UR Underlearning of Mandarin Chinese Tone 3 Sandhi Words[1]

Boer Fu
Massachusetts Institute of Technology

1 Introduction

This paper addresses a long-standing question in the field of phonological learning: can speakers learn UNDERLYING REPRESENTATIONS (UR) that are different from their SURFACE REPRESENTATIONS (SR) in the absence of morphological alternation (see Yip 1996, McCarthy 2005, Rasin et al. 2020, Richter 2021)? In languages that are rich in morphology, alternations offer help for speakers to learn non-identical UR-SR mappings. For example, in German final devoicing, a word-final voiced consonant in the UR is mapped to a voiceless consonant in the SR. Speakers of German have no trouble learning the non-identical UR-SR mapping, since they can rely on morphological alternations between singular and plural nouns (e.g. [ta:**k**] 'day.SG' and [ta:**g**ə] 'day.PL').

What happens when there is no morphological alternation? Can speakers still learn non-identical UR-SR mappings? I set out to address this question by exploring Mandarin Chinese. The language's shortage of inflectional morphology makes it an interesting test case for UR learning. I focus on the famous process of TONE 3 SANDHI, which results in a tonal mismatch between the UR /T3 T3/ and the SR [T2 T3]. Can Mandarin speakers learn the non-identical tonal UR-SR mapping in the absence of morphological alternations?

Previous literature on the phonological learning of tone 3 sandhi has mainly focused on how productive the process is, examining whether Mandarin speakers can consistently apply the sandhi rule in nonce words (see J. Zhang & Lai 2010, C. Zhang & Peng 2013). There has not been much work to investigate whether Mandarin speakers can learn the non-identical UR. In fact, it is usually assumed that the speaker URs of sandhi words are identical to their dictionary entry, *T3 T3*.

I challenge the assumption that all Mandarin disyllabic words that are purported to have undergone tone 3 sandhi are consistently learned with a /T3 T3/ UR by speakers. I show that many sandhi words are in fact stored with a surface-identical UR /T2 T3/ by some Mandarin speakers, contrary to the dictionary entry. This is demonstrated via a novel AABB reduplication diagnostic that I have designed. I conducted a pilot survey of 6 Mandarin speakers, making use of the AABB diagnostic. The results indicate that many sandhi words are learned with identical

[1] Many thanks to Adam Albright, Fulang Chen, Edward Flemming, Filipe Kobayashi, Michael Kenstowicz, Patrick Niedzielski, Donca Steriade, Stanislao Zompì for valuable discussions. The project also benefited from comments at the MIT reading groups Morphun and Phonology Circle, as well as audience feedback at CLS 58. All remaining mistakes are my own.

UR-SR mapping, or UNDERLEARNED. The term refers to the underapplication of tone 3 sandhi in the speaker's grammar. I argue that underlearning only takes place when the target sandhi word is compositionally opaque, because speakers cannot establish morphological alternations between the target word and related words elsewhere in the lexicon that might reveal the non-identical UR. Therefore, I conclude that morphological alternation is crucial for the speaker to learn non-identical UR-SR mappings.

The rest of the paper is organized as follows. Section 2 introduces Mandarin tone 3 sandhi and the resulting incomplete neutralization. Section 3 discusses why the UR learning of sandhi words is a problem for both learners and linguists alike. In Section 4, I delineate the AABB reduplication diagnostic. Results from a pilot survey are reported in Section 5. Section 6 and 7 discuss the implication in phonological learning and Chinese orthography. Section 8 concludes. In the paper, *italics* are used for dictionary entries, and slashes denote true UR.

2 Tone 3 sandhi and incomplete neutralization

Tone 3 sandhi is perhaps the most well-studied phonological phenomenon in Mandarin Chinese. It involves two tones: tone 2, or the rising tone, and tone 3, the low dipping tone that is often accompanied with creaky voice. The sandhi rule states that whenever a tone 3 is followed by another tone 3, it surfaces as a tone 2. Formally, the rule is written as T3 → T2 / __ T3.

