Testing the Role of Phonetic Knowledge in Mandarin Tone Sandhi

Jie Zhang[a]

Yuwen Lai[b]

[a]Department of Linguistics

The University of Kansas, USA

[b]Department of Foreign Languages and Literatures

National Chiao Tung University, Taiwan

Testing the Role of Phonetic Knowledge in Mandarin Tone Sandhi*

Abstract

Phonological patterns often have phonetic bases. But whether phonetic substance should be encoded in synchronic phonological grammar is controversial. We aim to test the synchronic relevance of phonetics by investigating native Mandarin speakers' applications of two exceptionless tone sandhi processes to novel words: the contour reduction 213→21/__T (T≠213), which has a clear phonetic motivation, and the perceptually neutralizing 213→35/__213, whose phonetic motivation is less clear. In two experiments, Mandarin subjects were asked to produce two individual monosyllables together as disyllabic words that are different types of novel words. Results show that speakers apply the 213→21 sandhi with a greater accuracy than the 213→35 sandhi in novel words, indicating a synchronic bias against the phonetically less motivated pattern. We also show that lexical frequency is relevant to the application of the sandhis to novel words, but it alone cannot account for the low sandhi accuracy of 213→35.

* This work could not have been done without the help of many people. We are grateful to Paul Boersma and Mietta Lennes for helping us with Praat scripts, Juyin Chen, Mickey Waxman, and Xiangdong Yang for helping us with statistics, and Hongjun Wang, Jiangping Kong, and Jianjing Kuang for hosting us at Beijing University during our data collection in 2007. For helpful comments on various versions of this work, we thank Allard Jongman, James Myers, three anonymous reviewers and an Associate Editor for *Phonology*, and audiences at the 2004 NYU Workshop on "Redefining Elicitation," Department of Linguistics at the University of Hawaii, Department of Psychology and the Child Language Program at the University of Kansas, and the 2005 Annual Meeting of the Linguistic Society of America. We owe a special debt to Hsin-I Hsieh, whose work on wug-testing Taiwanese tone sandhi (Hsieh 1970, 1975, 1976) inspired this research. All remaining errors are our own. This research was partly supported by research grants from the National Science Foundation (0750773) and the University of Kansas General Research Fund (2301760).

1. Introduction

1.1. The relevance of phonetics to phonological patterning

Phonological patterns are often influenced by phonetic factors. The influence manifests itself in a number of ways, the staple of which is the prevalence of patterns that have articulatory or perceptual bases and the scarcity of those that do not in crosslinguistic typology. For example, velar palatalization before high front vowels, postnasal voicing, and regressive assimilation for major consonant places have clear phonetic motivations and are extremely well-attested, while velar palatalization before low back vowels, postnasal devoicing, and progressive consonant place assimilation are nearly nonexistent. The typological asymmetry can also be manifested in terms of implicational statements. For example, in consonant place assimilation, if oral stops are targets of assimilation in a language, then *ceteris paribus*, nasal stops are also targets of assimilation (Mohanan 1993, Jun 1995, 2004). This is to be expected perceptually, as nasal stops have weaker transitional place cues and are thus more prone to losing their contrastive place than oral stops when articulatory economy is of concern (the Production Hypothesis; Jun 1995, 2004).

Evidence for the relevance of phonetics can also be found in the peripheral phonology of a language even when the phonetic effects are not directly evident in its core phonology. Such peripheral phonology may include the phonology of its established loan words (Fleischhacker 2001, Kang 2003, Kenstowicz 2007) and the speakers' judgments on poetic rhyming (Steriade and Zhang 2001). For example, Steriade and Zhang (2001) showed that although postnasal voicing is not neutralizing in Romanian, its phonetic effect is crucial in accounting for the poets' preference for /Vnt/~/Vnd/ as a semi-rhyme over /Vt/~/Vd/.

The parallels between the traditionally conceived categorical/phonological and gradient/phonetic patterns also indicate their close relation. Flemming (2001), for instance, outlines the similarity of patterning between phonological assimilation and phonetic coarticulation as well as a number of other processes present in both the traditional phonological and phonetic domains.

1.2. Where should phonetic explanations reside?

Although the existence of *some* form of relationship between phonological typology and phonetics is relatively uncontroversial, the precise way in which this relationship should be captured is a continuous point of contention among phonologists. One possibility is to consider the phonetic basis to be part of the intrinsic mechanism of the synchronic phonological grammar. Many theories have been proposed within rule-based phonology to encode this relation, from the abbreviation conventions of *SPE* (Chomsky and Halle 1968), the innateness of articulatory-based phonological processes in Natural Phonology (Stampe 1979), to the grounding conditions for universal constraints in Grounded Phonology (Archangeli and Pulleyblank 1994). Optimality Theory (Prince and Smolensky 1993) further invites phonetic explanations into synchronic phonology due to its ability to state phonetic motivations explicitly in the system as markedness constraints (Hayes and Steriade 2004). Works by Boersma (1998), Steriade (1999, 2001, 2008), Kirchner (2000, 2001, 2004), Flemming (2001), Zhang (2002, 2004), among others, have proposed constraints that directly encode phonetic properties and intrinsic rankings based on such properties in synchronic phonology to capture typological asymmetries. There are various approaches as to how the phonetic substance gets to be encoded in the grammar. The strongest position is that the phonetically based constraints, intrinsic rankings, grounding conditions, or other formal mechanisms are simply part of the design nature of the grammar on the level of the species (Universal Grammar, Chomsky 1986), and they predict true, exceptionless universals of phonological typology. It is also possible that the design scheme of the grammar only includes an analytical bias that draws from a type of grammar-external phonetic knowledge (Kingston and Diehl 1994) and restricts the space of constraints and constraint rankings (weightings) to be learned by the speaker (Hayes 1999, Wilson 2006). This type of approach predicts strong universal tendencies in favor of phonetically motivated patterns, but allows "unnatural" patterns to surface in grammars and be learned by speakers.

An alternative to the synchronic approach above is that the effect of phonetics on phonological typology takes place in the realm of diachronic sound change. The typological asymmetries in phonology are then due to different frequencies with which phonological patterns can arise through diachronic sound change, which is caused by phonetic factors such as misperception (e.g., Anderson 1981, Ohala 1981, 1990, 1993, 1997, Blevins and Garrett 1998, Hale and Reiss 2000, Buckley 2000, 2003, Hansson 2001, Hyman 2001, Blevins 2004, Yu 2004, Silverman 2006a, b).[1] Among researchers

---

[1] There are disagreements as to whether the speaker plays any active role in sound change: for example, Ohala considers sound change to be listener-based and non-teleological, while Bybee's (2001, 2006) usage-based model places great importance on the speaker's production in the

working in this framework, there are different positions on the role of UG in synchronic phonology in general, from categorical rejection (Ohala 1981, 1990, 1993, 1997, Silverman 2006a) to selective permission (Blevins 2004) to utmost importance (Hale and Reiss 2000). But all proponents of this approach agree that Occam's Razor dictates that if a diachronic explanation based on observable facts exists for typological asymmetries in phonological patterning, a UG-based synchronic explanation, which is itself hypothetical and unobservable, is not warranted (e.g., Hale and Reiss 2000: p.158, Blevin 2004: p.23, Hansson 2008: p. 882). For a comprehensive review on the diachronic explanations of sound patterns, see Hansson (2008).

The synchrony vs. diachrony debate is very often centered around the strongest form of the phonetics-in-UG hypothesis. Earlier proponents of the synchronic approach working in the OT framework were primarily concerned with establishing stringent implicational statements on phonological behavior from typological data, discovering the phonetic rationales behind the implications, and proposing Optimality-Theoretic models from which the implicational statements fall out as predictions (e.g., Jun 1995, Steriade 1999, Kirchner 2001, Zhang 2002). Conversely, critics of the synchronic approach, beyond proposing explicit frameworks for the evolution of phonological systems and how perception, and possibly production, may have shaped the evolution, have made efforts to identify counterexamples to the phonetically based typological asymmetries and provide explanations for the emergence of such "unnatural" patterns based on a chain of commonly attested diachronic sound changes (Hyman 2001, Yu 2004, Blevins 2006; see

---

initiation of sound change; Blevins's Evolutionary Phonology (Blevins 2004) also ascribes the speaker with a more active role in sound change than Ohala's model.

also Bach and Harms 1972, Anderson 1981). The debate between Blevins (2006) and Kiparsky (2006) on whether there exist true cases of coda voicing is a case in point. Regardless of the outcome of particular debates, this seems to be a losing battle for the synchronic approach in the long run, as the implicational statements gathered from crosslinguistic typology are necessarily inductive — provided that we have not looked at all languages, we cannot be certain that no counterexamples will ever emerge. Moreover, experimental studies showing that phonetically arbitrary patterns can in fact be readily learned by speakers (Dell et al. 2000, Onishi et al. 2002, Chambers et al. 2003, Buckley 2003, Seidl and Buckley 2005) also seem to provide additional arguments in the diachronic approach's favor.

However, as we have mentioned earlier, a synchronic approach does not necessarily assume the strongest form of phonetics as a design feature. The analytical bias approach (e.g., Wilson 2006, Moreton 2008), for example, only favors the learning of particular patterns in the process of phonological acquisition, but does not in principle preclude the emergence or learning of other patterns. Pitched against the diachronic approach, the form of argument from either approach should come from experimental studies that show whether speakers indeed exhibit any learning biases in favor of phonetically motivated patterns, not whether phonetically unmotivated patterns can be learned at all.[2]

---

[2] Hansson (2008: p.882) argues that the substantive bias approach is not necessarily incompatible with the diachronic approach, as such biases can be considered as one of the potential sources of “external errors” in language evolution. But we do think that there is a fundamental difference between the two approaches in terms the importance ascribed to phonetics in the grammar and the

Relatedly, there is a crucial difference between phonological patterns observed in a language and the speaker's internal knowledge of such patterns. Many recent works have shown that speakers may know both more and less than the lexical patterns of their language. For example, Zuraw (2007) demonstrated that Tagalog speakers possess knowledge of the splittability of word-initial consonant clusters that is absent in the lexicon, but projectable from perceptual knowledge, and that they can apply the knowledge to infixation in stems with novel initial clusters; Zhang and Lai (2008) and Zhang et al. (2009a, b), *pace* earlier works on Taiwanese tone sandhi such as Hsieh (1970, 1975, 1976) and Wang (1993), showed that the phonologically opaque "tone circle" is largely unproductive in wug tests despite its exceptionlessness in the language itself. This opens up a new area of inquiry for the synchrony vs. diachrony debate: provided that one is interested in the tacit knowledge of the speaker, then one needs to look beyond the typological patterns and see which approach provides a better explanation for experimental results that shed light on the speakers' internal knowledge. One likely fruitful comparison is to see whether there are productivity differences between two patterns that differ in the level of phonetic motivation, but are otherwise comparable.[3]

---

extent to which answers to asymmetries in both typological patterns and speakers' internal phonological knowledge lie in the formal grammatical module.

[3] Hansson (2008: p.881) worries that it would be "all too easy to explain away apparent counterexamples … as being lexicalized, morphologized, or in some other way not belonging to the 'real' phonology of the language." But one should insist that any claims about whether a pattern falls outside the "real" phonology of a language be supported by experimental evidence.

1.3. Experimental studies addressing the role of phonetics in learning

We provide a brief review of the relevant experimental literature on the role of phonetics in different types of learning situations in this section.

One possible line of investigation is to see that in a language with phonological patterns that differ in the degree of phonetic motivation, whether the patterns with stronger phonetic motivations are acquired more quickly and with greater accuracy in language acquisition.

The claim that phonetically motivated morphophonological processes are acquired earlier and with fewer errors has been made in the literature (e.g., MacWhinney 1978, Slobin 1985, Menn and Stoel-Gammon 1995). For example, Slobin (1985) compared the effortless acquisition of final devoicing by Turkish children and the error-ridden acquisition of stop-spirant alternation in Modern Hebrew by Israeli children and suggested that there is a hierarchy of acceptable alternation based on universal predispositions that favor assimilation and simplification in the articulatory output (p.1209).

Buckley (2002) pointed out that the role of such universal predispositions can only be established if the accessibility of the pattern, such as its distribution and regularity, is teased apart from the phonetic naturalness of the pattern. In demonstrating that many unnatural patterns are acquired readily due to their high regularity and high frequency of occurrence while many natural patterns are acquired with much difficulty

---

Moreover, the argument is more likely in a subtler form of whether there are any detectable differences between patterns, not whether a pattern is categorically in or out of “real” phonology.

due to their low accessibility, Buckley argued that accessibility, but not phonetic naturalness, determines the ease of learning. However, Buckley (2002) did not show that when accessibility is matched, a process lacking phonetic motivations can be acquired just as easily as a phonetically motivated process. The only such comparisons that can be found in Buckley (2002) are in Hungarian — the more natural backness harmony vs. the less natural /a/-lengthening, both of which are highly accessible, and the more natural rounding harmony vs. the less natural /e/-lengthening, both of which have low accessibility. MacWhinney (1978)'s original work, which Buckley cited, showed that backness harmony is acquired earlier than /a/-lengthening, but rounding harmony is acquired later than /e/-lengthening. Therefore, phonetic naturalness does seem to affect the order of acquisition, but it is unclear what the precise effect is. Moreover, given that these comparisons are only made under a crude control of "accessibility," the results cannot be deemed conclusive.

Another approach is to test the learning of patterns with different degrees of phonetic motivations in an artificial language. The artificial grammar paradigm (see Reber 1967, 1989, Redington and Chater 1996) has been widely used to investigate the learnability of phonological patterns in both children and adults. The paradigm typically involves two stages — the exposure stage, in which the subject is presented with stimuli generated by an artificial grammar, and the testing stage, in which the subject is tested on their learning of the patterns in the artificial grammar, measured by their ability to distinguish legal vs. illegal test stimuli, reaction time, or looking time in the head-turn paradigm for infant studies. It is particularly suited for the comparison of the learning of different patterns, as the relevant patterns can be designed to have matched regularity, lexical frequency, and transitional probability. This line of research has been actively pursued with conflicting results. Seidl and Buckley (2005) reported two experiments that tested whether nine-month-old infants learn patterns with different degrees of phonetic motivation differently. The first experiment tested whether the infants preferred a phonetically grounded pattern in which only fricatives and affricates, but not stops, occur intervocalically, or an arbitrary pattern in which only fricatives and affricates, but not stops, occur word-initially. The second experiment tested the difference between a grounded pattern in which a labial consonant is followed by a rounded vowel and a coronal consonant is followed by a front vowel and an arbitrary pattern in which a labial consonant is followed by a high vowel and a coronal consonant is followed by a mid vowel. In both experiments, the infants learned both patterns fairly well and showed no learning bias towards the phonetically grounded pattern, suggesting that phonetic grounding does not play a role in the learning of synchronic phonological patterns. But in experiment 1, all fricatives and affricates used were stridents, and as Kirchner (2001, 2004) showed, the precise articulatory control necessary for stridents in fact makes them less desirable in intervocalic position. Moreover, Thatte (2007) pointed out that there exist phonological generalizations other than the ones that Seidl and Buckley intended in their stimuli, and the infants might have responded to these generalizations (p.7, fn. 4). Therefore, Seidl and Buckley's claim that there is no learning bias towards phonetically grounded patterns is open to debate. In Jusczyk et al. (2003), 4.5-month-old infants were presented with sets of three words, or "triads," which consisted of two monosyllabic pseudowords in the forms of VC_1 and C_2V, followed by a disyllabic word in which either C_1 or C_2 assimilates in place to the adjacent consonant (*an, bi, ambi; an, bi, andi*). The

C_1-assimilation pattern is perceptually motivated and crosslinguistically extremely common, while the C_2-assimilation pattern has no clear perceptual grounding and crosslinguistically extremely rare. In a head-turn procedure, infants showed no difference in looking time towards the triads with regressive and progressive assimilations, indicating the lack of *a priori* preference for phonetically motivated phonological patterns. However, Thatte (2007)'s study, which compared intervocalic voicing (*pa, fi, pavi*) and devoicing (*pa, vi, pafi*) using a similar methodology, showed that 4.5-month-old infants exhibited a preference for the phonetically motivated intervocalic voicing, while 10.5-month-old infants preferred the phonetically unmotivated intervocalic devoicing. Thatte argued that the 4.5-month-olds' results support the view that infants have an innate preference for phonetically based patterns and tentatively interpreted the 10.5-month-olds' results as the combined effect of their overall lower threshold for boredom and their becoming bored with the phonetically motivated pattern earlier.

In addition to the conflicting results, as Seidl and Buckley (2005) pointed out, the A, B, AB triad procedure is quite novel in infant research, and the assumption that the infants take the AB string as a concatenation of A and B may not be valid. Therefore, the extent to which the phonetic bases of phonological patterns are directly relevant to first language acquisition remains an open question.

Pycha et al. (2003) tested adult English speakers' learning of three non-English patterns — "palatal vowel harmony" (stem and suffix vowels agree in [back]), "palatal vowel disharmony" (stem and suffix vowels disagree in [back]), and "palatal arbitrary" (arbitrary relation between stem and suffix vowels) — and found that although subjects exhibited better learning of the harmony and disharmony patterns than the arbitrary pattern, there was no difference between harmony and disharmony. Taking harmony to have a stronger phonetic motivation than disharmony, they concluded that phonetic naturalness is not relevant to the construction of the synchronic grammar. Wilson (2003), in two similar experiments with similar results, interpreted the results differently, however. Wilson argued that both assimilation and dissimilation can find motivations in phonetics and thus both have a privileged cognitive status in phonological grammar. Wilson (2006) showed that when speakers were presented with highly impoverished evidence of a new phonological pattern, they were able to extend the pattern to novel contexts predicted by a phonetically based phonology and linguistic typology, but not to other contexts; for instance, speakers presented with velar palatalization before mid vowels could extend the process before high vowels, but not *vice versa*. A phonology that encodes no substantive bias cannot predict these experimental observations.

