Gradient Acceptability in Mandarin Nonword Judgment

Shuxiao Gong, Jie Zhang
University of Kansas

1 Introduction

The study of phonotactics investigates the permissibility of sound combinations in a language. Acceptability judgments for nonexistent forms often serve as a proxy for the phonotactic grammar as they provide data about how native speakers generalize beyond linguistic forms they previously encountered (Myers, 2017; Sprouse, 2018). Results from these judgment tasks have shown that speakers possess phonotactic knowledge and use such knowledge to offer gradient acceptability ratings on novel words. For example, the acceptability of an English non-word like 'blick' is higher than 'bwick', and 'bnick' will be fully unacceptable (Chomsky & Halle, 1965). Using data from a syllable well-formedness judgment experiment, this paper explores the nature of gradient phonotactic acceptability in Mandarin, a language with considerably less complex syllable structure and phonotactics than English. The results showed that, in Mandarin, phonotactic judgment is also gradient, and the gradience is mainly explained by a number of grammatical factors such as principled phonotactic constraints, allophonic restrictions, and syllable-tone co-occurrence patterns.

1.1 *The Sources of Gradient Non-word Acceptability* In the past fifty years, numerous studies have come to suggest that phonotactic knowledge is gradient. Proposals to explain the gradience have argued that it arises from universal grammatical principles, particularly sonority sequencing profiles in consonant clusters (Berent, Steriade, Lennertz, & Vaknin, 2007; Coleman & Pierrehumbert, 1997) and similarity avoidance effects from the Obligatory Contour Principle (Frisch, Pierrehumbert, & Broe, 2004; Frisch & Zawaydeh, 2001). Non-words violating such principles will be marked as 'ungrammatical' or judged to be less acceptable than those that do not (Chomsky & Halle, 1965).

Another camp of arguments proposes that acceptability asymmetries among unattested structures originate from lexical statistics that measures how similar a non-word is to existing lexical entries (Daland et al., 2011). Some of these lexical statistics models are built on the segmental level. For example, the phonotactic probability of a non-word, as measured by Vitevitch & Luce (2004), calculates the cumulative bi-phone transitional probability; and the neighborhood density of a non-word, as defined in Bailey & Hahn (2001), counts the number of words generated by substituting, deleting, or adding a single phoneme. These measures are then used to predict the phonotactic acceptability of the non-word. Other models assume that phonotactic generalizations are extracted over features or natural classes. For example, both [bn] and [bd] are unattested onset sequences in English, but [bn] is judged better than [bd], because the combination of the natural classes [−continuant][+sonorant] is more frequent than [−continuant][−continuant] (Albright, 2009; Hayes & Wilson, 2008).

However, the boundary between grammatical explanations and lexical statistics is not always clear. For example, Daland et al. (2011) showed that the grammatical principle of sonority sequencing was learnable from the English lexicon, provided that the phonotactic learning algorithm can access the syllable structure and is able to generalize over features. Therefore, gradient acceptability perhaps is the result of the interaction between the grammatical and lexical factors (Coetzee, 2008; Shademan, 2007; White & Chiu, 2017).

There are additional factors that have not been systematically investigated previously in phonotactics research, but could also provide sources of gradience in phonotactic judgment. First, the studies cited above have all discussed phonotactic restrictions held on the *phonemic* level, and few have looked into the phonotactic effects of *allophonic* distributions. For instance, in English, plosives in the exclusive onset position are aspirated (e.g., [**p**$^{\mathbf{h}}$ik] *peak*), and they become unaspirated in onset sC clusters (e.g., [s**p**ik] *speak*). How will native English speakers respond to non-words violating such allophonic restrictions (e.g., *[s**p**$^{\mathbf{h}}$ik])? Second, most of the work on phonotactics has focused on segmental phonotactics, and we know

Proceedings of *AMP 2019*

preciously little about how suprasegmental properties, such as lexical tones, contribute to acceptability judgments.

In this paper, we first review the phonotactic properties of Mandarin, showing that Mandarin is an ideal language to investigate the contribution of the above factors to gradient acceptability. Experimental evidence from a syllable acceptability judgment task is then provided to support the hypothesis that grammatical factors like systematic phonotactic constraints, allophonic restrictions, syllable-tone co-occurrence constraints, are able to account for the variation in acceptability. Finally, the results also suggest that the effects of the grammatical factors cannot be reduced to lexical statistics as measured by neighborhood density.

1.2 *Mandarin Phonotactics and Allophony* To investigate the effects of aforementioned factors on speakers' gradient phonotactic judgment, Mandarin was chosen as the target language due to the following phonological properties of the language. First, Mandarin presents a clearer boundary between systematic gaps and accidental gaps than languages like English. To form a Mandarin syllable, consonants, glides, and vowels are drawn to fill in a restrictive syllable structure named CGVX (Duanmu, 1990, 2007). For example, in the word [tʰjen] 'sky', [tʰ] is the onset C, [j] is the glide G, [e] is the main vowel V, [n] is a coda X.

(1) Mandarin sound inventory
Onset **C**onsonant: p pʰ m f t tʰ n l ts tsʰ s tʂ tʂʰ ʂ ʐ̩ tɕ tɕʰ ɕ k kʰ x
Glide: j w ɥ
Surface **V**owel: i u y e ə o a ɑ
E**X**tra Ending Sound: i u n ŋ

Numerous proposals on principled phonotactic constraints in Mandarin have been made (Duanmu, 2007; Lin, 1989; Wiese, 1997; Yip, 1989). This study adopts the following four constraints adapted from Yi & Duanmu (2015) as the principled constraints.

(2) Yi & Duanmu's phonotactic constraints on Mandarin syllables
a. *HH: The vowel feature [+high] cannot occur in succession (e.g., *[l**ui**] *[t**yu**]).
b. *[Cor]_[Cor]: [Cor] cannot occur in both G and X (e.g., *[**ja**i] *[p**je**i]).
c. *[Lab]_[Lab]: [Lab] cannot occur in both G and X (e.g., *[**wo**u] *[n**wa**u]).
d. Identical articulators cannot occur in succession in C and G (e.g., *[**tʂj**an] *[**pw**aŋ]).[1]

Here, sounds bearing the [+high] feature are the three glides plus the three high vowels [j w ɥ i u y]. The natural class [Cor] includes [t tʰ n l ts tsʰ s tʂ tʂʰ ʂ ʐ̩ tɕ tɕʰ ɕ j ɥ i y], and [Lab] includes [p pʰ m f w ɥ u y o]. All constraints in (2) are varieties of the Obligatory Contour Principle, which is crosslinguistically well-attested and has been linked to the potential difficulty in the production planning of adjacent similar sounds (e.g., Frisch Pierrehumbert, & Broe, 2004). We therefore consider the violation of these four constraints as characterizing systematic gaps in Mandarin.