The tonal process leads to a case of neutralization in disyllabic words. Since all words with /T3 T3/ as their UR surface as [T2 T3], they become indistinguishable from words that have /T2 T3/ in UR. (1) shows an extreme case of neutralization. The minimal pair (1a) /wu3 pi3/ 'five stroke input system' and (1b) /wu2 pi3/ 'unparalleled' are neutralized into [wu2 pi3] after the application of tone 3 sandhi.

(1) *Neutralization in disyllabic words*

a.	/wu3 pi3/	[wu2 pi3]	五笔	'five stroke input system'
b.	/wu2 pi3/	[wu2 pi3]	无比	'unparalleled'

Mandarin tone 3 sandhi is reported to be a case of INCOMPLETE NEUTRALIZATION (see W. Wang & K. Li 1967, M. Lin et al. 1980, Liu 2013). The term is used to describe a phenomenon in which a contrast is maintained in production but lost in perception (see Port et al. 1981). Some Mandarin speakers are shown to be able to produce a small acoustic difference between /T3 T3/ and /T2 T3/ words, but it has been demonstrated that speakers cannot reliably perceive the difference. The acoustic difference is manifested in the pitch range of the rising tone on the initial syllable. Specifically, the rising tone [T2] derived from an underlying /T3/ begins at a lower pitch and ends at a lower pitch, compared to the rising tone that is underlyingly /T2/ (Zee 1980, Liu 2013, Yuan & Chen 2014). Mandarin listeners

fail to perform better than chance at identifying the correct UR (W. Wang & K. Li 1967, M. Lin et al. 1980, Liu 2013).

3 UR learning problem

The incomplete neutralization of tone 3 sandhi leads to a learning problem. How do children acquiring Mandarin learn two distinct URs for a pair of lexical items that share the same SR? In fact, can they learn non-identical UR at all?

As we have seen, however little pitch difference there is between /T3 T3/ and /T2 T3/ words, the acoustic information is of no use to Mandarin learners, since it cannot be reliably perceived. As to morphological alternations, Mandarin does not have much to offer. There is no productive word-internal morphological process, segmental or suprasegmental, that triggers change in the SR of sandhi words.

Nevertheless, there is morphological alternation in Mandarin, in the form of compounding. If a sandhi word is composed of two morphemes that appear elsewhere in the lexicon, the learner has a chance to observe the tonal UR of the individual morphemes. This can be said for the neutralized pair in (1). (1a) [wu2 pi3] 'five stroke input system' is made up of the individual morphemes [wu3] 'five' and [pi3] 'stroke'. Mandarin learners can establish a [wu2~wu3] alternation for the morpheme 'five' and infer that the UR for 'five stroke' is /wu3 pi3/. (1b) [wu2 pi3] 'unparalleled' is composed of [wu2] 'NEG' and [pi3] 'to compare'. The learner can see that the negation morpheme appears as [wu2] everywhere, concluding that 'unparalleled' is /wu2 pi3/ in UR.

The above examples show that COMPOSITIONAL TRANSPARENCY in compounds can help Mandarin learners acquire non-identical UR-SR mappings by providing morphological alternations. We should expect a learning asymmetry between compositionally transparent sandhi words and compositionally opaque sandhi words. If the target sandhi word is compositionally transparent, Mandarin learners ought to have no trouble learning a non-identical /T3 T3/ UR. However, if the target sandhi word is compositionally opaque, then not all learners will be able to posit a /T3 T3/ UR in accordance with the dictionary entry *T3 T3*. It is expected that a number of learners will store a /T2 T3/ UR, identical to the SR.