Two experiments on the learning of natural vs. unnatural allophonic rules in an artificial language conducted by Peperkamp and collaborators (Peperkamp et al. 2006, Peperkamp and Dupoux 2007) returned conflicting results. In both experiments, French subjects were exposed to alternations that illustrate intervocalic voicing (e.g., [p, t, k] → [b, d, g] / V__V) or a random generalization (e.g., [p, g, z] → [ʒ, f, t] / V__V). In the test phase, the subjects did not exhibit a learning difference between the two types of alternations in a phrase-picture matching task (Peperkamp and Dupoux 2007), but did show a strong bias in favor of intervocalic voicing in a picture naming task (Peperkamp et al. 2006). Peperkamp and colleagues surmised that the different results might be due to a ceiling effect in the cognitively less demanding phrase-picture matching task, and the difference between natural and unnatural alternations could lie in either the speed with which they are learned — natural alternations are learned faster — or the ease with which they can be used in processing once they have been learned — natural alternations can be used more easily, especially in cognitively demanding tasks. Either way, the account allows random alternations to be learned, but also admits that the phonetic nature of the alternation plays a role in its acquisition.

An additional issue with using the artificial language paradigm in adult research is that the learning at best approximates second language acquisition, whose mechanism is arguably very different from first language acquisition (Cook 1969, 1994, Dulay et al. 1982, Bley-Vroman 1988, Ellis 1994, among others); but the learning issue of interest here is the relevance of phonetics during the construction of *native* phonological grammars. Moreover, the artificial language paradigm often involves a heavy dose of explicit learning, while second language acquisition, like first language acquisition, often involves a significant amount of implicit learning. This furthers the distance between artificial language learning and real language acquisition.

1.4. The current study

The current study complements the experimental works above by using a nonce-probe paradigm ("wug" test) (Berko 1958) with adult speakers. In a typical wug test, subjects are taught novel forms in their language and then asked to provide morphologically complex forms, using the novel forms as the base. This paradigm has been widely used to test the productivity of regular and irregular morphological rules (e.g., Bybee and Pardo 1981, Albright 2002, Albright and Hayes 2003, Pierrehumbert

2006) and morphophonological alternations (e.g., Hsieh 1970, 1975, 1976, Wang 1993, Zuraw 2000, 2007, Albright et al. 2001, Hayes and Londe 2006). Our study wug-tests two patterns of tonal alternation (tone sandhi) that differ in the degree of phonetic motivation in Mandarin Chinese and compares the accuracies with which the sandhi patterns apply to nonce words.

This approach is in line with the assumption that the phonological patterns observed in the language may not be identical to the speakers' knowledge of the patterns and provides us with a novel opportunity to test the role of phonetics in synchronic phonology. It uses real phonological patterns that exist in the subjects' native language, which circumvents the learning-strategy problem with the artificial language paradigm. It also allows easier manipulations of confounding factors such as lexical frequency and thus minimizes the control problem in studying phonological learning in a naturalistic setting.

1.5. Organization of the paper

We discuss the details of two tone sandhi patterns under investigation in Mandarin in §2. The methodology and results for the two experiments that compare the productivity of the two sandhi patterns are discussed in §3 and §4. Theoretical implications of the results are further discussed in §5. §6 is the conclusion.

2. Tone sandhi in Mandarin Chinese and the general hypotheses

Mandarin Chinese is a prototypical tone language. The standard variety of Mandarin spoken in Mainland China, particularly Beijing, has four lexical tones — 55, 35, 213, and 51[4] — as shown in (1).[5] The pitch tracks of the four tones with the syllable [ma] pronounced in isolation by a male speaker, each averaged over five tokens, are given in Figure 1. Although Tone 2 is usually transcribed as a high rising tone 35, there is a small pitch dip at the beginning of the tone, creating a turning point, and research has shown that the perceptual difference between Tone 2 and Tone 3 mainly lies in the timing and pitch height of the turning point (Shen and Lin 1991, Shen et al. 1993, Moore and Jongman 1997). We can also notice in Figure 1 that the different tones have different durational properties; in particular, Tone 3 has the longest duration. These observations will become important in the discussion of Mandarin tone sandhi and the experimental results.

(1) Mandarin tones:

Tone 1:	ma55	'mother'
Tone 2:	ma35	'hemp'
Tone 3:	ma213	'horse'
Tone 4:	ma51	'to scold'

---

[4] Tones are marked with Chao tone numbers (Chao 1948, 1968) here. "5" indicates the highest pitch used in lexical tones while "1" indicates the lowest pitch. Contour tones are marked with two juxtaposed numbers. For example, 51 indicates a falling tone from the highest pitch to the lowest pitch.

[5] The variety of Mandarin spoken in Taiwan has a slightly different tonal inventory: the Third Tone is pronounced without the final rise as [21] even in prosodic final position. This is not the variety of Mandarin studied here.

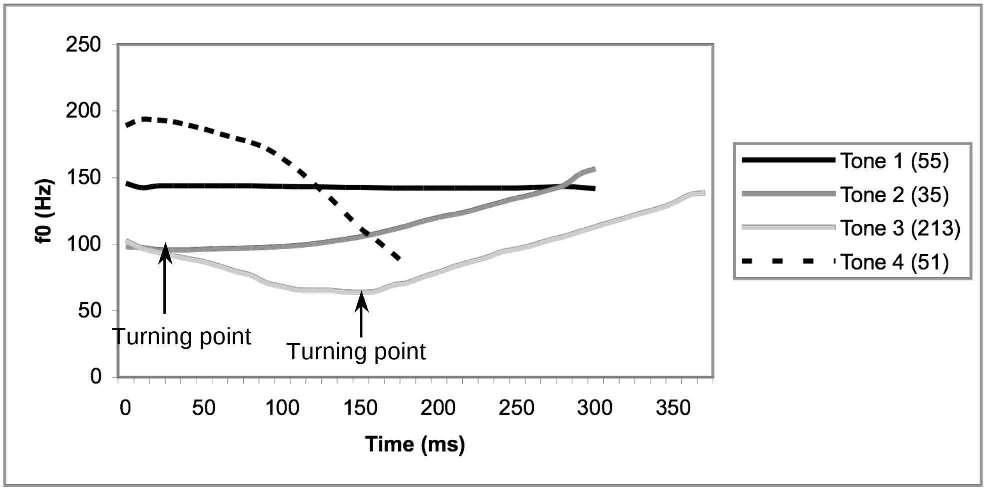


Figure 1. Representative pitch tracks for the four tones in Mandarin.

In tone languages, a tone may undergo alternations as conditioned by adjacent tones or the prosodic and/or morphosyntactic position in which the tone occurs. This type of alternation is often referred to as tone sandhi (Chen 2000, among others). Mandarin Chinese has two tone sandhi patterns, both of which involve the Third Tone 213. Specifically, 213 becomes 35 when followed by another 213 (the “Third-Tone Sandhi”); but 213 becomes 21 when followed by any other tone (the “Half-Third Sandhi”). These sandhis are exemplified in (2).

(2) Mandarin tone sandhi:

a. 213 → 35 / ___ 213

xɑu213-tɕjou213 → xɑu35-tɕjou213 ‘good wine’

b. 213 → 21 / ___ {55, 35, 51}

xɑu213-ʂu55 → xɑu21-ʂu55 ‘good book’

xɑu213-ɻən35 → xɑu21-ɻən35 ‘good person’ xɑu213-kʰan51 → xɑu21-kʰan51 ‘good-looking’

The pitch tracks for the four examples in (2) pronounced in isolation by a male speaker, each averaged over five tokens, are given in Figure 2.

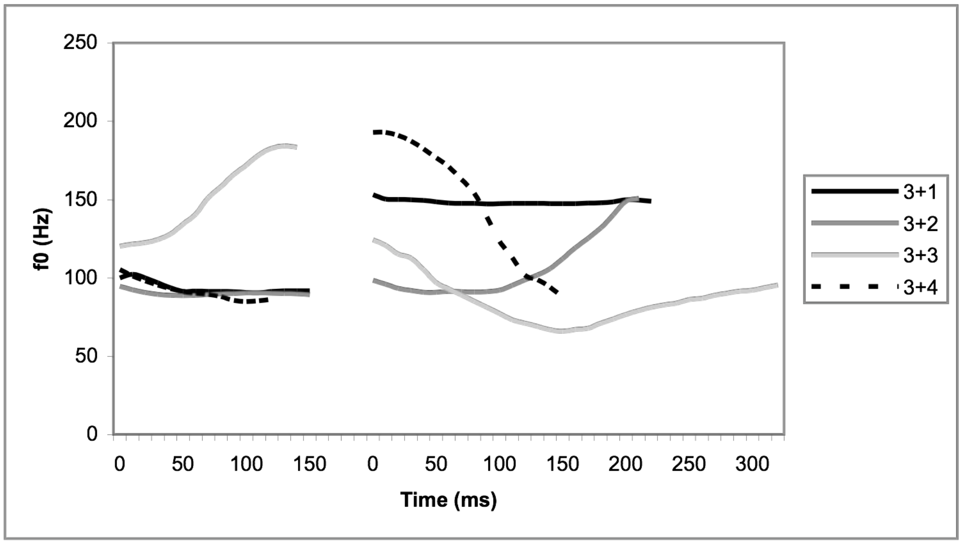


Figure 2. Representative pitch tracks for the tone sandhis in Mandarin.

Both of these sandhi patterns are fully productive in Mandarin disyllabic words and phrases, and they are both “phonological” in the traditional sense, in that they involve language-specific tone changes that cannot be predicted simply by tonal coarticulation. However, we consider the Half-Third Sandhi to have a stronger phonetic basis than the Third-Tone Sandhi. Our judgment is based on the following three reasons.

First, in terms of the phonetic mechanism of the tone change, although both sandhis involve simplification of a complex contour in prosodic nonfinal position, which has articulatory and perceptual motivations (Zhang 2002, 2004), the Third-Tone Sandhi also involves a raising of the pitch, which cannot be accounted for by the phonetic motivation of reducing pitch contours on syllables with insufficient duration. The Half-Third Sandhi, however, only involves truncation of the second half of the contour. The Third-Tone Sandhi is also structure-preserving, at least in perception.[6] Based on the closer phonetic relation between the base and sandhi pair in the Half-Third Sandhi and the contrastive status of the sandhi tone 35 in the Third-Tone Sandhi, we assume that the perceptual distance between the base and sandhi tones in the Half-Third Sandhi is smaller than that in the Third-Tone Sandhi. We take this as an argument for the stronger phonetic motivation for the Half-Third Sandhi (cf. the P-map, Steriade 2001, 2008).

Second, in the traditional Lexical Phonology sense, the Third-Tone Sandhi has lexical characteristics — it is structure-preserving (in perception), and its application to a polysyllabic compound is dependent on syntactic bracketing; but the Half-Third Sandhi is characteristic of a postlexical rule — it is allophonic and applies across the board. The syntactic dependency of the Third-Tone Sandhi is illustrated in the examples in (3) and (4). Examples in (3a) and (3b) show that for underlying Third-Tone sequences, the output tones differ depending on the syntactic branching structure: a right-branching sequence [213 [213 213]] renders two possible output forms [35 35 213] and [21 35 213], as in (3a), while a left-branching sequence [[213 213] 213] has only one output [35 35 213], as in (3b) (from Duanmu 2000: p.238). Examples (3c) and (3d), on the other hand, illustrate that the application of the Half-Third Sandhi is not influenced by the branching

---

[6] A detailed acoustic study by Peng (2000) has shown that this sandhi is non-neutralizing production-wise — the sandhi tone is lower in overall pitch than the lexical Tone 2. But this difference cannot be reliably perceived by native adult listeners. Therefore, /mai213 ma213/ ‘buy horse’ and /mai35 ma213/ ‘bury horse’ are in effect homophonous to native speakers.

structure: an underlying /213-35-213/ sequence, regardless of branching structure, is realized as [21-35-213].

(3) Left- vs. right-branching phrases:

a. [213 [213 213]] → 35-35-213 or 21-35-213:

[mai	[xɑu	tɕjou]]	
buy	good	wine	‘buy good wine’
213	213	213	Input
35	35	213	Output 1
21	35	213	Output 2

b. [[213 213] 213] → 35-35-213 only:

[[mai	xɑu]	tɕjou]	
buy	done	wine	‘finished buying wine’
213	213	213	Input
35	35	213	Output
*21	35	213	

c. [213 [35 213]] → 21-35-213:

[ɕjɑu	[xuŋ	ma]]	
little	red	horse	‘little red horse’
213	35	213	Input
21	35	213	Output

d. [[213 35] 213] → 21-35-213:

[[ɕjɑu	xuŋ]	pʰɑu]	
Xiao	hong	run	‘Xiaohong runs’
213	35	213	Input
21	35	213	Output

Examples in (4a) and (4b) show that there is a special status for prepositions in permitting the non-application of the Third-Tone Sandhi on them. Example (4a) illustrates that in a [213 [[213 213] 213]] sequence, if the second syllable is a preposition such as [wɑŋ] ‘to,’ there are three possible outputs: [35-35-35-213], [21-35-35-213], or [35-21-35-213]. But if the second syllable is not a preposition, as in (4b), there are only two possible outputs [35-35-35-213], [21-35-35-213]; *[35-21-35-213], where the Third-Tone Sandhi is blocked on the second syllable, is not a possible output (from Zhang 1997: 294-295). Examples (4c) and (4d), in contrast, illustrate that a [55 [[213 51] 213]] sequence, regardless of whether the second syllable is a preposition, is realized as [55-21-51-213], demonstrating again the irrelevance of grammatical structure to the application of the Half-Third Sandhi.[7]

(4) The special status of prepositions:

a. [213 [[213_{prep} 213] 213]] → 35-35-35-213, 21-35-35-213, or 35-21-35-213:

---

[7] For more discussion on the application of the Mandarin Third-Tone Sandhi, see Shih (1997), Zhang (1997), Duanmu (2000), and Lin (2007).

[ma	[[**wɑŋ**	pei]	tsou]]	
horse	to	north	walk	‘The horse walks to the north.’
213	213	213	213	Input
35	35	35	213	Output 1
21	35	35	213	Output 2
35	21	35	213	Output 3

b. [213 [[$213_{\text{non-prep}}$ 213] 213]] → 35-35-35-213 or 21-35-35-213:

[ma	[[**xən**	ʂɑu]	xou]]	
horse	very	rarely	roar	‘Horses rarely roar.’
213	213	213	213	Input
35	35	35	213	Output 1
21	35	35	213	Output 2
*35	21	35	213	

c. [55 [[213_{prep} 51] 213]] → 55-21-51-213:

[t^ha	[[**wɑŋ**	xou]	tsou]]	
he	to	back	walk	‘He walks backwards.’
55	213	51	213	Input
55	21	51	213	Output

d. [55 [[$213_{\text{non-prep}}$ 51] 213]] → 55-21-51-213:

[t^ha	[[**xən**	ai]	kou]]

he	very	love	dog	‘He loves dogs very much.’
55	213	51	213	Input
55	21	51	213	Output

We should recognize that the Third-Tone Sandhi is not truly lexical: it clearly applies across word boundaries ((3) and (4)), its application in long strings is affected by speech rate ((3) and (4)), and it is not structure preserving in production under careful acoustic scrutiny (fn. 6). What is uncontroversial, however, is the clear *difference* between the two sandhis in that the Third-Tone Sandhi exhibits certain lexical characters, while the Half-Third Sandhi does not. The close relation between the postlexical status of a phonological rule and its phonetic motivation has been well established in the Lexical Phonology literature (e.g., Kiparsky 1982, 1985, Mohanan 1982), and we take it as another piece of evidence that the Half-Third Sandhi has a stronger phonetic motivation than the Third-Tone Sandhi.

The third reason is that the Third-Tone Sandhi corresponds to a historical sandhi pattern in Chinese, namely, *shang* → *yang ping* / __ *shang*, where *shang* and *yang ping* refer to the historical tonal categories from which 213 and 35 descended, respectively. This historical sandhi pattern dates back to at least the 16th century (Mei 1977). And according to Mei’s reconstruction, the pitch values for *shang* and *yang ping* in 16th century Mandarin were low level (22) and low rising (13), respectively. The present-day rendition of the sandhi in Mandarin is the result of historical tone changes that morphed *shang* into low-falling-rising and *yang ping* into high rising. Therefore, the Mandarin Third-Tone Sandhi was not originally motivated by the phonetic rationale of avoiding complex pitch contour on short duration. The same point is made by the variable synchronic renditions of the same historical sandhi in related Mandarin dialects (Court 1985). For instance, in Tianjin, it is 13 → 45 / __ 13 (Yang et al. 1999); in Jinan, 55 → 42 / __ 55 (Qian and Zhu 1998); in Taiyuan, 53 → 11 / __ 53 (Wen and Shen 1999). The Half-Third Sandhi, on the other hand, does not have a similar historical origin; and due to the different tonal shapes of the historical *shang* tone in different present-day dialects, it does not have comparable synchronic renditions.

The differences between the Third-Tone Sandhi and the Half-Third Sandhi in their phonetic characteristics, morphosyntactic properties, and historical origins all point to the possibility that the Half-Third Sandhi has a stronger synchronic phonetic basis than the Third-Tone Sandhi. It is important to note that we have not committed to an absolute cut-off point for what is phonetically based and what is not — to identify patterns that are useful for testing the synchronic relevance of phonetics, such a threshold is not necessary, nor do we believe that it exists. However, it is crucial to be able to make *comparisons* between patterns along the lines that we have considered for Mandarin in order to identify the relevant ones for the test.