In the CGVX structure, only the vowel is obligatory. The factorial combination of all sounds plus empty slots gives rise to (21+1) * (3+1) * 8 * (4+1) = 3,520 possible syllables, among which 384 are existing syllables (Z. Chen & Li, 1994) and 3,136 are missing syllables. The four constraints together can rule out nearly 70% of the missing syllables in Mandarin (Yi & Duanmu, 2015). This is very different from English, in which a large number of constraints are needed to account for a good portion of the missing syllables.

Second, Mandarin has rich allophonic variations in vowels and hence provides plenty of opportunities to investigate the contribution of allophonic restrictions in the phonotactic grammar. There are multiple analyses of the Mandarin sound inventory, both in terms of surface phones and underlying phonemes. Lin (2007) proposed 22 consonants /p pʰ m f t tʰ n l ts tsʰ s tʂ tʂʰ ʂ ʐ̩ tɕ tɕʰ ɕ k kʰ x ŋ/ and 3 glides /j w ɥ/ for Mandarin. The alveolo-palatals [tɕ tɕʰ ɕ] only occur before high vowels [i y] and glides [j ɥ], and they are in complementary distribution with alveolars sibilants [ts tsʰ s], retroflex sibilants [tʂ tʂʰ ʂ], and velars [k kʰ x]. Other analyses treat alveolo-palatals either as allophones of alveolars (Duanmu, 2007), of velars (Chao, 1968), or as an independent series of phonemes (Cheng, 1973; Lin, 2007). This study maintains the

[1] Dentals [t tʰ n l ts tsʰ s] are allowed to combine with the coronal glides [j ɥ], e.g., [tʰjan] 'sky', because [Coronal] is underspecified in dentals and they receive the [Coronal] feature redundantly.

alveolo-palatals' independent phoneme identity because it is unclear, among the three series, which should serve as the underlying representations of these alveolo-palatals.

According to Cheng (1973), Mandarin has 12 surface vowels [ɿ ʅ i y u e ə ɤ o ɛ a ɑ]. Lin (1989) notes that [e o] are lowered to [ɛ ɔ] in open syllables; and therefore adds one more vowel [ɔ] to the surface inventory.[2] The number of vowel phonemes proposed by Chinese phonologists varies from zero (Pulleyblank, 1984) to eight (You, Qian, & Gao, 1980, p. 333). Wan & Jaeger (2003) collected 238 Mandarin phonological speech errors and examined which sounds were more likely to interchange with each other when the vowel's contiguous environment changed due to speech errors. A consistent observation of interchange between two sounds based on different contexts would suggest that they originate from the same category, i.e., belong to the same phoneme. For example, if a velar nasal /ŋ/ is mistakenly added after /a/, the output will be [ɑŋ] with the vowel realized as the allophonically appropriate form, instead of [aŋ]. Since [a] and [ɑ] interchange with each other depending on their differential environments caused by speech errors, they are assumed to derive from the same phoneme. Such substitution patterns suggest a five-vowel system /i y u ə a/ for Mandarin, with [e ə ɤ o ɛ ɔ] belonging to the mid vowel phoneme /ə/; and [ɛ a ɑ] to the low vowel phoneme /a/. Since Wan & Jaeger's (2003) data were based on Taiwan Mandarin, which does not clearly distinguish between dental and retroflex sibilants, their study could not provide direct evidence for the status of the apical vowels [ɿ] and [ʅ], which appear after these two series of sibilants, respectively. However, it is reasonable to believe that [ɿ] and [ʅ] are allophones of the high vowel phoneme /i/, as [i] does not appear after these sibilants and is hence in complementary distribution with these apical vowels (Duanmu, 2007; Li & Zhang, 2017; Lin, 1989, 2007). Based on these reasons, the allophonic rules for Mandarin vowels are shown below in (3). Notice that [ɛ] is an allophone for both /ə/ and /a/.[3]

(3) Mandarin vowel allophony

a. i → ɿ / [+anterior, +fricative] _
b. i → ʅ / [−anterior, +fricative] _
c. i → i / elsewhere
d. ə → o / _ u
e. ə → ɔ / w _ #
f. ə → e / _ i
g. ə → ɛ / j, ɥ _ #
h. ə → ə / _ n, ŋ
i. a → a / _ i, n, #
j. a → ɑ / _ u, ŋ
k. a → ɛ / j, ɥ _ n

Given that the main interest in the allophonic effects here is how they affect phonotactic judgment when they can be heard, to ensure that participants are able to hear the allophonic differences, this study sets aside the tenseness differences among the surface vowels and only selects eight forms [i y u e ə o a ɑ] generated from five underlying vowel phonemes /i y u ə a/. Acoustically, allophone pairs differing in tenseness, [ɔ]/[o], [ɛ]/[e], and [ə]/[ɤ], are more similar than pairs differing in other features, say, [e]/[o] (Howie, 1976). In addition, typologically, lax vowels occur in considerably fewer vowel inventories than their tense counterparts; and tenseness contrasts are also less common than height or backness contrasts. (Gordon, 2016; Ladefoged & Maddieson, 1996). Moreover, due to their limited distribution, [ɿ] and [ʅ] are often analyzed as continuations of their preceding sibilants (Duanmu, 2007; Lin, 2007). Therefore, this study will not put [ɿ] and [ʅ] in inappropriate contexts to create allophonic gaps. The vowel allophonic variations tested in the current study are given in (4).

(4) Mandarin vowel allophony used in this study

a. ə → o / w _ #, or _ u
b. ə → e / j, ɥ _ #, or _ i
c. ə → ə / _ n, ŋ, #
d. a → a / _ i, n, #

[2] The diminutive suffix [ɚ] can merge with the syllable it attaches to and create even more surface vowel forms (Lee & Zee, 2003). These forms will not be addressed here.