In order to test the hypothesis on the learning asymmetry between words that are compositionally transparent and opaque, we need to find out, as a first step, the tonal UR in the mental lexicon of adult Mandarin speakers. As phonologists, we have no opportunity to see the tonal UR of words that are purported to have undergone tone 3 sandhi, for all the reasons listed above that makes the same task difficult for the Mandarin learners. To tackle this challenge, I have designed a novel AABB reduplication diagnostic, which can reveal the tonal UR Mandarin speakers have posited for words with [T2 T3] SR.

4 AABB reduplication diagnostic

AABB reduplication is a semi-productive morphological process in Mandarin. It takes a disyllabic AB base, reduplicates each syllable, and forms a quadrisyllabic AABB word. The AB base can be an adjective, a verb, or a noun. The resulting AABB form expresses quantification of the AB base (N. Zhang 2015), as in (2).

(2) *AABB reduplication*

a.	[kan1 tɕiŋ4]	'clean'	e.	[kan1 kan1 tɕiŋ4 tɕiŋ4]	'spotless'
b.	[jow2 ɥy4]	'to hesitate'	f.	[jow2 jow2 ɥy4 ɥy4]	'indecisive'
c.	[ʂɨ2 kʰə4]	'moment'	g.	[ʂɨ2 ʂɨ2 kʰə4 kʰə4]	'every moment'

Only adjective-to-adjective AABB reduplication in (2a) is relatively productive. As a preview, the AABB reduplication diagnostic for tonal UR will make use of the unproductive noun-to-noun AABB reduplication, as seen in (2c).

4.1 AABB reduplication as a diagnostic for tonal UR

The tonal sequence in an AABB reduplicated form can help disambiguate between /T3 T3/ bases and /T2 T3/ bases, as exemplified by (3).

(3) *AABB reduplication with /T3 T3/ vs. /T2 T3/*

a.	/two3 ʂan3/	'to evade'	c.	[two2 **two2** ʂan2 ʂan3]	'evasive'
			d.	[two2 **two3** ʂan2 ʂan3]	
b.	/xʊŋ2 xwo3/	'flourishing'	e.	[xʊŋ2 **xʊŋ2** xwo2 xwo3]	'very
			f.	*[xʊŋ2 **xʊŋ3** xwo2 xwo3]	flourishing'

(3a) /two3 ʂan3/ 'to evade' has two variant AABB forms in (3c&d) (Feng 2003). The second syllable [**two2~3**] (in **boldface**) in the quadrisyllabic adjective is the site of tonal variation. (3b) /xʊŋ2 xwo3/ 'flourishing' only has one AABB reduplicated form, which is (3e) [xʊŋ2 **xʊŋ2** xwo2 xwo3]. The alternative tonal pattern of (3f), where the second syllable surfaces as [xʊŋ3], is illicit.

A /T3 T3/ base has two variant AABB forms, because the reduplicated form can either be derived from the SR or the UR of the base. As seen in Table 1. The SR route takes the base [two2 ʂan3] 'to evade' to [two2 **two2** ʂan2 ʂan3]. The UR route in Table 2, on the other hand, takes /two3 ʂan3/ and changes it into [two2 **two3** ʂan2 ʂan3]. We can also see why a /T2 T3/ base can only have one AABB form. Its SR and UR are identical, leading to the same reduplicated form.

AB Base SR	[two2 ʂan3] 'to evade'	[xʊŋ2 xwo3] 'flourishing'
AABB Reduplication	two2 two2 ʂan3 ʂan3	xʊŋ2 xʊŋ2 xwo3 xwo3
Tone 3 Sandhi	two2 two2 ʂan2 ʂan3	xʊŋ2 xʊŋ2 xwo2 xwo3
AABB reduplicated SR	[two2 **two2** ʂan2 ʂan3]	[xʊŋ2 **xʊŋ2** xwo2 xwo3]