Given the difference in phonetic grounding between the two sandhi patterns, the general question we pursue is whether Mandarin speakers exhibit different behaviors on the two sandhis in a wug test. Specifically, we test whether there is a difference in the productivity between the two sandhis. In line with the synchronic approach, we hypothesize that the sandhi with the stronger phonetic motivation — the Half-Third Sandhi — will apply more productively in wug words than the Third-Tone Sandhi. This greater productivity may be reflected in two ways. First, the Half-Third Sandhi may apply to a greater percentage of the wug tokens than the Third-Tone Sandhi. Second, there is no difference in the rate of application, but there is incomplete application for the Third-Tone Sandhi in wug words as compared to real words, while the Half-Third Sandhi applies to the wug words the same way as it applies to the real words. In light of the discussion on the difference between Tone 2 and Tone 3 earlier, we specifically expect a lower and later turning point and a longer duration for the sandhi tone in wug words than in real words for the Third-Tone Sandhi in this scenario.[8]

Finally, we must acknowledge that the Third-Tone Sandhi and the Half-Third Sandhi do not have the same lexical frequency in Mandarin. Calculations based on a syllable frequency corpus by Da (2004), which contains 192,647,157 syllables, indicate that the numbers of legal syllable types with Tones 1 through 4 are 258, 224, 254, and 318 respectively, and syllables with Tones 1 through 4 account for 16.7%, 18.4%, 14.8%, and 42.5% of all syllables in the corpus.[9] In other words, Tone 3 has the third-lowest type frequency and the lowest token frequency, which means that 3+3 disyllabic words may have relatively low frequency. Moreover, the Third-Tone Sandhi also has a limited environment as compared to the Half-Third Sandhi. The environments in which the Half-

---

[8] We also hypothesized that the reaction times for the two sandhis will be significantly different due to the potentially different types of processing for the two sandhis. But it was difficult to decouple the allophonic differences in rhyme duration among different tones from differences in reaction time, and this hypothesis was not borne out in the two experiments that we conducted. We do not report the reaction time results here. Interested readers can consult a previous version of this article in which reaction time results are reported, downloadable from the first author’s website at http://www2.ku.edu/~ling/faculty/zhang.shtml.

[9] The other 7.6% are syllables with a neutral tone.

Third Sandhi applies, which includes __55, __35, and __51, account for 75.9% of all sandhi environments by type frequency counts. Therefore, in our study, we need to pay special attention to whether the Half-Third Sandhi behaves as a unified process before 55, 35, and 51. If so, it will present a challenge to the goal of the study as any effect that conforms to our hypothesis may be due to the considerably higher lexical frequency of the Half-Third Sandhi. If not, the frequency profile of the four tones in Mandarin will provide us with an opportunity to study the potential effect of lexical frequency on sandhi productivity and its interaction with the effect of phonetics. If lexical frequency influences sandhi productivity, we primarily expect a type frequency effect (Baayen 1992, 1993, Bybee 1985, 2001, Ernestus and Baayen 2003, Pierrehumbert 2003, 2006, among others), and thus a low productivity of the Half-Third Sandhi for the 3+2 sequences. But we also cannot exclude the possible effect of token frequency, which has been shown to be relevant to the productivity of Taiwanese tone sandhi in Zhang and Lai (2008) and Zhang et al. (2009a, b). If the frequency effect is mainly based on token frequency, we would expect 3+3 to have the lowest productivity.

3. Experiment 1

3.1. Methods

3.1.1. Stimuli

Following Hsieh (1970, 1975, 1976)'s experimental design for a Taiwanese wug test, we constructed five sets of disyllabic test words in Mandarin. The first set includes real words, denoted by AO-AO (AO = actual occurring morpheme). This set serves as the control for the experiment and is the set with which results of wug words are compared. The other four sets are wug words: *AO-AO, where both syllables are actual occurring morphemes, but the disyllable is non-occurring; AO-AG (AG = accidental gap), where the first syllable is actual occurring, but the second syllable is an accidental gap in Mandarin syllabary; AG-AO, where the first syllable is an accidental gap and the second syllable is actual occurring; and AG-AG, where both syllables are accidental gaps. The AGs were selected by the authors, who are both native speakers of Mandarin Chinese. In each AG, both the segmental composition and the tone of the syllable are legal in Mandarin, but the combination happens to be missing. For example, [pʰan] is a legal syllable, and it occurs with tones 55 ‘to climb,’ 35 ‘plate,’ and 51 ‘to await eagerly;’ but it accidentally does not occur with the Third Tone 213. Therefore, [pʰan213] is a possible AG.

For each set of words, we used four different tonal combinations: the first syllable always has the Third Tone 213, and the second syllable has one of the four tones in Mandarin. Therefore, each tonal combination is in the environment to undergo either the Third-Tone or the Half-Third Sandhi. Eight words for each tonal combination were used, making a total of 160 test words (8×4×5).

The AO-AO words were all high frequency words selected from the Feng Hua Yuan character and digram frequency corpus of Da (1998). For the four wug sets, the digram frequencies are all zero, and we used the same first syllable to combine with the four different tones in the second syllable, e.g., in AG-AG, we used [**pʰiəŋ213** ʂwən55], [**pʰiəŋ213** tʰɤ35], [**pʰiəŋ213** tsɤŋ213], and [**pʰiəŋ213** tʂa51], along with seven other such sets. In the recorded stimuli, the same token was used to combine with different second syllables. The identity of the first syllable allows for the comparison between the two types of sandhi that the first syllable may undergo.

To avoid neighborhood effects in wug words to some extent, we ensured that any disyllabic wug word is not a real word with *any* tonal combinations, not just the tonal combination used for the disyllable. We specifically controlled for the tonal neighbors because research on homophony judgment (Taft and Chen 1992, Cutler and Chen 1997), phoneme (toneme) monitoring (Ye and Connine 1999), and legal-phonotactic judgment (Myers 2002) has all shown that phonemic tonal differences are perceptually less salient than segmental differences, which entails that tonal neighbors, in a sense, make closer neighbors.

Finally, to disguise the purpose of the experiment, we also used 160 disyllabic filler words. All filler syllables were real syllables in Mandarin; half of disyllabic fillers were real words, and the other half were wug words.

All test stimuli and fillers were read by the first author, who is a native speaker of Mandarin that grew up in Beijing. The Third-Toned syllables were all read with full Third Tones. The entire set of test stimuli as well as additional information about the test stimuli and fillers are given in the Appendix.

3.1.2. Experimental set-up

The experiment was conducted with SuperLab (Cedrus) in the Phonetics and Psycholinguistics Laboratory at the University of Kansas. There were 320 stimuli in total (160 test items + 160 fillers). Each stimulus consisted of two monosyllabic utterances separated by an 800ms interval. The stimuli were played through a headphone worn by the subjects. For each stimulus, the subjects were asked to put the two syllables together and pronounce them as a disyllabic word in Mandarin. Their response was collected by a Sony PCM-M1 DAT recorder through a 33-3018 Optimus dynamic microphone placed on the desk in front of them. The sampling rate for the DAT recorder was 44.1kHz. The digital recording was then down-sampled to 22kHz onto a PC hard-drive using Praat (Boersma and Weenink 2003). There was a 2000ms interval between stimuli. If the subject did not respond within 2000ms after the second syllable played, the next stimulus would begin. The stimuli were divided into two same-sized blocks (A and B) with matched stimulus types, and there was a five-minute break between the blocks. Half of the subjects took block A first, and the other half took block B first. Within each block, the stimuli were automatically randomized by SuperLab. Before the experiment began, there was a short introduction in Chinese that the subjects heard through the headphone and simultaneously read on a computer screen in front of them, which explained their task both in prose and through examples. There was then a practice session of 14 words (two of each of AO-AO, *AO-AO, AO-AG, AG-AO, and AG-AG, two real-word fillers, two wug fillers). The experiment began after a verbal confirmation from the subjects that they were ready. The entire experiment took around 45 minutes.

3.1.3. Subjects

Twenty native speakers of Mandarin (12 male, 8 female) recruited on KU campus participated in the study. All speakers were from northern areas of Mainland China and spoke Standard Mandarin natively without any noticeable accent as judged by the authors. Except for one speaker who was 45 years old and had been in the US for 20 years, all speakers ranged from 23 to 35 years in age and had been in the US for less than four years at the time of the experiment. Each subject was paid a nominal fee for participating in the study.

3.1.4. Data analyses

All test tokens from the subjects were listened to by the two authors. A token was not used in the analysis if there was a large enough gap between the two syllables that they clearly did not form a disyllabic word. For the rest of the tokens, it was judged that both the Third-Tone Sandhi and the Half-Third Sandhi applied 100% of the time. Non-applications of the sandhis should be easy to detect for native speakers, as they involve clear phonotactic violations (*213 nonfinally). Therefore, the test for the productivity of the sandhis lies in the accuracy of their applications to the wug words. To investigate the accuracy of sandhi application, we extracted the f_0 of the rhyme in the first syllable of the subjects' disyllabic response using Praat. We then took a f_0 measurement every 10% of the duration of the rhyme, giving eleven f_0 measurements for each rhyme. For each tonal combination (3+1, 3+2, 3+3, 3+4), we did two comparisons. The first is between AO-AO and the rest of the word groups *AO-AO, AO-AG, AG-AO, and AG-AG; i.e., real disyllables vs. wug disyllables. The other is between AO-AO, *AO-AO, AO-AG and AG-AO, AG-AG; i.e., real σ_1s vs. wug σ_1s. The rationale for the two comparisons is that lexical listing could be at the disyllabic word or monosyllabic morpheme level, and doing both comparisons will allow us to tease apart the two possibilities. Our hypothesis for these comparisons is that the difference in sandhi tones between real words and wugs should be greater for cases of Third-Tone Sandhi than Half-Third Sandhi due to the stronger phonetic motivation for the latter. In particular, we expect incomplete application of the Third-Tone Sandhi in wugs, i.e., Tone 3 in σ_1 will resist the change to Tone 2. Again, given the acoustic characteristics of Tone 2 and Tone 3 in Mandarin, the hypothesis translates into a lower and later turning point and a longer duration for the sandhi tone in wug words than in real words.

Among the twenty speakers, there were two speakers (one male and one female) whose f_0 values could not be reliably measured by Praat due to high degrees of creakiness in their voice. We discarded these speakers' data in the f_0 analysis.

Figure 3 illustrates how we compared two f_0 curves. We conducted a two-way Huynh-Feldt Repeated Measures ANOVA, which corrected for sphericity violations, with Word-Group and Point as independent variables. The Word-Group variable has two levels — Word-Group 1 and Word-Group 2, and a significant main effect would indicate that the two f_0 curves representing the two word groups have different average pitches. The Point variable has eleven levels, representing the eleven points where f_0 data are taken. A significant interaction between Word-Group and Point would indicate that the two curves have different shapes. This method of comparing two f_0 curves has been used by Peng (2000).

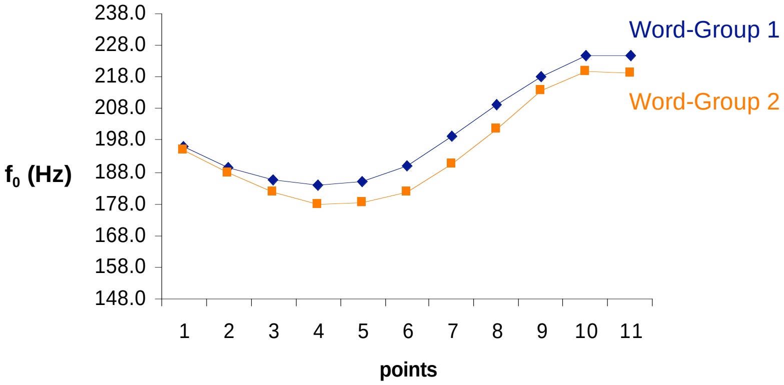


Figure 3. Comparing two f_0 curves.

For σ_1 in 3+3 combinations, we also measured the f_0 drop and the duration from the beginning of the rhyme to the pitch turning point, as shown in Figure 4. Comparisons between real and wug disyllables and between real and wug σ_1s on these measurements were made using one-way Huynh-Feldt Repeated-Measures ANOVAs. We expected the f_0 drop to be greater and the TP duration to be longer for wug words than real words.

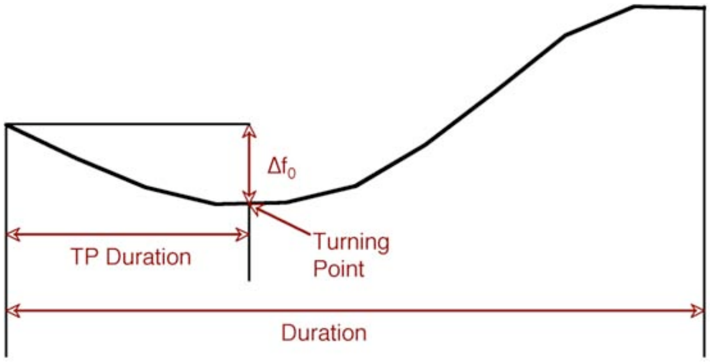


Figure 4. A schematic of the measurements taken from the pitch curve of the rhyme in σ_1 in 3+3 combinations. "Δf_0" and "TP Duration" are the pitch drop and duration from the beginning of the rhyme to the turning point, respectively. "Duration" is the entire rhyme duration.

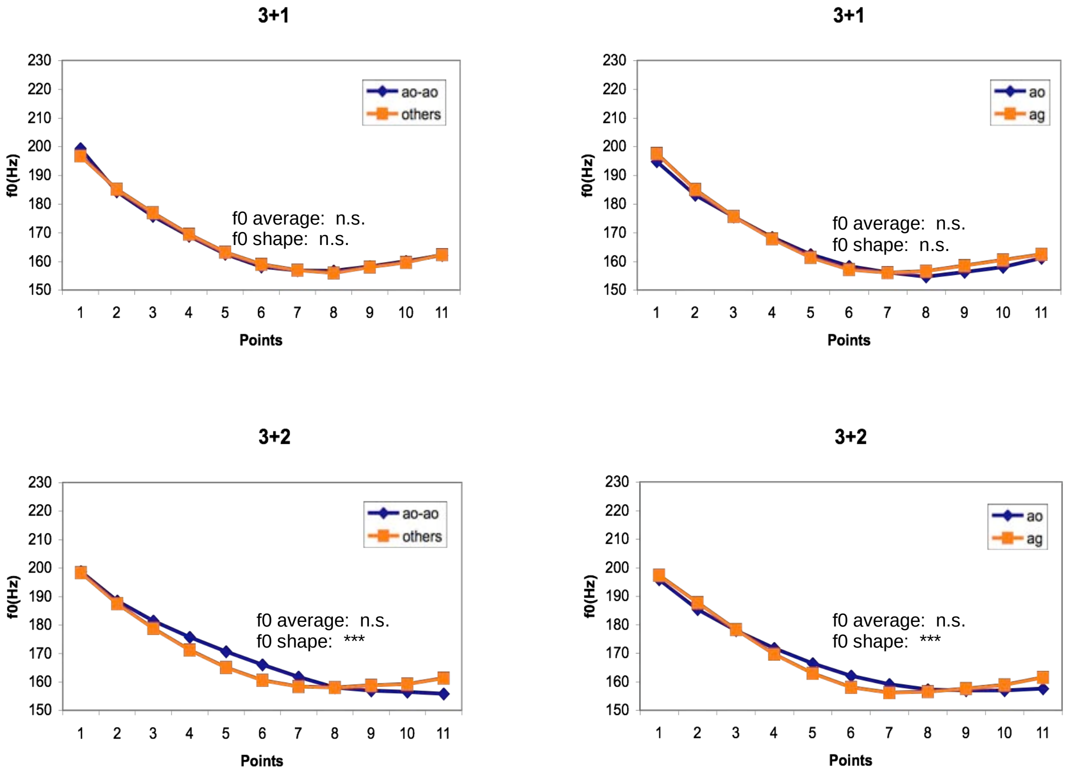
Finally, we measured the σ_1 rhyme duration for all the disyllabic combinations and compared real and wug disyllables and real and wug σ_1s for each tonal combination using one-way Huynh-Feldt Repeated-Measures ANOVAs. Based on the synchronic approach, we expected to find a longer rhyme duration for the wug words in 3+3 combination, but no difference between wug and real words in other combinations.

3.2. Results

3.2.1. f_0 contour

In this section, we report the results of comparison on the f_0 of the first syllable of the subjects’ response between real disyllables and wug disyllables and between real-σ_1 words and wug-σ_1 words.

The results from the Half-Third Sandhi comparisons are given in Figure 5. In this and all following figures, “n.s.” indicates no significant difference, “*”, “**”, and “***” indicate significant differences at $p<0.05$, $p<0.01$, and $p<0.001$ levels, respectively.


(a) AO-AO vs. others

(b) AO vs. AG in σ_1

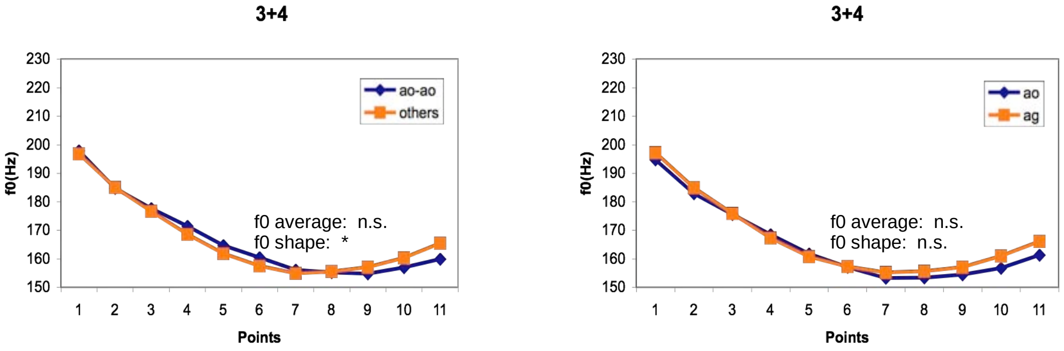


Figure 5. f_0 curves of the first syllable for the Half-Third Sandhi. The three graphs in (a) represent the real-disyllable vs. wug-disyllable comparisons for the first syllable in 3+1, 3+2, and 3+4. The three graphs in (b) represent the real-σ_1 vs. wug-σ_1 comparisons for the same tonal combinations.

As we can see in Figure 5, for both Tone 1 and Tone 4, the subjects' performance of the Half-Third Sandhi on wug words is generally identical to that on real words in terms of both the average f_0 and the f_0 contour shape. This is true for both the disyllabic and σ_1 comparisons for Tone 1 and the σ_1 comparisons for Tone 4. When σ_2 has Tone 2, the f_0 contour on σ_1 has a significantly different shape between real and wug words for both comparisons. The statistical results for these comparisons are given in Table 1.