[3] Wan and Jaeger (2003) argued that since [ɛ] interchanged with all of the other five mid vowel allophones when the contiguous environment changed in 20 errors, whereas there was only one error that suggested an affiliation between [ɛ] and the low vowel phoneme /a/. Therefore, [ɛ] should be treated as an allophone of the mid vowel, not the low vowel. However, their data and discussion are built on Taiwan Mandarin, which may or may not represent the situation for other Mandarin varieties.

e. a → ɑ / _ u, ŋ
f. a → e / j, ɥ _ n

Third, Mandarin is a tone language with four lexical tones that distinguish meanings, namely high-level (T1), rising (T2), low-dipping (T3), and falling (T4). However, these four tones occasionally do not occur with certain syllables. These missing syllable-tone combinations are known as tonal gaps. Many tonal gaps are the results of historical sound change and can be easily filled by loan words, neologism, and onomatopoetic words (Duanmu, 2011). Tonal gaps are not evenly distributed across the four lexical tones. Most tonal gaps are rising tone gaps, followed by high-level, low-dipping, and falling. The gradient acceptability across the four tones matched the distribution frequency of the four types of tonal gaps (Jin & Lu, 2018). Earlier results showed that, on the one hand, despite the accidental nature, tonal gaps received significantly lower acceptability than real words in non-word judgment tasks; on the other hand, their acceptability was also significantly higher than segmental gaps (Kirby & Yu, 2007; Myers, 2002; S. Wang, 1998). The current study on Mandarin allows us to provide further information on how tonal gaps compare with different types of segmental gaps in acceptability in a language.

Another reason that the current study focuses on Mandarin is that previous research has a heavy focus on English, a language with complex phonotactics and syllable structure and a large syllable inventory. The consequence is that the non-word rating results are highly gradient. Mandarin, on the other hand, has a restrictive syllable structure and a small syllable inventory: around 1,300 syllables. Without considering tonal contrasts (for example, [pā] with a high level tone and [pá] with a rising tone will be counted as the same syllable), this number further reduces to around 400 (Lin, 2007). It would be interesting to investigate how syllable inventory size affects speakers' non-word judgment. With a small inventory of syllables, Mandarin speakers may possess stronger intuitions on what is and what is not a good syllable/word than speakers of Indo-European languages such as English. If gradient acceptability is still observed in Mandarin, it will serve as stronger evidence for the presence of gradience in phonotactic knowledge. Meanwhile, in terms of experiment design and stimulus selection, Mandarin's restrictive syllable structure allows us to enumerate all theoretically possible syllables, which is impractical for English (Duanmu, 2008; Fudge, 1969).

1.3 *Summary* In the literature review above, we have shown that phonetically principled phonotactic constraints, allophonic restrictions, and syllable-tone co-occurrence constraints are all potential sources for gradient phonotactic acceptability, and that the properties of Mandarin phonotactics make it an excellent case study for the effects of these factors. The specific hypotheses on how these factors will impose gradience on Mandarin speakers' phonotactic judgment are as follows.

We first hypothesize that violations of principled phonotactic constraints, provided that they can be motivated on functional and typological grounds, will incur lower acceptability ratings than accidental phonotactic constraint violations. Moreover, given the experimental findings that listeners tend to be less attuned to allophonic differences than phonemic differences (e.g., Jaeger, 1980), and that the processing of lexical tones is disadvantaged compared to segmental information (e.g., Cutler & Chen, 1997), we also hypothesize that violations of allophonic and segmental-tonal cooccurrence restrictions will be more acceptable than principled and accidental phonotactic violations. In addition to these grammatical factors, lexical statistics may also contribute to gradient phonotactic acceptability, and the lexical statistics factor that we investigate is neighborhood density. According to previous studies, neighborhood density is hypothesized to be positively correlated with acceptability (Bailey & Hahn, 2001; Myers & Tsay, 2005; Vitevitch & Luce, 1999). But crucially, we hypothesize that the grammatical effects cannot be subsumed under the neighborhood density effect. The remainder of this paper presents a Mandarin syllable acceptability judgment experiment, a widely implemented test for phonological grammar (Myers, 2017).

2 Methods

2.1 *Participants* Thirty-one native Mandarin speakers (8 males and 23 females; mean age = 24.53 years old, SD = 6.68) born and raised in Northern China were recruited to participate in the current experiment. None of the participants reported any speech or hearing problem.

2.2 *Materials* To generate the stimulus syllables of the experiment, an exhaustive list of all theoretically possible Mandarin syllables (both existing and missing) was first made by the factorial combination of all possible surface sounds under the Mandarin CGVX syllable structure. In this structure, only the vowel is obligatory, so the C, V, and X slots can be filled by Ø. Tonal distinctions were not considered, and all syllables used in the study carried the high-level tone. The factorial combination of all sounds plus empty slots gave rise to (21+1) * (3+1) * 8 * (4+1) = 3,520 possible syllables, among which 384 were existing syllables (Z. Chen & Li, 1994) and 3,136 were missing syllables.[4]

Perceptual illusion and misperception are likely to occur when speakers hear stimuli containing sequences that are phonotactically illegal in their native language, where illegal sequences tend to be assimilated to sequences that are legal (Hallé, Segui, Frauenfelder, & Meunier, 1998; Massaro & Cohen, 1983). For example, Japanese listeners tend to perceive an additional vowel between the consonants in VC_1C_2V sequences when the C_1C_2 sequence is impossible in Japanese (e.g., [ebzo] heared as [ebuzo]) (Dupoux, Hirose, Kakehi, Pallier, & Mehler, 1999; Dupoux, Parlato, Frota, Hirose, & Peperkamp, 2011). Similarly, English listeners were also reported to hear an illusory schwa in illicit onset consonants (e.g., [bnif] heard as [bənif]) (Berent et al., 2007; Pitt, 1998). Speaker's native phonological system may prevent them from accurately perceiving a phonotactically illegal sequence.

