong: Zhong Hua Shuju.

Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag. DOI: https://doi.org/10.1007/978-0-387-98141-3

Wiener, S., & Turnbull, R. (2016). Constraints of tones, vowels and consonants on lexical selection in Mandarin Chinese. *Language and Speech*, *59*(1), 59–82. DOI: https://doi.org/10.1177/0023830915578000

Wilson, C., & Gallagher, G. (2018). Accidental gaps and surface-based phonotactic learning: A case study of South Bolivian Quechua. *Linguistic inquiry*, *49*(3), 610–623. DOI: https://doi.org/10.1162/ling_a_00285

Wu, F., & Kenstowicz, M. (2015). Duration reflexes of syllable structure in Mandarin. *Lingua*, *164*, 87–99. DOI: https://doi.org/10.1016/j.lingua.2015.06.010

Xu, Y. (2013). ProsodyPro—A tool for large-scale systematic prosody analysis. In: *Tools and Resources for the Analysis of Speech Prosody*. (pp. 7–10). Laboratoire Parole et Langage, France: Aix-en-Provence, France.

Yip, M. (2002). *Tone*. Cambridge University Press. DOI: https://doi.org/10.1017/CBO9781139164559

Zhang, J. (2000). Phonetic duration effects on contour tone distribution. Paper presented at the PROCEEDINGS-NELS.

Zhang, J. (2001). *The effects of duration and sonority on contour tone distribution–typological survey and formal analysis*. (Ph.D), UCLA.

Zuraw, K. (2000). *Patterned exceptions in phonology*. (Ph.D), UCLA.

Zuraw, K. (2002). Aggressive reduplication. *Phonology*, *19*(3), 395–439. DOI: https://doi.org/10.1017/S095267570300441X

Zuraw, K., & Hayes, B. (2017). Intersecting constraint families: an argument for Harmonic Grammar. *Language*, *93*(3), 497–548. DOI: https://doi.org/10.1353/lan.2017.0035