Real-disyllable vs. wug-disyllable			
	Tone 1	Tone 2	Tone 4
Wd-Gr (f_0 average)	F(1.000, 17.000) =0.005, p=0.945	F(1.000, 17.000) =0.805, p=0.382	F(1.000, 17.000) =0.000, p=1.000
Point	F(3.187, 54.180) =125.614, $p<0.001$	F(2.119, 36.023) =168.840, $p<0.001$	F(2.663, 45.263) =133.073, $p<0.001$
Wd-Gr × Point (f_0 shape)	F(3.574, 60.750) =0.880, p=0.472	F(2.824, 48.012) =13.036, $p<0.001$	F(3.436, 58.409) =3.535, p=0.016

Real-σ_1 vs. wug-σ_1			
	Tone 1	Tone 2	Tone 4
Wd-Gr (f_0 average)	F(1.000, 17.000) =0.061, p=0.808	F(1.000, 17.000) =0.000, p=0.997	F(1.000, 17.000) =0.189, p=0.670
Point	F(3.275, 55.680) =167.524, p<0.001	F(2.143, 36.439) =178.423, p<0.001	F(2.651, 45.059) =117.356, p<0.001
Wd-Gr × Point (f_0 shape)	F(2.545, 43.265) =2.178, p=0.113	F(3.150, 53.546) =9.072, p<0.001	F(2.942, 50.011) =2.265, p=0.093

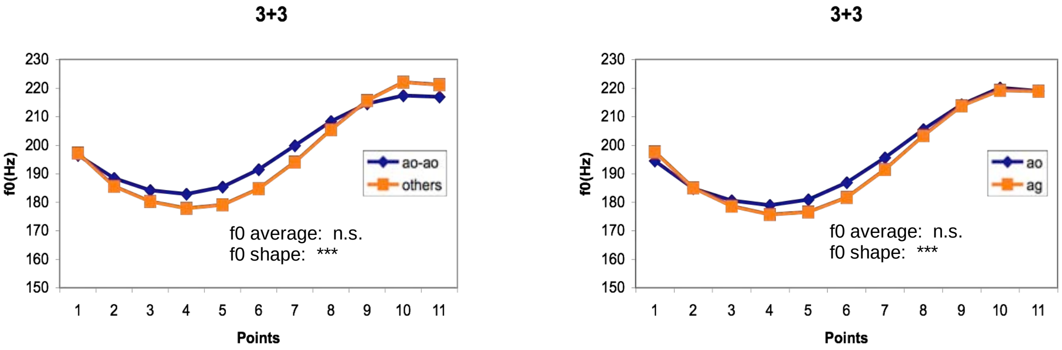
Table 1. Two-way Huynh-Feldt Repeated-Measures ANOVA results for the first syllable f_0 curves in the Half-Third Sandhi.

From Figure 5, we can also see that the f_0 shape difference between real and wug words for 3+2 lies in the fact that the f_0 shape for the wug words has a turning point at around 70% into the tone, while the f_0 shape for the real words is monotonically falling throughout the rhyme. This indicates that there may be incomplete application of the Half-Third Sandhi in 3+2, and hence its lower accuracy/productivity in this particular environment.[10] We currently have no account for why there is a significant f_0 shape difference between AO-AO and other word groups for 3+4.

The results from the Third-Tone Sandhi comparisons are given in Figure 6. Two-way Huynh-Feldt Repeated-Measures ANOVAs indicate that although the average f_0 is the same for both comparisons, the f_0 contour shape is significantly different between the real words and wug words for both comparisons. The ANOVA results are summarized in Table 2.

---

[10] The pitch rise at the end of the first syllable in 3+1 and 3+4 for real disyllable and real σ_1 words is likely due to coarticulation with the high pitch onset of the following tone (Tone 1 = 55, Tone 4 = 51).



(a) AO-AO vs. others

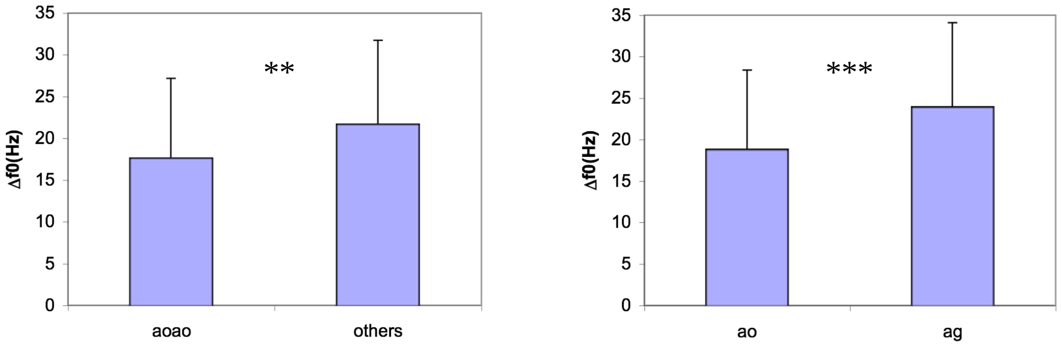
(b) AO vs. AG in σ_1

Figure 6. f_0 curves of the first syllable for the Third-Tone Sandhi. Graphs (a) and (b) represent the real-disyllable vs. wug-disyllable and real-σ_1 vs. wug-σ_1 comparisons, respectively.

Real-disyllable vs. wug-disyllable	
	Tone 3
Wd-Gr (f_0 average)	F(1.000, 17.000)=1.351, p=0.261
Point	F(2.371, 40.312)=73.135, p<0.001
Wd-Gr × Point (f_0 shape)	F(2.414, 41.031)=9.537, p<0.001
Real-σ_1 vs. wug-σ_1	
	Tone 3
Wd-Gr (f_0 average)	F(1.000, 17.000)=0.000, p=0.997
Point	F(2.143, 36.439)=178.423, p<0.001
Wd-Gr × Point (f_0 shape)	F(3.150, 53.546)=9.072, p<0.001

Table 2. Two-way Huynh-Feldt Repeated-Measures ANOVA results for the first syllable f_0 curves in the Third-Tone Sandhi.

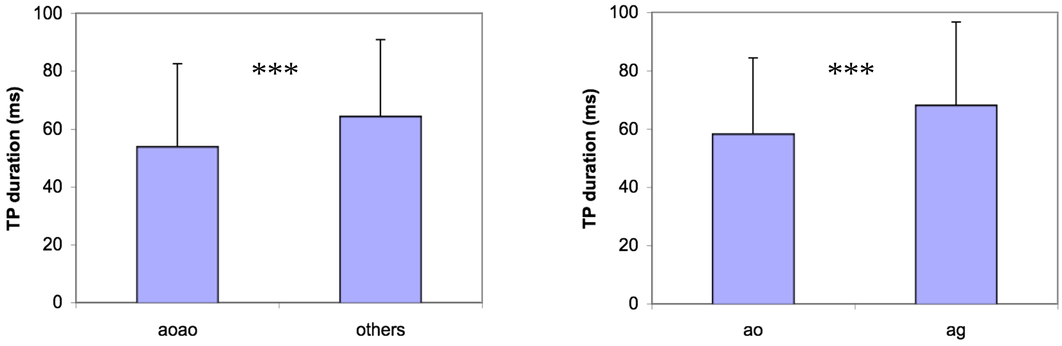
From Figure 6, we can also see that for the curves representing wug words ("others" in the first graph, "ag" in the second graph), the turning points are both lower and later than their counterparts for the curves representing real words, indicating that there may be incomplete application of the sandhi. To quantify these turning point differences in σ_1 of the 3+3 combination, we defined Δf_0 as the difference between the f_0 of the beginning of the rhyme and the f_0 turning point in the rhyme and TP duration as the duration from the beginning of the rhyme to the turning point. Results of comparisons between real and wug disyllables and between real and wug σ_1s on Δf_0 and TP duration for 3+3 are given in Figure 7 and Figure 8, respectively. In these and following figures, error bars indicate one standard deviation. One-way Repeated-Measures Huynh-Feldt ANOVAs with Word-Group as the independent factor indicate that for Δf_0, AO-AO is significantly different from other word groups ($F(1.000, 17.000)=8.543$, $p<0.01$), so is σ_1=AO from σ_1=AG ($F(1.000, 17.000)=48.254$, $p<0.001$); for TP duration, AO-AO is significantly different from other word groups ($F(1.000, 17.000)=19.561$, $p<0.001$), so is σ_1=AO from σ_1=AG ($F(1.000, 17.000)=21.343$, $p<0.001$). These results support our hypothesis: with a lower and later turning point, the sandhi tone on wug words is more similar to the original Tone 3 than that on real words, indicating incomplete application of the sandhi in wug words.



(a) AO-AO vs. others

(b) AO vs. AG in σ_1

Figure 7. Δf_0 results for 3+3. Graphs (a) and (b) represent the real-disyllable vs. wug-disyllable and real-σ_1 vs. wug-σ_1 comparisons, respectively.



(a) AO-AO vs. others

(b) AO vs. AG in σ_1

Figure 8. TP duration results for 3+3. Graphs (a) and (b) represent the real-disyllable vs. wug-disyllable and real-σ_1 vs. wug-σ_1 comparisons, respectively.

3.2.2. Rhyme duration

The results for σ_1 rhyme duration for all the tonal combinations are given in Figure 9, and the statistical results are summarized in Table 3. One-way Repeated-Measures Huynh-Feldt ANOVAs with Word-Group as the independent factor show that there are no significant differences between AO-AO and other word groups for any of the tonal combinations. But for 3+3, the difference approaches significance at $p<0.05$ ($F(1.000, 17.000)=4.218$, $p=0.056$), and the difference is in the expected direction, i.e., wug > real. For AO vs. AG, 3+3 is the only combination in which the wug words have a significantly longer σ_1 rhyme duration than the real words ($F(1.000, 17.000)=5.653$,

$p<0.05$). These results support our hypothesis: the durational property for the sandhi syllables is identical between real and wug words for the Half-Third Sandhi, but for the Third-Tone Sandhi, the sandhi syllable rhyme duration in wug words is longer than that in real words, indicating again incomplete application of the sandhi in wug words. These results are consistent with an approach that encodes phonetic biases in the grammar, but not with a frequency-only approach, as the latter expects a greater durational difference between real and wug words for 3+2 than 3+3 due to the former's lower lexical frequency.

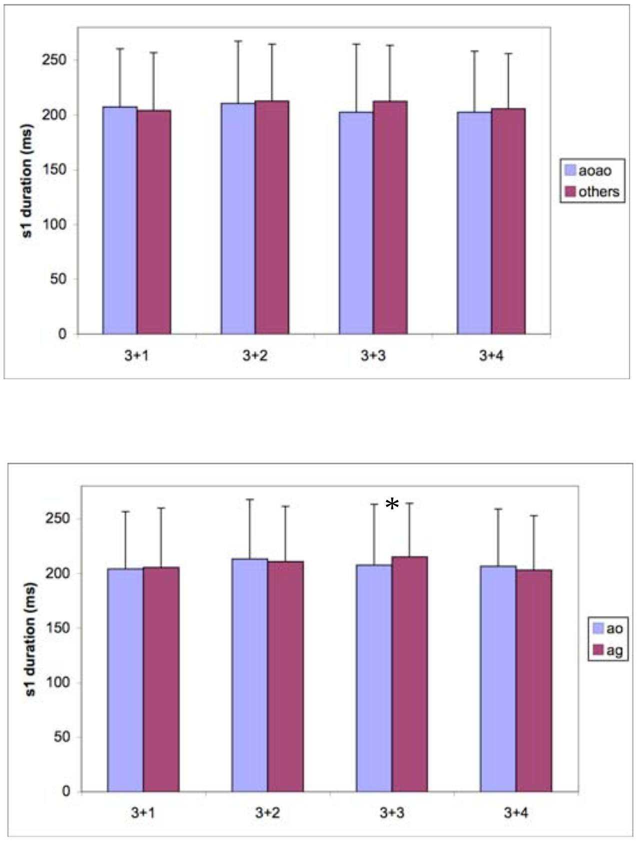


Figure 9. Rhyme duration of σ_1 for all tonal combinations. The two graphs represent the real-disyllable vs. wug-disyllable and real-σ_1 vs. wug-σ_1 comparisons, respectively.

Real-disyllable vs. wug-disyllable	
Tone 3 + Tone 1	F(1.000, 17.000)=0.660, p=0.428
Tone 3 + Tone 2	F(1.000, 17.000)=0.206, p=0.656
Tone 3 + Tone 3	**F(1.000, 17.000)=4.218, p=0.056**
Tone 3 + Tone 4	F(1.000, 17.000)=0.620, p=0.442
Real-σ_1 vs. wug-σ_1	
Tone 3 + Tone 1	F(1.000, 17.000)=0.097, p=0.759
Tone 3 + Tone 2	F(1.000, 17.000)=0.559, p=0.465
Tone 3 + Tone 3	**F(1.000, 17.000)=5.653, p=0.029**
Tone 3 + Tone 4	F(1.000, 17.000)=1.118, p=0.305

Table 3. One-way Huynh-Feldt Repeated-Measures ANOVA results for the σ_1 rhyme duration in all tonal combinations.

3.3. Discussion

Our results of the Third-Tone Sandhi indicate a significant difference between real words and wug words in the contour shape of the sandhi tone; in particular, the contour shape of the sandhi tone in wug words shares a greater similarity with the original Tone 3 by having a lower and later turning point and a longer tone duration. Given that we did not judge any 3+3 tokens in the data to have non-application of the Third-Tone Sandhi, the difference between real and wug words for the Third-Tone Sandhi was not due to the non-application of the sandhi to a limited number of tokens/speakers, but the *incomplete* application of the sandhi to a large number of tokens. The real vs. wug comparison for the Half-Third Sandhi, however, showed identical contour shape of the sandhi tone for Tone 1, an inconsistent contour shape difference for Tone 4 — a difference at $p<0.05$ level ($p=0.016$) for the disyllabic comparison, but no difference for the AO vs. AG comparison, and a significant contour shape difference for

Tone 2 that indicates incomplete application of the sandhi. This illustrates two points: (a) The Half-Third Sandhi behaves differently in different environments, and (b) the sandhi with the lowest type frequency (3+2) also applies less consistently to wug words than to real words.

The real-disyllable vs. wug-disyllable and real-σ_1 vs. wug-σ_1 comparisons returned similar results. But the difference between the two sandhis is more apparent in the real-σ_1 vs. wug-σ_1 comparison, as indicated by the equal or more significant difference for the Third-Tone Sandhi and the equal or less significant difference for the Half-Third Sandhi between the two groups for all f_0 measures.

Therefore, our hypothesis that the difference in sandhi tones between real words and wugs should be greater for cases of Third-Tone Sandhi than Half-Third Sandhi finds support in that (a) the difference between real and wug words for the Third-Tone Sandhi can be translated into incomplete application for the sandhi in wug words, and (b) there is no consistent difference between real and wug words for the Half-Third Sandhi. We have also found an effect that is potentially due to type frequency: the Half-Third Sandhi in 3+2 also applies incompletely to wug words. The effects overall, however, are not consistent with a frequency-only account, as the differences between real and wug words are more consistent for 3+3 than 3+2, as evidenced by the lack of rhyme duration difference in 3+2.

These results must be interpreted cautiously, however, for two reasons. First, the differences between real and wug words in the Third-Tone Sandhi, although statistically highly significantly, are quite small in magnitude. It is thus important for us to be able to replicate these results in a separate experiment. Second, although all of our participants came from northern areas of Mainland China and spoke Standard Mandarin natively without any noticeable accent, they did have backgrounds in different Northern Chinese dialects. This could potentially have an effect on the results. Experiment 2 was designed and conducted to address these issues.

4. Experiment 2

The goals of Experiment 2 are two-fold: first, it serves as a replication of Experiment 1; second, it includes only participants who grew up in Beijing and minimizes the potential dialectal effects on the results.

4.1. Methods

The methods of Experiment 2 are identical to those of Experiment 1 except that the experiment was conducted in the Phonetics Laboratory of the Department of Chinese Language and Literature at Beijing (Peking) University in China and that the recordings were made by a Marantz solid state recorder PMD 671 using a EV N/D 767a microphone. The sampling rate of the solid state recorder was 44.1kHz, and the digital recording was not further down-sampled.[11]

---

[11] We manipulated the duration of the second syllable of the stimuli in the following way in Praat. We took the median rhyme duration of the 160 second syllables in the test stimuli (454ms), and either expanded or shrank the duration of the rhymes of all second syllables to the same duration. We then calculated the expansion or shrinkage ratio of each rhyme and applied the same ratio to the VOT, frication duration, or sonorant duration of its onset consonant. The duration of the fillers remained unchanged. This duration manipulation was conducted in order to minimize the

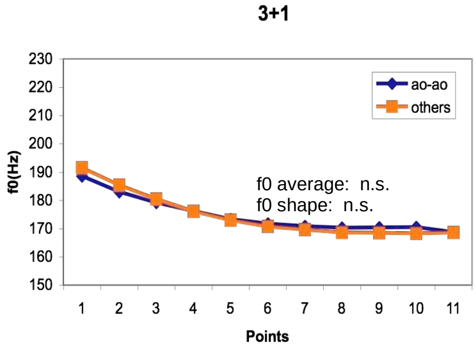
Thirty-one native speakers of Beijing Chinese (9 male, 22 female) recruited on Beijing University campus participated in the experiment. All subjects grew up and went through their primary and secondary schooling in Beijing, and none reported being conversant in any other dialects of Chinese. The subjects ranged from 19 to 37 years in age. Each subject was paid a nominal fee for participating in the study. Due to technical problems with Superlab, we were not able to use one male speaker's data. Therefore, we report data from 30 speakers.

4.2. Results

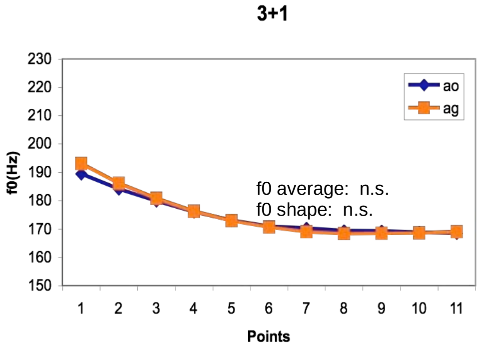
4.2.1. f_0 contour

The f_0 contour results for the Half-Third Sandhi comparisons are given in Figure 10.

(a) AO-AO vs. others



(b) AO vs. AG in σ_1



---

allophonic durational differences among different tones so that the reaction time hypothesis can be better tested (fn. 8).