To ensure that non-word stimuli are perceived as they are intended, not as legal forms or some other perceptually similar forms, we first ruled out the syllables that may lead to perceptual illusion from consideration based on the following criteria:

(5) No glide distinction before [y]: all glides before the vowel [y] are considered neutralized, i.e., [jy]=[wy]=[ɥy]. Only [jy] was preserved in the possible syllable list.
(6) No [+round] distinction before [u]: the glides [j] and [ɥ] before the vowel [u] are considered neutralized, i.e., [ju]=[ɥu]. Only [ju] was preserved in the possible syllable list.
(7) No distinction between [tɕ] and [tɕj] or between [tɕw] and [tɕɥ]; only [tɕ] and [tɕw] were preserved in the possible syllable list.
(8) No distinction between [oŋ] and [uŋ]. Only [uŋ] was preserved in the possible syllable list.
(9) No distinction between [an] and [aŋ], or between [ɑn] and [ɑŋ]. Only [an] and [ɑŋ] were preserved in the possible syllable list.

These criteria mark 1,273 syllables as indistinguishable from some other syllables. Therefore, the remaining list contains 1,863 missing syllables and 384 existing syllables. According to Chen & Li (1994), among the 384 existing syllables, 63 of them happen not to take the high-level tone; these will be referred to as tonal gaps. The remaining 321 syllables are real words.

The missing syllables were further divided into 434 allophonic gaps, which are gaps that only violate the allophonic rules of Mandarin; 1,041 systematic gaps, which are gaps that violate one or more of the four major phonotactic constraints of Mandarin (Yi & Duanmu, 2015); and 388 other segmental phonotactic gaps, which are the gaps that remain unexplained by the four constraints. These are referred to as segmental accidental gaps.

Table 1 below illustrates the different types of syllables discussed so far. For example, [wei] 'micro' is a real word. [ʐan] is a tonal accidental gap, because [ʐan] with a low-dipping tone is a real word 'to dye', but this syllable cannot bear a high-level tone. [ʂuŋ] is missing and does not violate any constraints listed in (2); therefore, it is a segmental accidental gap. [mui] is a systematic gap because it violates the constraint (2a) 'no adjacent high vowels'. [njeu] is an allophonic gap, because its only problem is the wrong mid vowel allophone, which should be [o] instead of [e]. [ljoi] is a gap violating both Mandarin phonotactics (2b) and an allophonic rule (the mid vowel should be the front vowel [e] before the off-glide [i], instead of being the back vowel [o]). According to the definitions above, it is counted as a systematic gap, not an allophonic gap.

The types shown in bold are the five stimulus groups of this study. 40 syllables were randomly selected for each group as the test stimuli, making a total of 200 stimulus syllables. The stimulus syllables were recorded in a high-level tone by a male native Mandarin speaker with phonetic training in an anechoic chamber. All stimuli were normalized for peak intensity using Praat (Boersma & Weenink, 2017). Pitch and duration were not normalized in order to preserve the naturalness of the stimuli. The stimuli had a mean duration of 555 ms (SD = 75).

[4] Chen & Li (1994) provide a syllable inventory based on 5,060 frequent Chinese characters.

All Possible Syllables (3,520)					
Existing Syllables (384)		Missing Syllables (3,136)			
Real Words (321)	**Tonal Gaps** (63)	**Allophonic Gaps** (434)	**Segmental Accidental Gaps** (388)	**Systematic Gaps** (1,041)	Forms Indistinguishable from Other Forms (1,273)

Table 1 Different types of syllables in Mandarin

2.3 *Procedure* The main task was an auditory syllable well-formedness judgment task for the 200 test stimuli described in the previous section. The test was carried out with the Paradigm software (Tagliaferri, 2005) on a Lenovo laptop. Participants listened to the stimuli using earphones connected to the laptop and were asked to decide how good the test stimuli sound as Mandarin syllables on a Likert scale from 1 (bad) to 7 (good). No written forms were given because no orthographic system can represent allophonic gaps. During each trial, the stimulus was first played, and then seven buttons with 1-7 number tags appeared on the screen, together with a text instruction asking the participants to click on one of the buttons to rate the acceptability of the syllable they just heard. The task was self-paced without any time limit. Between two trials there was a 500 ms pause, and the screen was left blank during the pause. Five practice trials, one for each syllable type, were provided prior to the 200 main stimuli presented in a randomized order. Participants' rating responses were recorded.

2.4 *Data Analysis* One participant's data deviated from all the others: he gave a score of 1 (the lowest rating score) for 196 out of all 200 test items (98%), including many real words. His data were excluded from analysis. To reduce the impact of the varying uses of the rating scale by subjects and to reach better normalization, the raw rating scores were transformed to z-scores for each subject out of all test data of this subject (Cowart, 1997). This may also facilitate the convergence of the computationally intensive mixed-effects models (Bates, Mächler, Bolker, & Walker, 2014).

Numerous studies have shown that neighborhood density plays an essential role in spoken word perception and production (see Vitevitch & Luce, 2016 for a review). Specifically, neighborhood density has an inhibitory effect in lexical decision tasks; stimuli in a dense neighborhood are responded to more slowly due to more competition from their lexical neighbors (Bailey & Hahn, 2001; Yao & Sharma, 2017). For acceptability judgment tasks, neighborhood density is positively correlated with ratings: stimuli in a dense neighborhood are judged as more acceptable because they are more similar to other existing lexical entries (Kirby & Yu, 2007; Myers & Tsay, 2005). Therefore, neighborhood density was introduced as a covariate to represent the lexical statistics effects on non-word judgment in the current study. It is defined as the number of words generated by substituting, deleting, or adding a single phoneme together with their summed frequency (Greenberg & Jenkins, 1964). For example, the form [lat] has abundant lexical neighbors in English (e.g. *cat, lap*), while [zev] has a low neighborhood density. Diphthong vowels were counted as sequences of two phonemes, so that [ai] would have [ei], [a], [i], etc. as its neighbors, but it is not a neighbor of [u] or [y]. Even though the stimulus construction process ignored tonal distinctions, they were taken into consideration when searching for lexical neighbors, as previous work has shown that including tonal neighbors in neighborhood density counts improves the correlation between neighborhood density and reaction time in lexical decision tasks (Yao & Sharma, 2017). Lexical tones were indicated by a digit at the end of each syllable. For example, the form [ku1] would have [ku3] and [ku4] as its neighbors. The neighborhood density was also weighted by each neighbor's homophone density[5] in a lexicon based on 5,060 common Chinese characters (Z. Chen & Li, 1994). For example, the non-word stimulus [pyŋ1] has two neighbors, [pəŋ1] and [pɑŋ1]. According to the list, [pəŋ1] has two homophones and [pɑŋ1] has three. Therefore, the final neighborhood density for [pyŋ1] is 2 + 3 = 5. Allophonic differences were encoded in the lexicon as well, so that [pan1] and [pɑŋ1] were not counted as neighbors, even though underlyingly they are (/pan1/ ~ /paŋ1/).