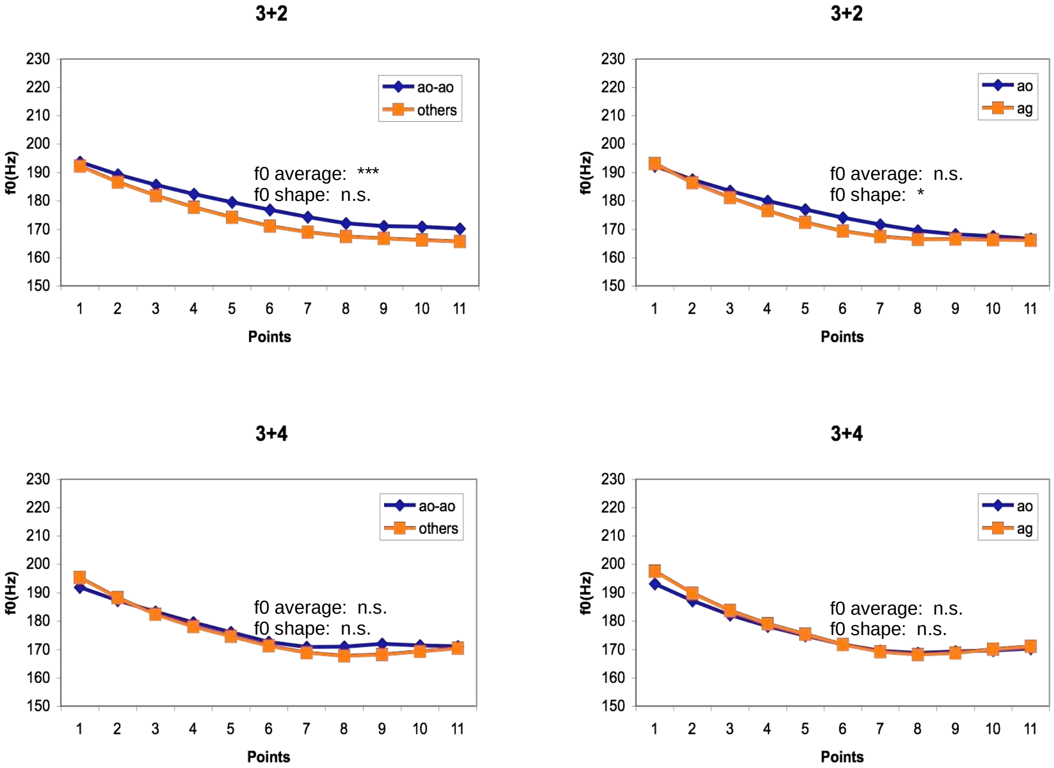


Figure 10. f_0 curves of the first syllable for the Half-Third Sandhi.

For both Tone 1 and Tone 4, the subjects' performance of the Half-Third Sandhi on wug words is generally identical to that on real words in terms of both the average f_0 and the f_0 contour shape. This is true for both the disyllabic and σ_1 comparisons for Tone 1 and the σ_1 comparisons for Tone 4. For the disyllabic comparison for Tone 1, however, the *p* value is right at 0.05 and needs to be acknowledged. When σ_2 has Tone 2, the average f_0 pitch on σ_1 is significantly lower for wug words than real words for the disyllabic comparison, and the f_0 shape between real and wug words is significantly different for the AO vs. AG comparisons. The statistical results for these comparisons are given in Table 4.

Real-disyllable vs. wug-disyllable			
	Tone 1	Tone 2	Tone 4
Wd-Gr (f_0 average)	F(1.000, 29.000) =0.024, p=0.878	F(1.000, 29.000) =19.561, p<0.001	F(1.000, 29.000) =0.616, p=0.439
Point	F(1.773, 51.431) =68.996, p<0.001	F(1.460, 42.348) =128.525, p<0.001	F(1.855, 53.807) =121.127, p<0.001
Wd-Gr × Point (f_0 shape)	F(2.466, 71.504) =2.905, p=0.050	F(1.930, 55.958) =2.581, p=0.087	F(1.387, 40.243) =2.506, p=0.111
Real-σ_1 vs. wug-σ_1			
	Tone 1	Tone 2	Tone 4
Wd-Gr (f_0 average)	F(1.000, 29.000) =0.110, p=0.743	F(1.000, 29.000) =3.007, p=0.094	F(1.000, 29.000) =0.745, p=0.395
Point	F(1.597, 46.324) =81.352, p<0.001	F(1.618, 46.918) =119.066, p<0.001	F(1.683, 48.807) =118.023, p<0.001
Wd-Gr × Point (f_0 shape)	F(1.454, 42.165) =1.655, p=0.207	F(2.319, 67.242) =4.646, p=0.010	F(1.804, 52.323) =1.954, p=0.156

Table 4. Two-way Huynh-Feldt Repeated-Measures ANOVA results for the first syllable f_0 curves in the Half-Third Sandhi.

The f_0 shape difference between real-σ_1 and wug-σ_1 words for 3+2 lies in the fact that the f_0 shape for the wug words has a turning point at around 80% into the tone, while the f_0 shape for the real words is monotonically falling throughout the rhyme. This is similar to the f_0 shape difference in both real vs. wug comparisons in Experiment 1. It again indicates that there may be incomplete application, and hence lower accuracy/productivity of the Half-Third Sandhi in 3+2.

The results from the Third-Tone Sandhi comparisons are given in Figure 11. Two-way Huynh-Feldt Repeated-Measures ANOVA's indicate that both the average f_0 and the f_0 contour shape are significantly different between the real words and wug words for both comparisons. The ANOVA results are summarized in Table 5.

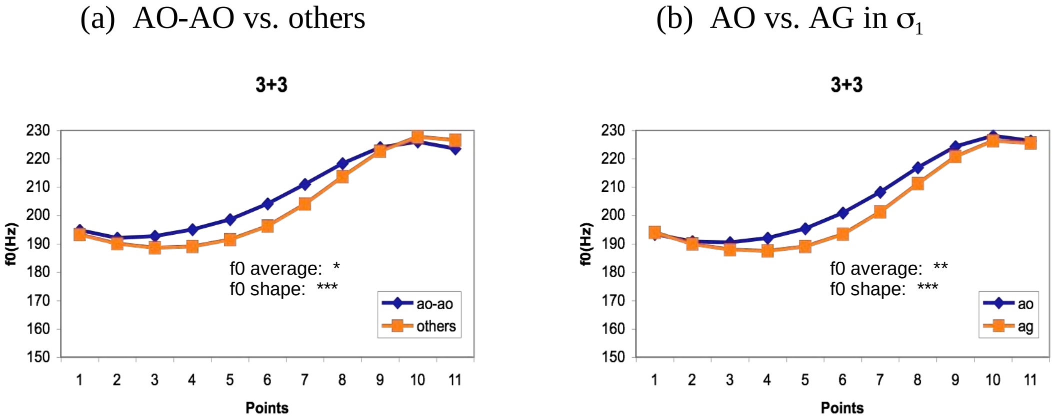

(a) AO-AO vs. others (b) AO vs. AG in σ_1

Figure 11. f_0 curves of the first syllable for the Third-Tone Sandhi.

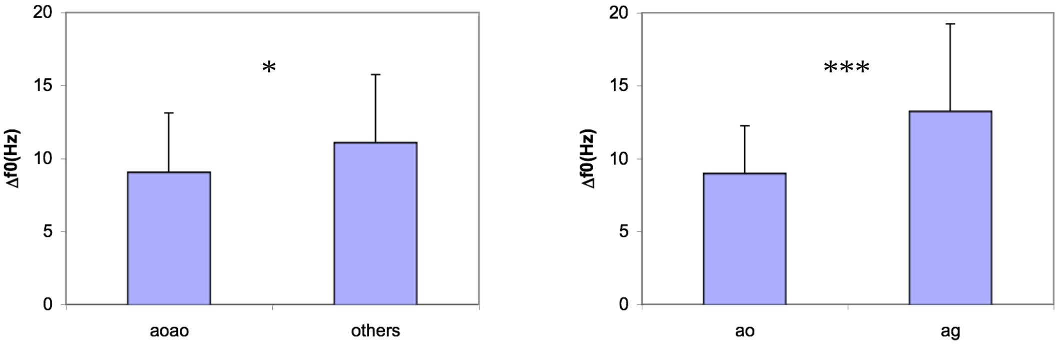
Real-disyllable vs. wug-disyllable	
	Tone 3
Wd-Gr (f_0 average)	F(1.000, 29.000)=4.946, p=0.034
Point	F(1.643, 47.654)=154.695, p<0.001
Wd-Gr × Point (f_0 shape)	F(2.161, 62.678)=12.291, p<0.001
Real σ_1 vs. wug σ_1	
	Tone 3
Wd-Gr (f_0 average)	F(1.000, 29.000)=11.153, p=0.002
Point	F(1.720, 49.893)=192.180, p<0.001
Wd-Gr × Point (f_0 shape)	F(2.319, 67.250)=18.352, p<0.001

Table 5. Two-way Huynh-Feldt Repeated-Measures ANOVA results for the first syllable f_0 curves in the Third-Tone Sandhi.

We have replicated our major finding regarding the f_0 contours in Experiment 1: the σ_1 in 3+3 sequences show consistent contour shape differences between the real and wug words in the two comparisons. This experiment also shows that there is an average pitch difference between real and wug words for 3+3. Moreover, other tonal sequences do not show differences between real words and wug words except for 3+2 — the tonal combination that has the lowest type frequency. However, 3+2 differences between real and wug words are less consistent than the 3+3 differences. This would not be consistent with a frequency-only account, but would be consistent with an account in which both phonetics and frequency are relevant.

From Figure 11, we can see that the contour shape difference between real and wug words for 3+3 is similar to that in Experiment 1: the turning points for wug words are both lower and later than their counterparts in real words, indicating that there may be incomplete application of the sandhi in the wug words.

The comparisons between real and wug disyllables and between real and wug σ_1s on Δf_0 for 3+3 are given in Figure 12. A one-way Repeated-Measures Huynh-Feldt ANOVA indicates that AO-AO has a significantly smaller Δf_0 than other word groups ($F(1.000, 29.000)=4.457$, $p<0.05$), so does σ_1=AO than σ_1=AG ($F(1.000, 29.000)=28.523$, $p<0.001$).

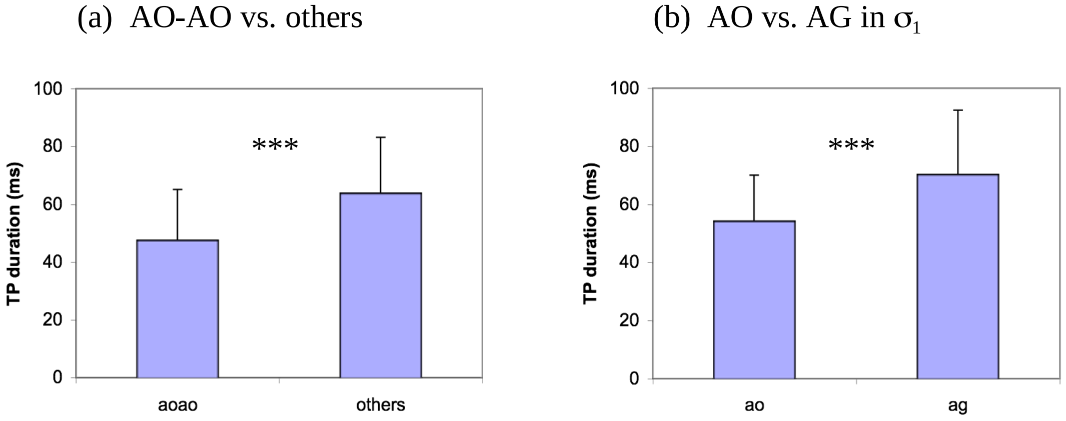


(a) AO-AO vs. others

(b) AO vs. AG in σ_1

Figure 12. Δf_0 results for 3+3.

Comparisons between real and wug words for TP duration of 3+3 are given in Figure 13. A one-way Repeated-Measures Huynh-Feldt ANOVA indicates that AO-AO has a significantly shorter TP duration than other word groups (F(1.000, 29.000)=28.793, $p<0.001$), so does σ_1=AO than σ_1=AG (F(1.000, 29.000)=56.235, $p<0.001$).


(a) AO-AO vs. others

(b) AO vs. AG in σ_1

Figure 13. TP duration results for 3+3.

Given that we will later see in §4.2.3 that the wug words generally have a longer σ_1 rhyme duration than real words, to ensure that the longer TP duration in wug words is not simply due to the longer σ_1 duration, we also calculated the percentages of the TP duration to the entire σ_1 rhyme duration and compared the real words with wug words. These comparisons are shown in Figure 14. ANOVA results show that percentage-wise, AO-AO's turning point is still significantly earlier than that of other word groups (F(1.000, 29.000)=5.082, $p<0.05$), so is σ_1=AO than σ_1=AG (F(1.000, 29.000)=34.617, $p<0.001$).

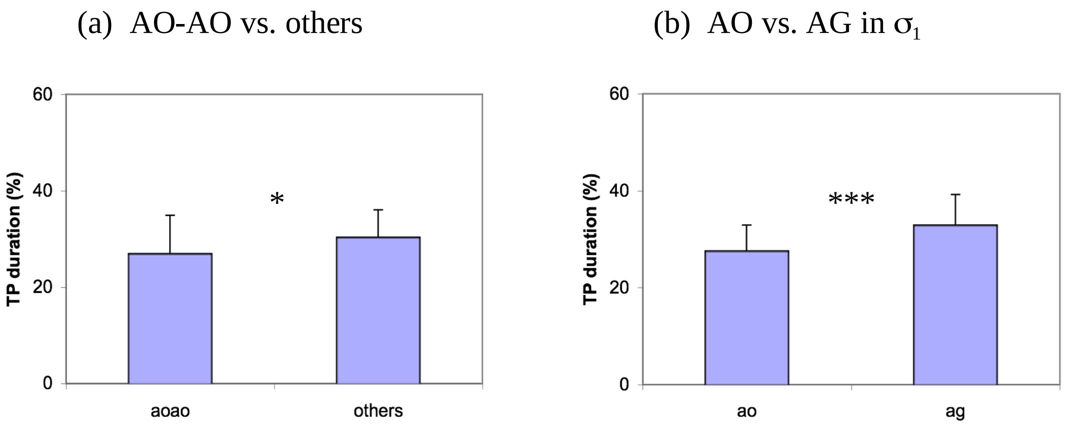
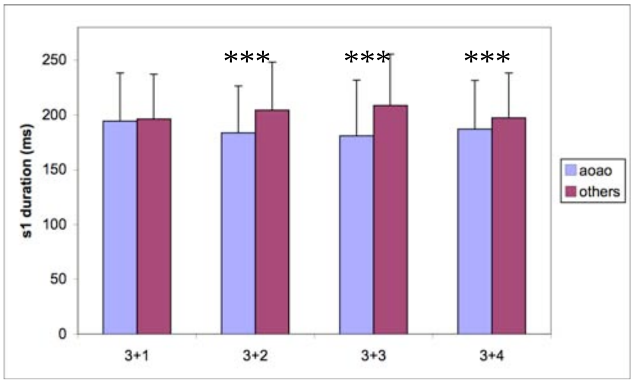

(a) AO-AO vs. others (b) AO vs. AG in σ_1

Figure 14. TP duration as a percentage of the entire σ_1 rhyme duration in 3+3.

We have replicated our turning point results in Experiment 1: the σ_1 turning point in 3+3 sequences is significantly lower and later in wug words than real words, which makes the tone more similar to the original Tone 3 in wug words, indicating incomplete application of the sandhi in wug words.

4.2.2. Rhyme duration

The results for σ_1 rhyme duration for all the tonal combinations are given in Figure 15. For the AO-AO vs. other-word-groups comparison, a Huynh-Feldt Repeated-Measures ANOVA shows that there is a significant Word-Group effect: F(1.000, 29.000)=58.058, p<0.001; the ANOVA results within each tone, summarized in Table 6, show that except for 3+1, the wug words have a significantly longer σ_1 rhyme duration than AO-AO. For the AO vs. AG comparison, the ANOVA again shows a significant Word-Group effect: F(1.000, 29.000)=58.576, p<0.001; the ANOVA results within each tone, also summarized in Table 6, show that the AG words have a significantly longer σ_1 rhyme duration than AO words for all of the tonal combinations.



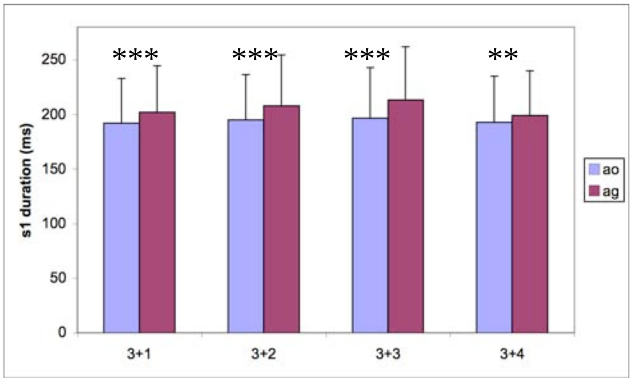


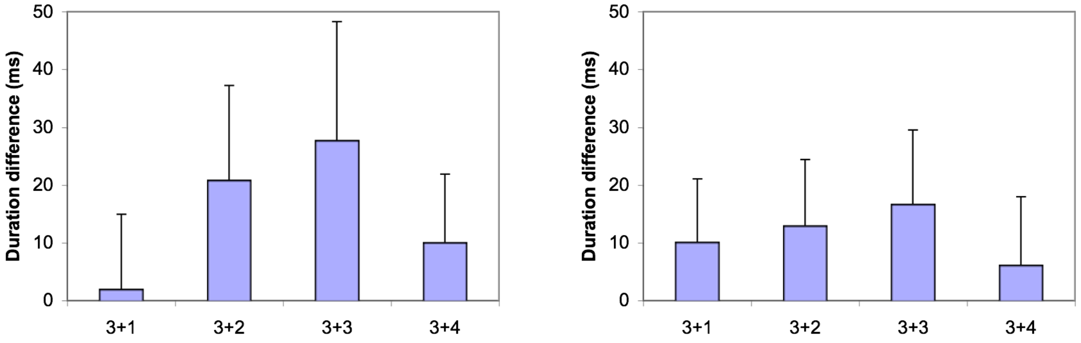
Figure 15. Rhyme duration of σ_1 for all tonal combinations. The two graphs represent the real-disyllable vs. wug-disyllable and real-σ_1 vs. wug-σ_1 comparisons, respectively.

Real-disyllable vs. wug-disyllable	
Tone 3 + Tone 1	F(1.000, 29.000)=0.698, p=0.410
Tone 3 + Tone 2	F(1.000, 29.000)=48.128, p<0.001
Tone 3 + Tone 3	F(1.000, 29.000)=54.432, p<0.001
Tone 3 + Tone 4	F(1.000, 29.000)=21.346, p<0.001

Real-σ_1 vs. wug-σ_1	
Tone 3 + Tone 1	F(1.000, 29.000)=25.382, p<0.001
Tone 3 + Tone 2	F(1.000, 29.000)=38.187, p<0.001
Tone 3 + Tone 3	F(1.000, 29.000)=50.444, p<0.001
Tone 3 + Tone 4	F(1.000, 29.000)=7.962, p=0.009

Table 6. One-way Huynh-Feldt Repeated-Measures ANOVA results for the σ_1 rhyme duration in all tonal combinations.

To compare the real vs. wug durational difference in different tonal combinations, we calculated the durational difference between AO-AO and other word groups as well as between σ_1=AO and σ_1=AG for each tonal combination, shown in Figure 16, and we conducted a one-way Huynh-Feldt Repeated-Measures ANOVA with Tone as the independent variable and the durational difference as the dependent variable for each real vs. wug comparison. The ANOVA results show that for the AO-AO vs. other-word-groups comparison, Tone has a significant effect on the durational difference between the two word groups (F(2.441, 70.783)=22.032, p<0.001), and post-hoc tests show that the 3+3 and 3+2 sequences exhibit significantly greater durational differences than 3+1 and 3+4 (p<0.001 for all comparison except for 3+2 vs. 3+4, which is at p<0.01). No other pairwise differences were found. For the σ_1=AO vs. σ_1=AG comparison, Tone also has a significant effect on the durational difference between the two word groups (F(3.000, 87.000)=6.174, p<0.005), and post-hoc tests show that 3+3 and 3+2 exhibit significantly greater durational differences than 3+4 (p<0.005 for 3+3 vs. 3+4; p<0.05 for 3+2 vs. 3+4).


(a) AO-AO vs. others

(b) AO vs. AG in σ_1

Figure 16. σ_1 rhyme duration differences between wug and real words for all tonal combinations.

The σ_1 rhyme duration data here differ from that of Experiment 1 in that there is an overall significantly longer duration for wug words than real words regardless of the tonal combination. But the durational *difference* in σ_1 rhyme between real and wug words is dependent on the tonal combination. 3+3 and 3+2 sequences induced significantly greater durational differences between real and wug words than the other tonal sequences. The numerical differences between 3+3 and 3+2 observed in Figure 16, though in the expected direction, did not reach statistical significance. These results indicate that in wug words, 3+3 and 3+2 sequences may have involved incomplete sandhi application, which would give the first syllable a longer duration. They are again consistent with a synchronic approach that take into account both phonetics and lexical frequency.

4.3. Discussion

Our data provide converging evidence with Experiment 1 for the lower application accuracy of the Third-Tone Sandhi than the Half-Third Sandhi. In all f_0 comparisons between real and wug words in both Experiment 1 and Experiment 2, the contour shape of 3+3 sequences is the only comparison that consistently shows a significant difference. Moreover, the properties of the difference are consistent across comparisons and experiments: the turning point of the sandhi tone is significantly lower and later in wug words than in real words, and similarly to Experiment 1, these differences are not caused by the *non-application* of the sandhi to a limited number of tokens/speakers, indicating that the sandhi is incompletely applied to a large number of wug words.

The potential frequency effects observed in Experiment 1 are also replicated here. The 3+2 sequences exhibited differences between real and wug words in that the sandhi tones in wug words showed properties of non-application — existence of a turning point and a longer duration. But the difference in f_0 shape is less consistent than in 3+3. This is consistent with an approach that encodes the effects of both phonetics and frequency, but not with a frequency-only approach, which would predict a more consistent difference between real and wug words in 3+2 than 3+3.

Also as in Experiment 1, the difference between the Third-Tone Sandhi and the Half-Third Sandhi is more apparent in the real-σ_1 vs. wug-σ_1 comparison, as indicated by the equal or more significant difference for the Third-Tone Sandhi and the equal or less significant difference for the Half-Third Sandhi between the two word groups for all f_0 measures.

5. General discussion

5.1. The relevance of phonetics to synchronic phonology

Our f_0 data from the two wug test experiments, including the pitch track, turning point, and duration, collectively support our hypothesis that there is a difference in productivity between the two tone sandhi patterns in Mandarin: the more innovative sandhi that has a stronger phonetic basis — the Half-Third Sandhi — applies accurately to wug words except for 3+2, which has the lowest type frequency; the sandhi with a longer history and more opaque phonetic basis — the Third-Tone Sandhi — applies incompletely to wug words, as evidenced by the significantly lower and later turning point in the sandhi tone.

The f_0 data suggest that phonological patterns with different degrees of phonetic basis have different synchronic statuses: there is a bias that favors the pattern that has a stronger phonetic basis. Lexical frequency by itself cannot account for the data patterns due to two reasons. First, the Half-Third Sandhi behaves differently in different environments, indicating that speakers do not pool these environments together in how they have internalized the sandhi. Therefore, it is inaccurate to say that the Third-Tone Sandhi has an overall lower frequency than the Half-Third Sandhi; more appropriately, it has a lower type frequency than the Half-Third Sandhi in 3+1 and 3+4, but higher type frequency than the Half-Third Sandhi in 3+2. Second, the difference between real words and wug words is more consistently observed in 3+3 than 3+2, as indicated by the rhyme duration data in Experiment 1 and the f_0 data in Experiment 2. A frequency-only account would have expected the opposite.

The phonetic effect manifests itself here gradiently in the following sense: the sandhi with a weaker phonetic motivation applies without fail to the wug words, but the application is incomplete, in that the sandhi tone bears more resemblance to the base tone than the sandhi tone in real words. This, in a way, is a more subtle gradient effect than the one in which the pattern applies to only a certain percentages of the structure that satisfies its environment, as shown by many other works on gradience and exceptionality in phonology (e.g., Zuraw 2000, 2007, Frisch and Zawaydeh 2001, Ernestus and Baayen 2003, Hayes and Londe 2006, Pierrehumbert 2006, Coetzee 2008a, Coetzee and Pater 2008, Zhang and Lai 2008, Zhang et al. 2009a, b). Methodologically, this result points to the importance of detailed phonetic studies that can reveal patterns that traditionally escaped the attention of phonologists, but could potentially shed light on issues of theoretical contention. This finds parallel in the discovery of incomplete neutralization in many processes thought to be neutralizing, such as final devoicing in a host of languages (e.g., Charles-Luce 1985, Slowiaczek and Dinnsen 1985, Port and Crawford 1989, Warner et al. 2004), English flapping (Dinnsen 1984, Patterson and Connine 2001, Zue and Laferriere 1979), and Mandarin Third-Tone Sandhi (Peng 2000).[12]

---

[12] An anonymous reviewer points out that the results here are in fact the opposite of what is expected of a comparison between a "phonological" and a "phonetic" process, as conventional wisdom would have us believe that a more "phonological" process tends to be more categorical, while a "phonetic" process is more likely to exhibit gradient properties (e.g., Keating 1984, 1990, Pierrehumbert 1990, Cohn 1993). However, as we mentioned in §2, the difference between the two sandhis in question lies in the degree of their phonetic motivation and not in a binary "phonological" vs. "phonetic" distinction. Both of the sandhis are "phonological" in that they involve language-specific tone changes that cannot be predicted simply by tonal coarticulation.

5.2. Frequency effects

As we have argued above, frequency effects *alone* cannot account for our data. But frequency does seem to correlate positively with sandhi productivity: the Half-Third Sandhi in 3+2, which has the lowest type frequency, has the lowest application accuracy in wug words among all Half-Third Sandhi environments, and the inaccurate application can be characterized as incomplete application of the sandhi, just like what we have observed for the Third-Tone Sandhi. The frequency effects here are also of a slightly different nature than the frequency matching of patterned exceptionality in the lexicon in wug tests (Zuraw 2000, Albright 2002, Albright and Hayes 2003, Ernestus and Baayen 2003, Hayes and Londe 2006, among others) — the pattern here is exceptionless in the lexicon, but is of lower frequency than other non-competing patterns. The effects are also subtler than a comparable case — Taiwanese tone sandhi — documented in Zhang and Lai (2008) and Zhang et al. (2009a, b), in which frequency differences in the lexicon cause application rate differences in wug tests: the application rates here are consistently 100%; but the degree of application differs.[13]

But in the wug test results, both patterns show gradience — Third-Tone Sandhi in 3+3, and Half-Third Sandhi in 3+2. This mirrors the results from the incomplete neutralization literature.

[13] An anonymous reviewer questions whether the lexical frequency differences between Tone 2 and other tones are big enough to have noticeable effects in productivity. It is difficult, and possibly impractical to quantify a minimum difference in lexical frequency that can elicit an effect on productivity. Studies that illustrate the effects of frequency on phonological productivity (e.g., Zuraw 2000, Ernestus and Baayen 2003, Hayes and Londe 2006, Zhang and Lai 2008, Zhang et al. 2009a, b) and production (e.g., Bybee 2000, Jurafsky et al. 2001, Ernestus et al. 2006) typically

5.3. Alternative interpretations

Finally, we consider four other alternative interpretations to our results here, all of which were suggested by anonymous reviewers, to whom we are grateful.

An important alternative to consider is whether it is possible to treat the Third Tone as underlyingly 21 and insert a High pitch to the right when the tone occurs phrase finally. The insertion of a pre- or post-[-αT] is crosslinguistically attested and referred to as a "bounce" effect by Hyman (2007). The tone sandhi in the Third-Tone Sandhi can then be considered as OCP avoidance, and the Half-Third Sandhi is simply nonexistent. The 21 underlying form for the Third Tone is a particularly attractive option for Taiwan Mandarin, in which the Third Tone is pronounced as [21] even in final position. This position is technically workable for Beijing Mandarin, but difficult to defend from a typological perspective. First, Northern Chinese dialects, to which Mandarin belongs, are known to have "right-dominant" sandhis that protect domain-final tones and change nonfinal tones (Yue-Hashimoto 1987, Zhang 2007). It is not clear why Mandarin would be an exception. Second, while contour simplification in nonfinal positions is extremely common crosslinguistically, contour complication, even in final position, is quite rare.

---

use regression analyses or binary comparisons between high vs. low frequencies. However, in Hayes and Londe's (2006) study on variable backness harmony in Hungarian, a less than 8% harmony rate difference between two types of stems (N and NN, N=neutral) in a web-based corpus does translate into comparable a productivity difference in a wug test; in Zhang and Lai's (2008) and Zhang et al's (2009a, b) studies on tone sandhi productivity in Taiwanese, type and token frequencies differences that are smaller than what is observed here are also shown to correlate significantly with the productivity results.

Yue-Hashimoto's (1987) typology of Chinese tone sandhi systems identified close to 100 cases of contour leveling or simplification, but only three cases of contour complication. It is not clear why we would want to entertain a typologically odd analysis when a better attested option is available. These points are also made in Zhang (2007) (p. 260, fn. 2).

The second alternative relates to our discussion earlier that the Third-Tone Sandhi is sensitive to syntactic information, while the Half-Third Sandhi is not. Another manifestation of this is that the Third-Tone Sandhi sometimes does not apply across a [NP][VP] boundary, as shown in (5a): the [li] syllable has the option of not undergoing the Third-Tone Sandhi, and as a consequence, a [21 21] sequence obtains in the output. This potentially makes the processing of the Third-Tone Sandhi more difficult, as the speaker needs to access the syntactic information in order to determine whether the Third-Tone Sandhi should apply. However, the stimuli that we used in the experiments were all disyllabic, and 3+3 disyllabic sequences do not have the option of not undergoing the sandhi, even if the syntactic configuration is [NP][VP], as shown in (5b). Therefore, the syntactic information is immaterial to the stimuli that we used in the experiments.

(5) Third-Tone Sandhi in [NP][VP]:

a.	[[lɑu	li]	[mai	ɕjɛ]]	
	old	Li	buy	shoes	'Old Li buys shoes'
	213	213	213	35	Input
	35	**21**	**21**	35	Output 1
	35	35	21	35	Output 2

b.	[[ni]	[xɑu]][14]	
	you	good	‘How are you?’
	213	213	Input
	35	213	Output
	*21	213	

The third alternative is that the productivity difference stems from the nature of lexical listing in that the Third-Tone Sandhi is lexically listed, while the Half-Third Sandhi is productively derived by the markedness and faithfulness interactions in an OT grammar. This is consistent with the fact that the Third-Tone Sandhi has a long history and thus may have a higher degree of lexicalization. Therefore, even if the two sandhis do differ in phonetic motivation, synchronically speaking, it is their difference in lexical listing that causes the productivity difference.

There are two arguments against this alternative. First, if the nature of lexical listing is truly different between the two sandhis, then we would expect the Third-Tone Sandhi to be entirely unproductive while the Half-Third Sandhi to be entirely productive regardless of lexical frequency. However, we observed a gradient difference between the two sandhis, and the Half-Third Sandhi is affected by lexical frequency. These gradient effects, we believe, are better captured by an analysis that is gradient in nature rather than one that imposes a categorical distinction between the two sandhis based on the presence

---

[14] The adjective [xɑu] ‘good’ is traditionally treated as an adjectival verb in Chinese syntax (see Li and Thompson 1981).

vs. absence of lexical listing. Second, despite the long history of the Third-Tone Sandhi, its application to disyllabic words in Mandarin is in fact exceptionless, just like the Half-Third Sandhi. Therefore, learners of Mandarin cannot conclude purely from input statistics that the former has a higher degree of lexicality than the latter. In order to reach this conclusion, it seems that the learner still has to access the phonetic nature of the sandhis, indicating the synchronic relevance of phonetics.

The final alternative capitalizes on the observation that the subjects produced the Half-Third Sandhi after hearing only one full Third Tone in σ_1 position followed by a different tone, but produced the Third-Tone Sandhi after hearing two identical full Third Tones. It is thus possible that the production of the Third-Tone Sandhi is influenced by a greater perceptual perseveration effect from the input than that of the Half-Third Sandhi, which causes the nonce syllable in σ_1 position of 3+3 to have more characteristics of Tone 3.

Although this approach correctly predicts incomplete neutralization in both real and wug words (see fn. 6 for results on incomplete neutralization between 3+3 and 2+3 in real word productions), it cannot predict the *difference* between them, as it is not clear why the perceptual perseveration effect should be stronger for wug words than for real words. But more importantly, the approach assumes tone priming irrespective of segmental contents, as it assumes that the two Third Tones both have an effect on the subjects' production of the sandhied Third Tone even though the second syllable has completely different segmental contents from the syllable undergoing sandhi. However, whether tone by itself is an effective prime in a tone language is a controversial issue. Although Culter and Chen (1995) showed that in Cantonese, tone and segments behave similarly as primes for lexical decision, more studies on Mandarin (Chen et al. 2002, Lee 2007) and Cantonese (Yip et al. 1998, Yip 2001) showed that priming effects in lexical decision and production latency are only found when the prime and the target share either segmental contents or segmental contents + tone. Tone by itself is an ineffective prime. This casts further doubt on the workability of this alternative.

6. Conclusion

In this paper, we have proposed a novel research paradigm to test the relevance of phonetics to synchronic phonology — wug testing of patterns differing in phonetic motivations that coexist in the same language. By directly addressing existing native patterns and allowing easier control of confounding factors such as lexical frequency, the wug test paradigm provides converging evidence with other research paradigms that have been used to test this issue, such as the study of phonological acquisition in a first language and the artificial language paradigm. The language we used was Mandarin Chinese, which has two tone sandhi patterns that differ in their degrees of phonetic motivation, and our wug tests showed that Mandarin speakers applied the sandhi with a stronger phonetic motivation — the Half-Third Sandhi — to wug words with a greater accuracy than the phonetically more opaque sandhi — the Third-Tone Sandhi, thus supporting the direct relevance of phonetics to synchronic phonology. We have also shown that lexical frequency is relevant to the application of the Half-Third Sandhi in wug words, as reflected in the lower accuracy of the sandhi in the 3+2 environment. However, lexical frequency alone cannot account for the low sandhi accuracy of 3+3, as the sandhi tone differences between real and wug words are more consistent for 3+3 than 3+2, even though 3+2 has a lower lexical frequency.

We recognize that our position that phonetics, likely in the form of substantive biases, is part of the design feature of grammar construction complicates the search for phonological explanations in the following sense: it potentially creates a duplication problem for patterns whose explanation may come from either the substantive bias or misperception; how does one, then, tease apart which one truly is the explanatory factor? This problem is pointed out by Hansson (2008: p.886), for example. We surmise that the answer will not come from these individual cases for which the explanation may truly be ambiguous, but comprehensive experimental studies on many different patterns to see which approach makes better predictions on both the speakers' internal knowledge and the evolution of these patterns in general. Therefore, the study reported here can be simply viewed as fodder for future research that investigates the phonetics-phonology relationship. For example, to conduct similar studies, we need the two patterns under comparison to satisfy the following conditions: (a) they have comparable triggering environments; (b) they are of comparable productivity in the native lexicon; (c) they have comparable frequencies of occurrence in the native lexicon; and (d) they differ in their degrees of phonetic motivation. There are many other Chinese dialects, especially the Wu and Min dialects, that have considerably more intricate patterns of tone sandhi than Mandarin, and we can often find differences in the degree of phonetic motivation among the sandhi patterns in these dialects. We hope our study on Mandarin will lead to similar research in other Chinese dialects, which will make further contributions to the phonetics-phonology interface debate.