A tendency that ungrammatical syllables were produced with longer duration than real words was observed (Figure 1). Therefore, syllable duration was introduced as another covariate in the statistic model.

[5] Results of correlation tests suggested that homophone-weighted neighborhood density better correlated with the judgment data than plain neighborhood density.

The z-scores of the rating judgments were then fitted with a mixed-effects linear regression model, using the five stimuli groups (*type*), homophone-weighted neighborhood density (*ND*), and *duration* as independent variables and *item* as a random intercept. For the categorical variable *type*, Real Word was set as the baseline for comparison. *Duration* was also rescaled to z-scores to avoid excessively distinct scaling among the variables. *Participant* was not included in the random effects because the individual variations were already captured by the z-score transformation. The random slope by *type* for *item* was also excluded because the resulting model failed to converge. All analyses were conducted using the *lme4* package (Bates et al., 2014) in R, and p-values were obtained using the *afex* package (Singmann, Bolker, Westfall, & Aust, 2016).

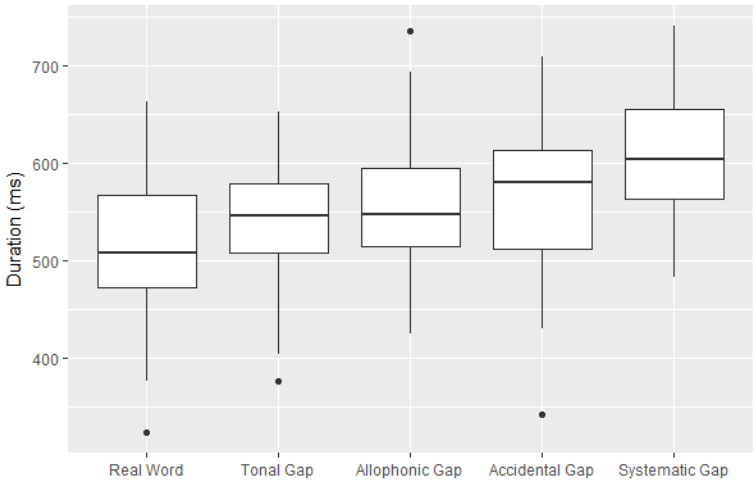


Figure 1 Duration by stimulus types

3 Results

Both forward (starting with the model with only random effects, gradually adding fixed effects and interactions) and backward (starting with the most complicated model with full interactions, gradually deducting fixed effects and interactions) algorithm were attempted for searching for the best model, and the two types of algorithms agreed on the same final model, which includes *type*, *ND*, *duration*, *type* * *ND*, and *type* * *duration*. Its parameter estimates are shown in Table 1.

	Estimate	Std Error	t value	p value
(Intercept)	0.9396	0.1449	6.485	<.0001
Tonal Gap	-1.1405	0.2023	-5.637	<.0001
Allophonic Gap	-1.1582	0.1720	-6.734	<.0001
Accidental Gap	-1.5142	0.1697	-8.924	<.0001
Systematic Gap	-1.7807	0.1746	-10.198	<.0001
Duration	-0.0982	0.0675	-1.455	.1472
Neighborhood Density	0.0019	0.0023	0.826	.4100
Tonal Gap : Duration	-0.1089	0.0993	-1.096	.2744
Allophonic Gap : Duration	0.0371	0.0956	0.388	.6988
Accidental Gap : Duration	-0.0983	0.0983	-1.000	.3186
Systematic Gap : Duration	0.1881	0.0998	1.886	.0609
Tonal Gap : ND	0.0048	0.0039	1.256	.2106
Allophonic Gap : ND	0.0091	0.0066	1.367	.1733
Accidental Gap : ND	0.0111	0.0053	2.089	.0381
Systematic Gap : ND	0.0168	0.0010	1.682	.0942

Table 2 The best model for subjects' acceptability ratings

The effect of *type* stands out even with *ND* and *duration* in the model. Figure 2 illustrates that the acceptability of real words is the highest, followed by tonal gaps, allophonic gaps, accidental gaps, and systematic gaps. A one-way ANOVA using only *type* to predict the ratings was fitted, and the effect of *type* on ratings was significant ($F(4,5995) = 955.97$, $p < .0001$). Post-hoc multiple comparisons with Bonferroni p-value adjustments suggested that the ratings of all five stimulus types were significantly different from each other (all p-values < .0001).

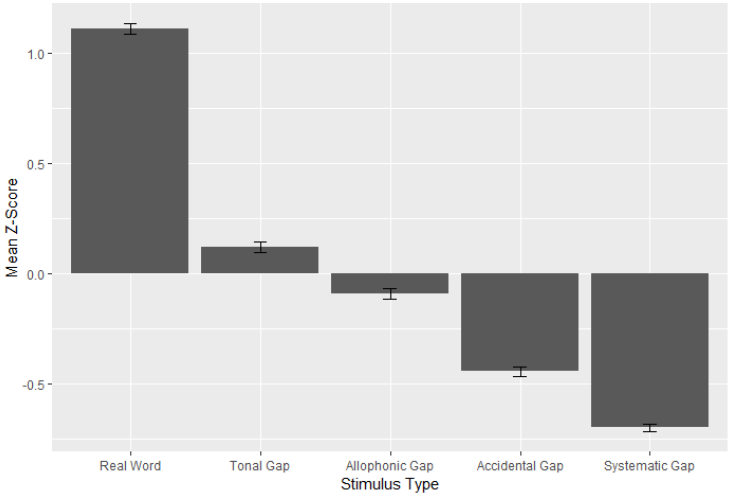
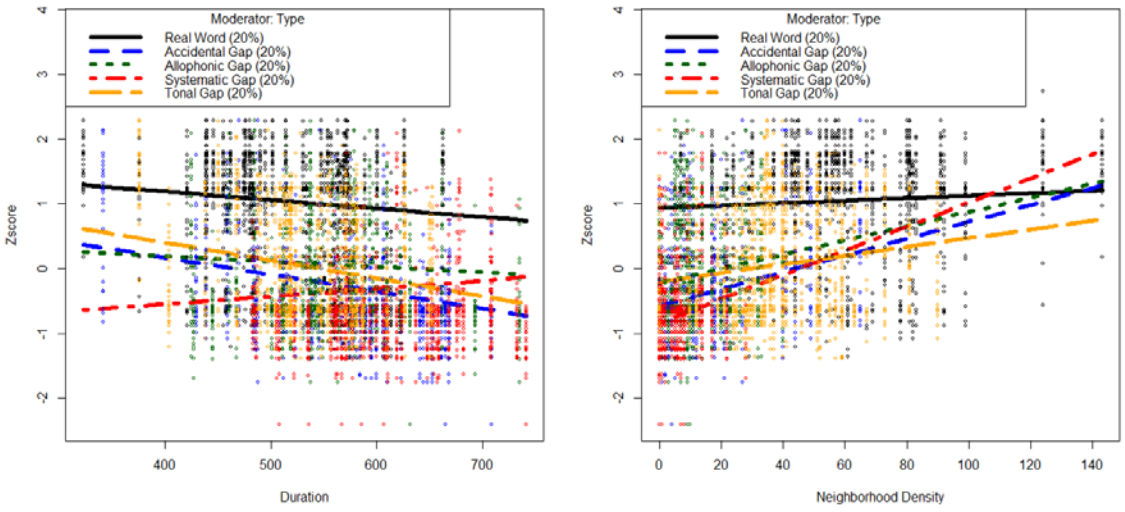


Figure 2 Mean z-scores of well-formedness ratings by stimulus types

The interactions between *type* and *duration* and between *type* and *ND* are illustrated in Figure 3a and Figure 3b, respectively. The effects of *duration* on acceptability ratings vary in different stimulus types, whereas neighborhood density is always positively correlated with the ratings, except that the effect is weaker for real words, consistent with previous findings (Kirby & Yu, 2007; Myers & Tsay, 2005).



(a) (b)

Figure 3 Different effects of (a) duration and (b) neighbourhood density on ratings by stimulus types

Due to the interaction terms, the main effects of the factors cannot be directly interpreted from Table 2. The following analysis of variance table indicates that neighborhood density and duration can explain significant amounts of variances in the model. But more importantly, *type* stands out as the most significant variable, even with duration and neighborhood density in the model.

	Sum Sq.	Mean Sq.	NumDF	DenDF	F value	p value
Type	51.284	12.8211	4	185	28.5454	<.0001
Duration	6.905	6.9049	1	185	15.3733	.0001
Neighborhood Density	4.029	4.0291	1	185	8.9706	.0031
Type:Duration	4.929	1.2323	4	185	2.7437	.0299
Type:Neighborhood Density	3.455	0.8636	4	185	1.9228	.1084

Table 3 Type III analysis of variance table of the ratings model with Satterthwaite's method

4 Discussion

The results of the experiment confirmed that Mandarin speakers' non-word judgment is gradient, and a large portion of such gradience is explained by grammatical factors (the five stimulus types); lexical statistics in the form of neighbourhood density alone cannot explain all the variance. In addition, this study investigated the knowledge of allophonic restrictions that has largely been ignored in previous studies and showed that allophonic gaps behave like neither real words nor systematic gaps, but more similarly to tonal gaps. This indicates that the phonotactic grammar is not blind to allophonic variations, yet in the meantime, speakers are not as sensitive to allophonic restrictions as to principled phoneme-level phonotactic violations even when the allophonic violations can be reliably heard.

Regarding the effect of lexical statistics in the form of neighborhood density, generally, neighborhood density is positively correlated with acceptability ratings, replicating previous findings (Myers & Tsay, 2005). It is interesting to note that the correlation between neighborhood density and the judgment rating is considerably weaker for real words (Bailey & Hahn, 2001). It is possible that due to the high grammaticality of real words, they are directly judged as good without referring to lexical statistics in acceptability judgment. For non-words, the phonotactic grammar does not offer a clear answer, and lexical statistics (i.e., neighborhood density) is consulted in the acceptability judgment. But the point that lexical statistics alone cannot explain all variation in phonotactic judgment stands (Shademan, 2007).

In the previous literature, gradient phonotactic acceptability was mainly reported in Indo-European languages with complicated syllable structures and large syllable inventories. These properties are more likely to induce gradient judgment on non-words because under these circumstances, native speakers' intuition about whether a non-word exists in the lexicon is presumably not clear. For languages with a simple syllable structure and a small syllable inventory, such as Mandarin, native speakers' non-word judgment is likely to be more polarized because it is easier to distinguish words from non-words. Even so, the results of the current experiment still indicate that non-word judgment in Mandarin is gradient, modulated by grammatical factors like principled phonotactic constraints, allophonic restrictions, and syllable-tone co-occurrence restrictions. This provides even stronger support for the gradience of the phonotactic grammar.

Further divisions of the five stimulus types may reveal more fine-grained gradience in acceptability. For example, the tonal gaps in the current study can be divided into two groups. Modern Mandarin sonorant onsets [m n l ʐ] are predominantly derived from Middle Chinese voiced sounds, yet the syllables carrying the high-level tone (Tone 1) in modern Mandarin descended from Middle Chinese syllables with voiceless onsets only (M. Y. Chen, 1976). Consequently, Tone 1 tends not to occur on syllables starting with [m n l ʐ], due to the lack of historical sources. This explains why 40 out of total 63 Tone 1 tonal gaps start with a sonorant onset. The large number of Tone 1 tonal gaps starting with a sonorant onset is an interesting trend in the Mandarin lexicon (Myers, 2007). The O/E ratio of Tone 1 and sonorant onset 'sequences' is only 0.08, indicating that this co-occurrence pattern is highly underrepresented by the lexicon. This led us to examine the acceptability of these sonorant onset Tone 1 gaps separately from other gaps, and the results showed that this lexical bias against T1's cooccurrence with a sonorant onset was indeed noticed by the speakers: the tonal gaps with a sonorant onset were judged to be worse than the other tonal gaps ($t(1198) = -4.9588$, $p < .0001$).

In addition, a number of additive effects among different stimulus types can be observed from the rating data. Systematic gaps and accidental gaps may or may not violate allophonic restrictions. For example, the systematic gap [nwau] not only violates the labial co-occurrence constraint (2c), but is also allophonically inappropriate since the low vowel /a/ before the off-glide [u] should surface as the back [ɑ], not the front [a]. The rating data suggest that gaps that do not violate allophonic restrictions are judged to be better than those who do (Figure 4).