Starting from Hsieh's seminal works on wug-testing Taiwanese tone sandhi, the productivity of complicated tone sandhi patterns has been a long-standing question in Chinese phonology. This is especially true for sandhi patterns that involve phonological opacity (e.g., the tone circle in Southern Min; see Chen 2000 for examples) and syntactic dependency (e.g., the different sandhi patterns that Subject-Predicate and Verb-Object compounds undergo in Pingyao; see Hou 1980). We hope that our research will inspire more psycholinguistic testing of these patterns that will shed light on this question. Some results on how sandhi productivity is gradiently influenced by phonological opacity have in fact been obtained for Taiwanese (Zhang and Lai 2008, Zhang et al. 2009a, b).

Finally, our results here shed additional light on the nature of gradience in phonology. Not only are the phonetic and frequency effects observed here gradient, they are gradient in an interesting way: the sandhis may apply 100% to all wug words, but they apply *incompletely* in that the sandhi tone bears more resemblance to the base tone than the sandhi tone in real words. This complements the well-attested gradient effects whereby a phonological pattern only applies to a certain percentage of the experimental test items.[15] This observation is both methodologically and theoretically significant: methodologically, it further demonstrates the importance of careful acoustic studies, which can reveal phonological patterns that have hitherto escaped our attention;

---

[15] As one anonymous reviewer suggests, whether any predictions can be made about the nature of gradience in productivity is an independently interesting question. Previous works have shown that it may be influenced by multiple factors, including the nature of the gradience in the lexicon (Zuraw 2000, 2007, Pierrehumbert 2006, Hayes and Londe 2006, among others) and phonological opacity (Zhang and Lai 2008, Zhang et al. 2009a, b). But more empirical research is needed to identify both the factors and the mechanism with which the factors interact with each other.

theoretically, it forces us to rethink theoretical models of phonology, which need to provide a viable explanation for the multiple layers of gradience.

References

Albright, Adam (2002). Islands of reliability for regular morphology: Evidence from Italian. *Language* 78.4: 684-709.

Albright, Adam, Argelia Andrade, and Bruce Hayes (2001). Segmental environments of Spanish diphthongization. In Adam Albright and Taehong Cho (eds.), *UCLA Working Papers in Linguistics* 7 (*Papers in Phonology* 5): 117-151.

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

Anderson, Stephen (1981). Why phonology isn't "natural." *Linguistic Inquiry* 12: 493-539.

Archangeli, Diana and Douglas Pulleyblank (1994). *Grounded phonology*. Cambridge, MA: The MIT Press.

Baayen, R. Harald (1992). Quantitative aspects of morphological productivity. In Geert E. Booij and Jaap van Marle (eds.), *Yearbook of Morphology* 1991. Dordrecht: Kluwer Academic Publishers. 109-149.

Baayen, R. Harald (1993). On frequency, transparency, and productivity. In Geert E. Booij and Jaap van Marle (eds.), *Yearbook of Morphology* 1992. Dordrecht: Kluwer Academic Publishers. 181-208.

Bach, Emmon and Robert Harms (1972). How do languages get crazy rules? In Robert Stockwell and Ronald Macauley (eds.), *Linguistic change and generative theory*. Bloomington, IN: Indiana University Press. 1-21.

Berko, Jean (1958). The child's learning of English morphology. *Word* 14: 150-177.

Blevins, Juliette (2004). *Evolutionary phonology: The emergence of sound patterns.* Cambridge, UK: Cambridge University Press.

Blevin, Juliette (2006). A theoretical synopsis of Evolutionary Phonology. *Theoretical Linguistics* 32.2: 117-166.

Blevins, Juliette and Andrew Garrett (1998). The origins of consonant-vowel metathesis. *Language* 74: 508-556.

Bley-Vroman, Robert (1988). The fundamental character of foreign language learning. In William E. Rutherford and Michael Sharwood-Smith (eds.), *Grammar and second language teaching: A book of readings.* New York: Newbury House/Harper and Row. 19-30.

Boersma, Paul (1998). *Functional phonology: Formalizing the interaction between articulatory and perceptual drives.* The Hague: Holland Academic Graphics.

Boersma, Paul and David Weenink (2003). *Praat: doing phonetics by computer* [Computer program]. Version retrieved September 2003 from http://www.praat.org/.

Buckley, Eugene (2000). On the naturalness of unnatural rules. *Proceedings from the Second Workshop on American Indigenous Languages, UCSB Working Papers in Linguistics*, Vol. 9.

Buckley, Eugene (2002). *Rule naturalness and the acquisition of phonology.* Talk presented at the 2nd North American Phonology Conference, Concordia University, Montréal.

Buckley, Eugene (2003). Children's unnatural phonology. *BLS 29 Proceedings*: 523-534.

Bybee, Joan (1985). *Morphology: a study of the relation between meaning and form.* Philadelphia: Benjamins.

Bybee, Joan (2000). The phonology of the lexicon: evidence from lexical diffusion. In Michael Barlow and Suzanne Kemmer (eds.), *Usage-based models of language*. Stanford, CA: CSLI Publications.

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

Bybee, Joan (2006). *Frequency of use and the organization of language*. Oxford: Oxford University Press.

Bybee, Joan and Elly Pardo (1981). On lexical and morphological conditioning of alternations: A nonce-probe experiment with Spanish verbs. *Linguistics* 19: 937-968.

Chambers, Kyle E., Kristine H. Onishi, and Cynthia Fisher (2003). Infants learn phonotactic regularities from brief auditory experience. *Cognition* 87: B69-B77.

Chao, Yuen Ren (1948). *Mandarin primer: An intensive course in spoken Chinese.* Cambridge, MA: Harvard University Press.

Chao, Yuen Ren (1968). *A grammar of spoken Chinese.* Berkeley and Los Angeles: University of California Press.

Charles-Luce, Jan (1985). Word-final devoicing in German: Effects of phonetic and sentential contexts. *Journal of Phonetics* 13: 309-324.

Chen, Jenn-Yeu, Train-Min Chen, and Gary S. Dell (2002). Word-form encoding in Mandarin Chinese as assessed by the implicit priming task. *Journal of Memory and Language* 46.4: 751-781.

Chen, Matthew Y. (2000). *Tone sandhi: Patterns across Chinese dialects.* Cambridge, UK: Cambridge University Press.

Chomsky, Noam and Morris Halle (1968). *The sound pattern of English.* New York: Harper and Row.

Chomsky, Noam (1986). *Knowledge of Language: Its Nature, Origin, and Use.* New York: Praeger.

Coetzee, Andries W. (2008) Grammaticality and ungrammaticality in phonology. *Language* 84: 218-257.

Coetzee, Andries W. and Joe Pater (2008). Weighted constraints and gradient restrictions on place co-occurrence in Muna and Arabic. *Natural Language and Linguistic Theory* 26: 289-337.

Cohn, Abigail (1993). Nasalization in English: Phonology or phonetics. *Phonology* 10.1: 43-82.

Cook, Vivian J. (1969). The analogy between first and second language learning. *International Review of Applied Linguistics* 7.3: 207-216.

Cook, Vivian J. (1994). The metaphor of access to Universal Grammar. In Nick C. Ellis (ed.), *Implicit and explicit learning of languages.* London: Academic Press. 477-502.

Court, Christopher (1985). Observations on some cases of tone sandhi. In Graham Thurgood, James A. Matisoff, and David Bradley (eds.), *Linguistics of the Sino Tibetan area: The state of the art.* Canberra: Department of Linguistics, Research School of Pacific Studies, Australian National University. 125-137.

Cutler, Anne and Hsuan-Chih Chen (1995). Phonological similarity effects in Cantonese word recognition. *Proceedings of the 13th international Congress of Phonetic Sciences*. Stockholm, Sweden. 106-109.

Cutler, Anne and Hsuan-Chih Chen (1997). Lexical tone in Cantonese spoken-word processing. *Perception and Psychophysics* 59.2: 165-179.

Da, Jun (1998). *Chinese text computing — bigrams and statistical measures*. http://lingua.mtsu.edu/chinese-computing/old-version1/statistics/mi.html.

Da, Jun (2004). *Chinese text computing — syllable frequencies with tones*. http://lingua.mtsu.edu/chinese-computing/phonology/syllabletone.php.

Dell, Gary S., Kristopher D. Reed, David R. Adams, and Antje S. Meyer (2000). Speech errors, phonotactic constraints, and implicit learning: A study of the role of experience in language production. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 26: 1355-1367.

Dinnsen, Daniel A. (1984). A re-examination of phonological neutralization. *Journal of Linguistics* 21: 265-279.

Duanmu, San (2000). *The phonology of Standard Chinese*. Oxford: Oxford University Press.

Dulay, Heidi C., Marina Burt, and Stephen Krashen (1982). *Language two*. Oxford: Oxford University Press.

Ellis, Rod (1994). *The study of second language acquisition*. Oxford: Oxford University Press.

Ernestus, Mirjam and R. Harald Baayen (2003). Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language* 79: 5-38.

Ernestus, Mirjam, Mybeth Lahey, Femke Verhees, and R. Harald Baayen (2006).

Lexical frequency and voice assimilation. *Journal of the Acoustical Society of America* 120: 1040-1051.

Fleischhacker, Heidi (2001). Cluster-dependent epenthesis asymmetries. In Adam Albright and Taehong Cho (eds.), *UCLA Working Papers in Linguistics* 7, *Papers in Phonology* 5: 71-116.

Flemming, Edward (2001). Scalar and categorical phenomena in a unified model of phonetics and phonology. *Phonology* 18.1: 7-44.

Frisch, Stefan A. and Bushra Adnan Zawaydeh (2001). The psychological reality of OCP-Place in Arabic. *Language* 77.1: 91-106.

Hale, Mark and Charles Reiss (2000). Substance abuse and dysfunctionalism: current trends in phonology. *Linguistic Inquiry* 31: 157-169.

Hansson, Gunnar Ólafur (2001). *Theoretical and Typological Issues in Consonant Harmony*. Ph.D. dissertation, University of California, Berkeley.

Hansson, Gunnar Ólafur (2008). Diachronic explanations of sound patterns. *Language and Linguistics Compass* 2.5: 859-893.

Hayes, Bruce and Zsuzsa Cziráky Londe (2006). Stochastic phonological knowledge: the case of Hungarian vowel harmony. *Phonology* 23: 59-104.

Hayes, Bruce and Donca Steriade (2004). Introduction: The phonetic bases of phonological markedness. In Hayes et al. (2004). 1-33.

Hayes, Bruce (1999). Phonetically driven phonology: the role of Optimality Theory and inductive grounding. In Michael Darnell, Edith Moravcsik, Michael Noonan, Frederick Newmeyer, and Kathleen Wheatly (eds.), F*unctionalism and formalism in linguistics, vol. 1.* Amsterdam: John Benjamins. 243-285.

Hayes, Bruce, Donca Steriade, and Robert Kirchner (2004). *Phonetically based phonology*. Cambridge, UK: Cambridge University Press.

Hou, Jing-Yi (1980). Pingyao fangyan de liandu biandiao (Tone sandhi in the Pingyao dialect). *Fangyan* (Dialects) 1980.1: 1-14.

Hsieh, Hsin-I (1970). The psychological reality of tone sandhi rules in Taiwanese. In *Papers from the 6th Meeting of the Chicago Linguistic Society*. Chicago: Chicago Linguistic Society. 489-503.

Hsieh, Hsin-I (1975). How generative is phonology. In E. F. Koerner (ed.), *The transformational-generative paradigm and modern linguistic theory*. Amsterdam: John Benjamins. 109-144.

Hsieh, Hsin-I (1976). On the unreality of some phonological rules. *Lingua* 38: 1-19.

Hyman, Larry (2001). On the limits of phonetic determinism in phonology: *NC revisited. In Elizabeth Hume and Keith Johnson (eds.), *The role of speech perception in phonology*. New York: Academic Press. 141-186.

Hyman, Larry (2007). Universals of tone rules: 30 years later. In Tomas Riad and Carlos Gussenhoven (eds.), *Tones and tunes, vol. 1: Typological studies in word and sentence prosody*. Berlin and New York: Mouton de Gruyter. 1-34.

Jun, Jongho (1995). *Perceptual and articulatory factors in place assimilation: An Optimality-Theoretic approach*. Ph.D. dissertation, UCLA.

Jun, Jongho (2004). Place assimilation. In Hayes et al. (2004). 58-86.

Jurafsky, Daniel, Alan Bell, Michelle Gregory, and William D. Raymond (2001). Probabilistic relations between words: Evidence from reduction in lexical production. In Joan Bybee and Paul J. Hopper (eds.), *Frequency and the* *emergence of linguistic structure*. Amsterdam: Benjamins. 229–254.

Jusczyk, Peter W., Paul Smolensky, Karen Arnold, and Elliott Moreton (2003). Acquisition of nasal place assimilation by 4.5-month-old infants. In Derek Houston, Amanda Seidl, George J. Hollich, Elizabeth Johnson, and Ann Marie Jusczyk (eds.), *Jusczyk Lab Final Report*. Retrieved from http://hincapie.psych.purdue.edu/Jusczyk

Kang, Yoonjung (2003). Perceptual similarity in loanword adaptation: English postvocalic word-final stops in Korean. *Phonology* 20.3: 219-273.

Keating, Patricia A. (1984). Phonetic and phonological representation of stop consonant voicing. *Language* 60: 286-319.

Keating, Patricia A. (1990). Phonetic representations in a generative grammar. *Journal of Phonetics* 18: 321-334.

Kenstowicz, Michael (2007). Salience and similarity in loanword adaptation: A case study from Fijian. *Language Sciences* 29: 316-340.

Kingston, John and Randy Diehl (1994). Phonetic knowledge. *Language* 70: 419-454.

Kiparsky, Paul (1982). Lexical morphology and phonology. In I. S. Yang (ed.), *Linguistics in the morning calm*. Seoul: Hanshin. 3-91.

Kiparsky, Paul (1985). Some consequences of lexical phonology. *Phonology Yearbook* 2: 85-138.

Kiparsky, Paul (2006). The amphichronic program vs. evolutionary phonology. *Theoretical Linguistics* 32.2: 217-236.

Kirchner, Robert (2000). Geminate inalterability and lenition. *Language* 76.3: 509-545.

Kirchner, Robert (2001). *An effort-based approach to consonant lenition.* New York: Routledge.

Kirchner, Robert (2004). Consonant lenition. In Hayes et al. (2004). 313-345.

Lee, Chao-Yang (2007). Does horse active mother? Processing lexical tone in form priming. *Language and Speech* 50.1: 101-123.

Li, Charles N. and Sandra A. Thompson (1981). Mandarin Chinese: a functional reference grammar. Berkeley, Los Angeles, London: University of California Press.

Lin, Yen-Hwei (2007). *The sounds of Chinese*. Cambridge, UK: Cambridge University Press.

MacWhinney, Brian (1978). The acquisition of morphophonology. *Monographs of the Society for Research in Child Development*, Serial No. 174, Vol. 43, Nos. 1-2.

Mei, Tsu-lin (1977). Tones and tone sandhi in 16th century Mandarin. *Journal of Chinese Linguistics* 5: 237-260.

Menn, Lise and Carol Stoel-Gammon (1995). Phonological development. In Paul Fletcher and Brian MacWhinney (eds.), *The handbook of child language*. Oxford: Blackwell. 335-359.

Mohanan, K. P. (1982). *Lexical phonology*. Ph.D. dissertation, MIT. Distributed by Bloomington: Indiana University Linguistics Club.

Mohanan, K. P. (1993). Fields of attraction in phonology. In John Goldsmith (ed.), *The last phonological rule: Reflections on constraints and derivations.* Chicago & London: The University of Chicago Press. 61-116.

Moore, Corinne B. and Allard Jongman (1997). Speaker normalization in the perception of Mandarin Chinese tones. *Journal of the Acoustical Society of America* 102: 1864-1877.

Moreton, Elliott (2008). Analytic bias and phonological typology. *Phonology* 25: 83-127.

Myers, James (2002). An analogical approach to the Mandarin syllabary. *Chinese Phonology* 11, *Special Issue — Glide, Syllable and Tone*: 163-190.

Oakes, Michael P. (1998). *Statistics for corpus linguistics.* Edinburgh: Edinburgh University Press.

Ohala, John J. (1981). The listener as a source of sound change. *Papers from the Parasession on Language and Behavior, Chicago Linguistic Society* 18: 178-203.

Ohala, John J. (1990). The phonetics and phonology of aspects of assimilation. In John Kingston and Mary E. Beckman (eds.), *Papers in laboratory phonology I: between the grammar and the physics of speech.* Cambridge, UK: Cambridge University Press. 258-275.

Ohala, John J. (1993). Coarticulation and phonology. *Language and Speech* 36: 155-170.

Ohala, John J. (1997). The relation between phonetics and phonology. In William Hardcastle and John Laver (eds.), The *handbook of phonetic sciences.* Oxford: Blackwell. 674-694.

Onishi, Kristine H., Kyle E. Chambers, and Cynthia Fisher (2002). Learning phonotactic constraints from brief auditory experience. *Cognition* 83: B13-B23.

Patterson, David and Cynthia M. Connine (2001). Variant frequency in flap production: a corpus analysis of variant frequency in American English flap production. *Phonetica* 58: 254-275.

Peng, Shu-Hui (2000). Lexical versus ‘phonological’ representations of Mandarin sandhi tones. In Michael B. Broe and Janet B. Pierrehumbert (eds.), *Language acquisition and the lexicon: Papers in laboratory phonology V*. Cambridge, UK: Cambridge University Press. 152-167.

Peperkamp, Sharon and Emmanuel Dupoux (2007). Learning the mapping from surface to underlying representations in an artificial language. In Jennifer Cole and Jose Hualde (eds.), *Laboratory Phonology* 9. Berlin: Mouton de Gruyter. 315-338.

Peperkamp, Sharon, Katrin Skoruppa, and Emmanuel Dupoux (2006). The role of phonetic naturalness in phonological rule acquisition. In David Bamman, Tatiana Magnitskaia, and Colleen Zaller (eds.), *Proceedings of the 30th Annual Boston University Conference on Language Development.* Somerville, MA: Cascadilla Press. 464-475.

Pierrehumbert, Janet B. (1990). Phonological and phonetic representation. *Journal of Phonetics* 18: 375-94.

Pierrehumbert, Janet B. (2003). Probabilistic phonology: Discrimination and robustness. In Rens Bod, Jennifer Hay, and Stefanie Jannedy (eds.), *Probabilistic linguistics*. Cambridge, MA: MIT Press. 177-228.