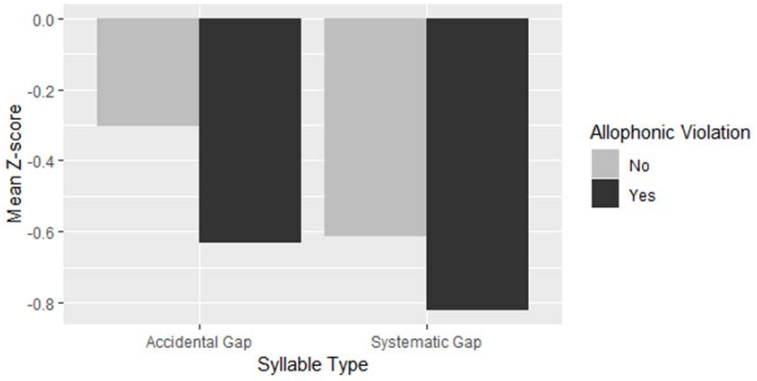


Figure 4 Ratings of accidental and systematic gaps by allophonic violation

Within allophonic gaps, if we fixed an allophonic gap by using the correct allophone instead, the result would be either a real word (e.g., [xwə] → [xwo], with Tone 1) or a tonal gap (e.g., [ʐau] → [ʐɑu], with Tone 1). Our results showed that allophonic gaps that can be fixed as real words are more acceptable than those that are fixed as tonal gaps ($t(1198) = 4.4219$, $p < .0001$), suggesting an additive effect between the violations of allophonic restrictions and segmental-tonal co-occurrence restrictions. This finding also suggests that the dispreference for allophonic gaps compared to real words cannot simply be due to the dispreference for tonal gaps.

5 Conclusion

Using syllable well-formedness judgment as the experimental paradigm, this study explores the nature of Mandarin speakers' non-word acceptability judgment. Five types of syllables were tested in an acceptability judgment experiment: real words, tonal gaps, allophonic gaps, segmental accidental gaps, and systematic gaps. Stimulus type was used as a fixed factor to predict participants' rating data; the duration and neighborhood density of the stimuli were also added into the model as covariates. The results showed that for Mandarin, a language with a comparatively simple syllable structure, non-word judgment is also gradient. Non-words were judged to be significantly poorer than real words, and within non-words, the acceptability varied across the stimulus types. Tonal gaps and accidental gaps patterned together and received significantly higher acceptability ratings than accidental gaps, and systematic gaps received the lowest acceptability rating. The regression model suggests that even after duration and neighborhood density were taken into account, the stimulus type still stood out as a significant predictor for non-word judgment. The lower acceptability of systematic gaps than other types of gaps indicates that some functionally and typologically grounded phonotactic constraints guide the speakers to distinguish systematic gaps from accidentally missing forms. Additional analyses among the stimulus types also revealed further gradient and additive effects on acceptability judgments. Finally, this study makes an additional novel contribution by including non-words violating allophonic restrictions as test stimuli. The fact that the acceptability of allophonic gaps is significantly lower than real words suggests that the phonotactic grammar is surface-based and sensitive to allophonic restrictions.

References

Albright, A. (2009). Feature-based generalisation as a source of gradient acceptability. *Phonology*, *26*(01), 9.

Bailey, T. M., & Hahn, U. (2001). Determinants of Wordlikeness: Phonotactics or Lexical Neighborhoods? *Journal of Memory & Language*, *44*(4), 568.

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2014). *Fitting linear mixed-effects models using lme4*. Retrieved from https://arxiv.org/abs/1406.5823

Berent, I., Steriade, D., Lennertz, T., & Vaknin, V. (2007). What we know about what we have never heard: Evidence from perceptual illusions. *Cognition*, *104*(3), 591–630.

Boersma, P., & Weenink, D. (2017). *Praat: doing phonetics by computer*. Retrieved from http://www.praat.org

Chao, Y.-R. (1968). *A Grammar of Spoken Chinese*. Berkeley and Los Angeles: University of California Press.

Chen, M. Y. (1976). From Middle Chinese to Modern Peking. *Journal of Chinese Ling*, *4*(2), 113–277.

Chen, Z., & Li, X. (Eds.). (1994). *Putonghua Jichu Fangyan Jiben Cihui Ji (Yuyin Juan) "Fundamental Vocabulary of Basic Mandarin Dialects (Pronunciation Volume)."* Beijing: Yuwen Chubanshe.

Cheng, C.-C. (1973). *A Synchronic Phonology of Mandarin Chinese*. The Hague: Mouton.

Chomsky, N., & Halle, M. (1965). Some controversial questions in phonetical theory. *Journal of Linguistics*, *1*(2), 97–138.

Coetzee, A. W. (2008). Grammaticality and Ungrammaticality in Phonology. *Language*, *84*(2), 218–257.

Coleman, J., & Pierrehumbert, J. B. (1997). Stochastic phonological grammars and acceptability. *Computational Phonology Third Meeting of the ACL Special Interest Group in Computational Phonology*.

Cowart, W. (1997). *Experimental Syntax: Applying Objective Methods to Sentence Judgments*. London: SAGE Publications Ltd.

Cutler, A., & Chen, H.-C. (1997). Lexical tone in Cantonese spoken-word processing. *Perception & Psychophysics*, *59*(2), 165–179.

Daland, R., Hayes, B., White, J., Garellek, M., Davis, A., & Norrmann, I. (2011). Explaining sonority projection effects. *Phonology*, *28*(2011), 197–234.

Duanmu, S. (1990). *A Formal Study of Syllable, Tone, Stress, and Domain in Chinese Languages*. Massachusetts Institute of Technology.

Duanmu, S. (2007). *The Phonology of Standard Chinese* (2nd ed.). Oxford ; New York: Oxford University Press.

Duanmu, S. (2008). The "spotty-data problem" and boundaries of grammar. In *Interfaces in Chinese Phonology* (pp. 261–278). Taipei.

Duanmu, S. (2011). Chinese Syllable Structure. In M. van Oostendorp, C. J. Ewen, E. Hume, & K. Rice (Eds.), *The Blackwell Companion to Phonology* (Vol. 5). Blackwell Publishing.