Pierrehumbert, Janet B. (2006). The statistical basis of an unnatural alternation. In Louis Goldstein, Douglas H. Whalen, and Catherine Best (eds.), *Laboratory Phonology 8, Varieties of Phonological Competence*. Berlin: Mouton de Gruyter. 81-107.

Port, Robert, and Penny Crawford (1989). Incomplete neutralization and pragmatics in German. *Journal of Phonetics* 17: 257-282.

Prince, Alan and Paul Smolensky (1993). *Optimality Theory: Constraint interaction in generative grammar*. Ms., Rutgers University and University of Colorado, Boulder.

Pycha, Anne, Pawel Nowak, Eurie Shin, and Ryan Shosted (2003). Phonological rule-learning and its implications for a theory of vowel harmony. In Gina Garding and Mimu Tsujimura (eds.), *WCCFL 22 Proceedings*. Somerville, MA: Cascadilla Press. 423-435.

Qian, Zeng-Yi, Guang-Qi Zhu (1998). *Jinanhua yindang* (*A record of the Jinan dialect*). Shanghai: Shanghai Jiaoyu Chubanshe (Shanghai Education Press).

Reber, Arthur S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior* 6: 855–863.

Reber, Arthur S. (1989). Implicit learning and tacit knowledge. *Journal of Experimental Psychology: General* 118: 219-235.

Redington, Martin and Nick Chater (1996). Transfer in artificial grammar learning: A reevaluation. *Journal of Experimental Psychology: General* 125: 123–138.

Seidl, Amanda and Eugene Buckley (2005). On the learning of arbitrary phonological rules. *Language Learning and Development* 1: 289-316.

Shen, Xiaonan Susan and Maocan Lin (1991). A perceptual study of Mandarin tones 2 and 3. *Language and Speech* 34: 145-156.

Shen, Xiaonan Susan, Maocan Lin, and Jingzhu Yan (1993). F0 turning point as an F0 cue to tonal contrast: A case study of Mandarin tones 2 and 3. *Journal of the Acoustical Society of America* 93: 2241-2243.

Shih, Chilin (1997). Mandarin Third Tone Sandhi and prosodic structure. In Wang Jialing and Norvin Smith (eds.), *Studies in Chinese phonology*. Berlin: Mouton de Gruyter. 81-123.

Silverman, Daniel (2006a). *A critical introduction to phonology: of sound, mind, and body*. London, UK: Continuum.

Silverman, Daniel (2006b). The diachrony of labiality in Trique, and the functional relevance of gradience and variation. In Louis M. Goldstein, Douglas H. Whalen, and Catherine T. Best (eds.), *Papers in laboratory phonology VIII*. Berlin: Mouton de Gruyter. 135-154.

Slobin, Dan Isaac (1985). *The crosslinguistic study of language acquisition, Vol. 2: Theoretical issues*. Hillsdale, NJ: Lawrence Erlbaum Associates. 1157-1256.

Slowiaczek, Louisa M. and Daniel A. Dinnsen (1985). On the neutralization status of Polish word-final devoicing. *Journal of Phonetics* 13: 325-341.

Stampe, David (1979). *A dissertation on Natural Phonology*. New York: Garland Publishing.

Steriade, Donca (1999). Phonetics in phonology: The case of laryngeal neutralization. In Matthew K. Gordon (ed.), *UCLA Working Papers in Linguistics, Papers in Phonology* 3: 25-146.

Steriade, Donca (2001). Directional asymmetries in place assimilation: A perceptual account. In Elizabeth Hume and Keith Johnson (eds.), *The Role of Speech Perception in Phonology*. New York: Academic Press. 219-250.

Steriade, Donca (2008). The phonology of perceptibility effects: the P-map and its consequences for constraint organization. In Kristin Hanson and Sharon Inkelas (eds.), *The nature of the word: Essays in honor of Paul Kiparsky*. Cambridge, MA: MIT Press. 151-180.

Steriade, Donca and Jie Zhang (2001). *Context-dependent similarity and the Romanian Semi-rime*. Invited talk at the 37th annual meeting of the Chicago Linguistic Society. Chicago, IL.

Taft, Marcus and Hsuan-Chih Chen (1992). Judging homophony in Chinese: the influence of tones. In Hsuan-Chih Chen and Ovid J. L. Tzeng (eds.), *Language processing in Chinese*. North-Holland: Elsevier Science Publishers B. V. 151-172.

Thatte, Victoria Anne (2007). *Phonetic motivation as a learning bias in phonological acquisition: an experimental study*. MA thesis, UCLA.

Wang, H. Samuel (1993). Taiyu biandiao de xinli texing (On the psychological status of Taiwanese tone sandhi). *Tsinghua Xuebao* (*Tsinghua Journal of Chinese Studies*) 23: 175-192.

Warner, Natasha, Allard Jongman, Joan Sereno, J., and Rachèl Kemper (2004). Incomplete neutralization of sub-phonemic durational differences in production and perception of Dutch. *Journal of Phonetics* 32: 251-276.

Wen, Duan-Zheng, Ming Shen (1999). *Taiyuanhua yindang* (*A record of the Taiyuan dialect*). Shanghai: Shanghai Jiaoyu Chubanshe (Shanghai Education Press).

Wilson, Colin (2003). Experimental investigation of phonological naturalness. In Gina Garding and Mimu Tsujimura (eds.), *WCCFL 22 Proceedings*. Somerville, MA: Cascadilla Press. 533-546.

Wilson, Colin (2006). Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science* 30: 945-982.

Yang, Zi-Xiang, He-Tong Guo, and Xiang-Dong Shi (1999). *Tianjinhua yindang* (*A record of the Tianjin dialect*). Shanghai: Shanghai Jiaoyu Chubanshe (Shanghai Education Press).

Ye, Yun and Cynthia M. Connine (1999). Processing spoken Chinese: The role of tone information. *Language and Cognitive Processes* 14: 609-630.

Yip, Michael C. W. (2001). Phonological priming in Cantonese spoken-word processing. *Psychologia* 44: 223-229.

Yip, Michael C. W., Po-Yee Leung, and Hsuan-Chih Chen (1998). Phonological similarity effects in Cantonese spoken-word processing. *Proceedings of ICSLP'98*, Vol.5: 2139-2142. Sydney, Australia.

Yu, Alan C. L. (2004). Explaining final obstruent voicing in Lezgian: Phonetics and history. *Language* 80.1: 73-97.

Yue-Hashimoto, Anne O. (1987). Tone sandhi across Chinese dialects. In Chinese Language Society of Hong Kong (ed.), Wang Li memorial volumes, English volume. Hong Kong: Joint Publishing Co. 445-474.

Zhang, Jie (2002). *The effects of duration and sonority on contour tone distribution.* New York: Routledge.

Zhang, Jie (2004). The role of contrast-specific and language-specific phonetics in contour tone distribution. In Hayes et al. (2004). 157-190.

Zhang, Jie (2007). A directional asymmetry in Chinese tone sandhi systems. *Journal of East Asian Linguistics* 16.4: 259-302.

Zhang, Jie and Yuwen Lai (2008). Phonological knowledge beyond the lexicon in Taiwanese double reduplication. In Yuchau E. Hsiao, Hui-Chuan Hsu, Lian-Hee Wee, and Dah-An Ho (eds.), *Interfaces in Chinese Phonology: Festschrift in Honor of Matthew Y. Chen on His 70th Birthday.* Academia Sinica, Taiwan. 183-222.

Zhang, Jie, Yuwen Lai, and Craig Sailor (2009a). Opacity, phonetics, and frequency in Taiwanese tone sandhi. In *Current issues in unity and diversity of languages: Collection of papers selected from the 18th International Congress of Linguists*. Linguistic Society of Korea. 3019-3038.

Zhang, Jie, Yuwen Lai, and Craig Sailor (2009b). Effects of phonetics and frequency on the productivity of Taiwanese tone sandhi. *Proceedings of the 43rd Annual Meeting of the Chicago Linguistic Society*, Vol. 1. 273-286.

Zhang, Ning (1997). The avoidance of the Third Tone Sandhi in Mandarin Chinese. *Journal of East Asian Linguistics* 6: 293-338.

Zue, Victor W. and Martha Laferriere (1979). Acoustic study of medial /t, d/ in American English. *Journal of the Acoustical Society of America* 66: 1039-1050.

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

Zuraw, Kie (2007). The role of phonetic knowledge in phonological patterning: corpus and survey evidence from Tagalog infixation. *Language* 83.2: 277-316.

Appendix: Additional Stimuli Information

This appendix provides additional information and entire word lists for the five word groups — AO-AO, *AO-AO, AO-AG, AG-AO, and AG-AG — used in the experiments.

I. AO-AO

For AO-AO words, we controlled both the frequency and the Mutual Information Score for the disyllables using the Feng Hua Yuan character and digram frequency corpus of Da (1998), which contains 4,718,131 characters and 4,159,927 digrams. All AO-AO disyllables fall within the raw frequency (raw number of occurrence) range of 31-62 and are relatively common words. The Mutual Information Score is calculated as $I(x,y) = \log_2 \frac{p(x,y)}{p(x)p(y)}$, where *p(x,y)* represents the digram frequency, and *p(x)* and *p(y)* represent the frequencies of the two characters respectively. A higher Mutual Information Score indicates a higher likelihood for the two characters to cooccur, and hence form real words. All AO-AO words fall within the range of 8-17 for the Mutual Information Score, which is a range that indicates that all these digrams are common words.[16]

---

[16] Da (1998) provides the following guidelines on how to interpret Mutual Information Scores: a Mutual Information Score greater than 3 indicates that the two words have a strong collocation; a Mutual Information Score less than 1 indicates that they are unlikely to be related; and a Mutual Information Score between 1 and 3 is in the gray area. For more information on Mutual Information Scores, see Oakes (1998).

Base tones	Chinese digram	IPA	Gloss	Digram freq.	Mutual info. score
3+1	鼓吹	ku tʂʰwei	‘to advocate’	45	9.11
	锦标	tɕin pjɑu	‘trophy’	44	9.51
	陕西	ʂan ɕi	Province name	39	8.79
	崭新	tʂan ɕin	‘brand new’	38	8.90
	脑筋	nɑu tɕin	‘brains’	34	9.68
	眼眶	jan kʰwɑŋ	‘eye socket’	34	9.09
	纺织	fɑŋ tʂɨ	‘to spin and weave’	33	11.45
	洒脱	sa tʰwɔ	‘free and easy’	31	8.97
3+2	沈阳	ʂən jɑŋ	City name	45	9.89
	谎言	hwɑŋ jɛn	‘lie’	41	8.95
	赌博	tu pwɔ	‘to gamble’	39	10.30
	补偿	pu tʂɑŋ	‘to compensate’	38	10.46
	礼仪	li ji	‘etiquette’	36	8.79
	减肥	tɕjɛn fei	‘to lose weight’	35	10.00
	野蛮	jɛ man	‘barbaric’	32	10.89
	饮食	jin ʂɨ	‘food intake’	32	9.50
3+3	展览	tʂan lan	‘exhibit’	60	8.90
	检讨	tɕjɛn tʰɑu	‘self-criticism’	62	9.20
	苦恼	kʰu nɑu	‘worried’	41	8.62
	拇指	mu tʂɨ	‘thumb’	34	10.50
	甲板	tɕja pan	‘ship deck’	41	8.50
	阻挡	tsu tɑŋ	‘to obstruct’	39	10.48
	洗碗	ɕi wan	‘to wash dishes’	34	9.15
	蚂蚁	ma ji	‘ant’	33	16.69
3+4	拯救	tʂəŋ tɕju	‘to rescue’	41	12.15
	粉碎	fən swei	‘to shatter’	39	10.32
	掩护	jɛn xu	‘to cover’	38	8.53
	忍耐	ɻən nai	‘to tolerate’	36	9.28
	巧妙	tɕʰjɑu mjɑu	‘ingenious’	35	9.56
	绑架	pɑŋ tɕja	‘to kidnap’	34	11.08
	饮料	jin ljɑu	‘drinks’	33	8.82
	尺寸	tʂʰɨ tsʰwən	‘size’	31	10.98

II. *AO-AO

Base tones	Chinese digram	IPA
3+1	尺仓	tʂʰɨ tsʰɑŋ
	宇章	ɥy tʂɑŋ
	写终	ɕjɛ tʂuŋ
	拢叉	luŋ tʂʰa
	榜中	pɑŋ tʂuŋ
	拇村	mu tsʰwən
	井披	tɕiəŋ pʰi
	减苍	tɕjan tsʰɑŋ
3+2	尺玩	tʂʰɨ wan
	宇零	ɥy liəŋ
	写拳	ɕjɛ tɕʰɥɛn
	拢宅	luŋ tʂai
	榜连	pɑŋ ljɛn
	拇挪	mu nwɔ
	井菩	tɕiəŋ pʰu
	减和	tɕjan xɤ
3+3	尺洒	tʂʰɨ sa
	宇览	ɥy lan
	写五	ɕjɛ wu
	拢法	luŋ fa
	榜洒	pɑŋ sa
	拇饮	mu jin
	井免	tɕiəŋ mjɛn
	减也	tɕjɛn jɛ
3+4	尺葬	tʂʰɨ tsɑŋ
	宇耀	ɥy jɑu
	写逆	ɕjɛ ni
	拢料	luŋ ljɑu
	榜报	pɑŋ pɑu
	拇葬	mu tsɑŋ
	井妙	tɕiəŋ mjɑu
	减会	tɕjan xwei

III. AO-AG

Base tones	Chinese digram	IPA
3+1	闯 shun	tʂʰwɑŋ ʂwən
	火 mu	xwɔ mu
	领 lan	liəŋ lan
	巧 re	tɕʰjɑu ɻɤ
	本 mai	pən mai
	苦 liang	kʰu ljɑŋ
	款 lang	kʰwan lɑŋ
	损 rao	swən ɻɑu
3+2	闯 te	tʂʰwɑŋ tʰɤ
	火 ka	xwɔ kʰa
	领 pie	liəŋ pʰjɛ
	巧 jiu	tɕʰjɑu tɕju
	本 mie	pən mjɛ
	苦 geng	kʰu kəŋ
	款 dui	kʰwan twei
	损 duan	swən twan
3+3	闯 zeng	tʂʰwɑŋ tsəŋ
	火 suan	xwɔ swan
	领 huai	liəŋ xwai
	巧 hang	tɕʰjɑu xɑŋ
	本 xun	pən ɕɥyn
	苦 heng	kʰu xəŋ
	款 pan	kʰwan pʰan
	损 cuo	swən tswɔ
3+4	闯 zhua	tʂʰwɑŋ tʂwa
	火 sen	xwɔ sən
	领 dei	liəŋ tei
	巧 shua	tɕʰjɑu ʂwa
	本 dei	pən tei
	苦 keng	kʰu kʰəŋ
	款 mang	kʰwan mɑŋ
	损 diu	swən tjəu

IV. AG-AO

Base tones	Chinese digram	IPA
3+1	ping 八	pʰiəŋ pa
	pan 昭	pʰan tʂɑu
	xia 凶	ɕja ɕjuŋ
	cang 黑	tsʰɑŋ xei
	zhui 咪	tʂwei mi
	chua 单	tʂʰwa tan
	run 邱	ɻwən tɕʰjou
	shuan 君	ʂwan ɕɥyn
3+2	ping 豪	pʰiəŋ xɑu
	pan 胡	pʰan xu
	xia 林	ɕja lin
	cang 原	tsʰɑŋ ɥyɛn
	zhui 伦	tʂwei lwən
	chua 林	tʂʰwa lin
	run 盘	ɻwən pʰan
	shuan 葵	ʂwan kʰwei
3+3	ping 马	pʰiəŋ ma
	pan 海	pʰan xai
	xia 哪	ɕja na
	cang 尺	tsʰɑŋ tʂʰɨ
	zhui 法	tʂwei fa
	chua 轨	tʂʰwa kwei
	run 起	ɻwən tɕʰi
	shuan 老	ʂwan lɑu
3+4	ping 套	pʰiəŋ tʰɑu
	pan 玉	pʰan ɥy
	xia 类	ɕja lei
	cang 率	tsʰɑŋ ly
	zhui 半	tʂwei pan
	chua 路	tʂʰwa lu
	run 费	ɻwən fei
	shuan 怒	ʂwan nu

V. AG-AG

Base tones	Chinese digram	IPA
3+1	ping shun	pʰiəŋ ʂwən
	pan mai	pʰan mai
	xia mei	ɕja mei
	cang re	tsʰɑŋ ɻɤ
	zhui mai	tʂwei mai
	chua liang	tʂʰwa ljɑŋ
	run lang	ɻwən lɑŋ
	shuan kuo	ʂwan kʰwɔ
3+2	ping te	pʰiəŋ tɤ
	pan ka	pʰan kʰa
	xia kong	ɕja kʰuŋ
	cang mie	tsʰɑŋ mjɛ
	zhui mie	tʂwei mjɛ
	chua geng	tʂʰwa gəŋ
	run dui	ɻwən twei
	shuan ta	ʂwan tʰa
3+3	ping zeng	pʰiəŋ tsəŋ
	pan seng	pʰan səŋ
	xia lue	ɕja lɥɛ
	cang xia	tsʰɑŋ ɕja
	zhui kuang	tʂwei kʰwɑŋ
	chua heng	tʂʰwa xəŋ
	run pan	ɻwən pʰan
	shuan sai	ʂwan sai
3+4	ping zhua	pʰiəŋ tʂwa
	pan sen	pʰan sən
	xia dei	ɕja tei
	cang shua	tsʰɑŋ ʂwa
	zhui dei	tʂwei tei
	chua keng	tʂʰwa kʰəŋ
	run mang	ɻwən mɑŋ
	shuan sen	ʂwan sən

VI. Fillers

All filler syllables are real syllables in Mandarin; half of disyllabic fillers were real words, and the other half were wug words. The tonal combinations of the fillers were chosen randomly and are given below.

σ_2 σ_1	Real fillers				Wug fillers			
	T1	T2	T3	T4	T1	T2	T3	T4
T1	7	4	5	7	3	3	0	6
T2	3	6	6	11	2	8	4	14
T3	2	4	2	6	2	1	1	3
T4	2	2	5	8	6	10	0	17