Dupoux, E., Hirose, Y., Kakehi, K., Pallier, C., & Mehler, J. (1999). Epenthetic vowels in Japanese: A perceptual illusion? *Journal of Experimental Psychology: Human Perception and Performance*.

Dupoux, E., Parlato, E., Frota, S., Hirose, Y., & Peperkamp, S. (2011). Where do illusory vowels come from? *Journal of Memory and Language*.

Frisch, S. A., Pierrehumbert, J. B., & Broe, M. B. (2004). Similarity avoidance and the OCP. *Natural Language and Linguistic Theory*, *22*(1), 179–228.

Frisch, S. A., & Zawaydeh, B. A. (2001). The Psychological Reality of OCP - Place in Arabic. *Language*, *77*(1), 91–106.

Fudge, E. (1969). Syllables. *Journal of Linguistics*, *5*(2), 253–286.

Gordon, M. (2016). *Phonological Typology*. Oxford: Oxford University Press.

Greenberg, J. H., & Jenkins, J. J. (1964). Studies in the psychological correlates of the sound system of American English. *Word*, *20*(July), 157–177.

Hallé, P. A., Segui, J., Frauenfelder, U., & Meunier, C. (1998). Processing of Illegal Consonant Clusters: A Case of Perceptual Assimilation? *Journal of Experimental Psychology: Human Perception and Performance*, *24*(2), 592–608.

Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, *39*(3), 379–440.

Howie, J. M. (1976). *Acoustical Studies of Mandarin Vowels and Tones*. Cambridge: Cambridge University Press.

Jaeger, J. J. (1980). Testing the Psychological Reality of Phonemes. *Language and Speech*, *23*(3), 233–253.

Jin, S., & Lu, Y. (2018). Accidental Gaps in Mandarin Tones. *176th Meeting of the Acoustic Society of America*.

Kirby, J. P., & Yu, A. C. L. (2007). Lexical and phonotactic effects on wordlikeness judgments in Cantonese. *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS 2007)*, *21*(August), 1389–1392.

Ladefoged, P., & Maddieson, I. (1996). *The Sounds of the World's Languages*. Oxford: Blackwell Publishing.

Lee, W.-S., & Zee, E. (2003). Standard Chinese (Beijing). *Journal of the International Phonetic Association Journal of the International Phonetic Association Journal of the International Phonetic Association*, *33*(33), 109–112.

Li, M., & Zhang, J. (2017). Perceptual distinctiveness between dental and palatal sibilants in different vowel contexts and its implications for phonological contrasts. *Laboratory Phonology*, *8*(18), 1–27.

Lin, Y.-H. (1989). *Autosegmental treatment of segmental processes in Chinese phonology*.

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

Massaro, D. W., & Cohen, M. M. (1983). Phonological context in speech perception. *Perception & Psychophysics*, *34*(4), 338–348.

Myers, J. (2002). An analogical approach to the Mandarin syllabary. *Journal of Chinese Phonology*, *11*(Special Issue), 163–190.

Myers, J. (2007). Bridging the gap: MiniCorp analyses of Mandarin phonotactics. In R. Colavin, K. Cooke, K. Davidson, S. Fukuda, & A. Del Guidice (Eds.), *Proceedings of the thirty-seventh Western Conference on Linguistics* (pp. 137–147). San Diego: University of California, San Diego.

Myers, J. Acceptability Judgments. , Oxford research encyclopedia of linguistics § (2017).

Myers, J., & Tsay, J. (2005). The processing of phonological acceptability judgments. *Proceedings of Symposium on 90-92 NSC Projects*, 26–45.

Pitt, M. A. (1998). Phonological processes and the perception of phonotactically illegal consonant clusters. *Perception and Psychophysics*, *60*(6), 941–951.

Pulleyblank, E. G. (1984). Vowelless Chinese? An Application of the Three-Tiered Theory of Syllable Structure. *Sixteenth International Conference on Sino-Tibetan Languages and Linguistics*. Seatle: University of Washington.

Shademan, S. (2007). Grammar and analogy in phonotactic well-formedness judgments. UCLA.

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2016). *afex: Analysis of Factorial Experiments*. Retrieved from https://cran.r-project.org/package=afex

Sprouse, J. (2018). Acceptability judgments and grammaticality, prospects and challenges. In N. Hornstein, H. Lasnik, P. Patel-Grosz, & C. Yang (Eds.), *Syntactic Structures after 60 Years* (pp. 195–224). Berlin: De Gruyter Mouton.

Tagliaferri, B. (2005). *Paradigm*. Retrieved from http://www.paradigmexperiments.com

Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language*, *40*(3), 374–408.

Vitevitch, M. S., & Luce, P. A. (2004). A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 481–487.

Vitevitch, M. S., & Luce, P. A. (2016). Phonological Neighborhood Effects in Spoken Word Perception and Production. *Annual Review of Linguistics*, *2*(1), annurev-linguist-030514-124832.

Wan, I.-P., & Jaeger, J. (2003). The Phonological Representation of Taiwan Mandarin Vowels: A Psycholinguistic Study. *Journal of East Asian Linguistics*, *12*, 205–257.

Wang, S. (1998). An experimental study on the phonotactic constraints of Mandarin Chinese. In B. K. T'sou (Ed.), *Studia Linguistica Serica* (pp. 259–268). Hong Kong: Language Information Sciences Research Center, City University of Hong Kong.

White, J., & Chiu, F. (2017). Disentangling phonological well-formedness and attestedness: An ERP study of onset clusters in English. *Acta Linguistica Academica*, *64*(4), 513–537.

Wiese, R. (1997). Underspecification and the description of Chinese vowels. In J. Wang & N. Smith (Eds.), *Studies in Chinese Phonology* (pp. 219–249). Berlin; New York: Mouton de Gruyter.

Yao, Y., & Sharma, B. (2017). What is in the neighborhood of a tonal syllable? Evidence from auditory lexical decision in Mandarin Chinese. *Proceedings of the Linguistic Society of America*, *2*, 45.

Yi, L., & Duanmu, S. (2015). Phonemes, Features, and Syllables: Converting Onset and Rime Inventories to Consonants and Vowels. *Language and Linguistics*, *16*(6), 819–842.

Yip, M. (1989). Feature geometry and cooccurrence restrictions. *Phonology*, *6*(2), 349–374.

You, R., Qian, N., & Gao, Z. (1980). Lun Putonghua de Yinwei Xitong. *Zhongguo Yuwen*.