Table 1: AABB reduplication via SR

AB Base UR	/two3 ʂan3/ ‘to evade’	/xʊŋ2 xwo3/ ‘flourishing’
AABB Reduplication	two3 two3 ʂan3 ʂan3	xʊŋ2 xʊŋ2 xwo3 xwo3
Tone 3 Sandhi	two2 two3 ʂan2 ʂan3	xʊŋ2 xʊŋ2 xwo2 xwo3
AABB reduplicated SR	[two2 **two3** ʂan2 ʂan3]	[xʊŋ2 **xʊŋ2** xwo2 xwo3]

Table 2: AABB reduplication via UR

The different behavior of /T3 T3/ and /T2 T3/ base in AABB reduplication means that we can utilize the morphological process as a diagnostic for tonal UR of the AB base. Given that the AABB reduplication is a semi-productive process, with the appropriate instruction, Mandarin speakers can apply it to any existing disyllabic word, forming a novel AABB reduplicated form. If the AB base word is one that is purported to have undergone tone 3 sandhi, what the speaker accepts as the AABB form can inform us of how they store the lexical item in UR. Specifically, if the sandhi word is truly stored as /T3 T3/ in the UR, then [T2 **T2** T2 T3] and [T2 **T3** T2 T3] are both possible as the reduplicated form. However, if the sandhi word is learned with a /T2 T3/ UR, then only [T2 **T2** T2 T3] is licit as the AABB form. The alternative AABB form [T2 **T3** T2 T3] will be rejected by the speaker, as schematized in Table 3. The diagnostic is stated in (4).

	AABB: [T2 **T2** T2 T3]	AABB: [T2 **T3** T2 T3]
Base UR: /T3 T3/	*Accept (via SR)*	*Accept (via UR)*
Base UR: /T2 T3/	*Accept (via UR or SR)*	***Reject!***

Table 3: AABB form judgement table

(4) **The AABB reduplication diagnostic**: For a word with surface [T2 T3], if the speaker rejects [T2 **T3** T2 T3] as an AABB reduplicated form, then they have posited a /T2 T3/ UR for the lexical item. Otherwise, the UR is /T3 T3/.

5 Speaker judgement survey

I conducted a survey on Mandarin speakers’ judgement on the tones of AABB reduplicated forms. The survey serves two purposes. Firstly, it is a proof-of-concept trial run for the novel AABB reduplication diagnostic. The aim is to find out if the diagnostic can reliably reveal the tonal UR of sandhi words. Secondly, it is designed to detect if there is an asymmetry in the phonological learning of sandhi words, depending on the compositional transparency of the target word.

5.1 Methods

The speaker judgement survey makes use of the AABB reduplication diagnostic. Mandarin speakers are given a list of disyllabic words with surface [T2 T3] and asked to judge the tones of the AABB reduplicated form.

The survey is framed as a language game. At the beginning of the survey, the speakers are given the instruction that there is a novel AABB reduplication

process that takes in any noun of the shape AB and turns it into an AABB noun, which now has the meaning ‘every AB’. The speakers are also provided with examples sentences. In addition to regular nouns, there are place names in the survey. They are given a slightly different novel meaning of ‘everywhere in AB’.

For each word, the speaker is prompted with a question: “The word ____, after AABB reduplication, is pronounced as?” They are asked to choose between three options: (a) [T2 **T2** T2 T3], (b) [T2 **T3** T2 T3], and (c) Both forms are fine. The question prompt is written in Simplified Chinese. Each option has the AABB form written in PINYIN, the romanization script used in China, which transcribes lexical tones with diacritics. The pinyin transcription is also accompanied by the AABB form written in Chinese characters. (5) is a sample question, with English translation in italics.

(5) “蚂蚁”一词， 组成 AABB 的叠词后，读音是：
The word “ant”, after AABB reduplication, is pronounced as?
a. má má yí yǐ 蚂蚂蚁蚁 [ma2 **ma2** ji2 ji3]
b. má mǎ yí yǐ 蚂蚂蚁蚁 [ma2 **ma3** ji2 ji3]
c. 两种形式都可以。 *Both forms are fine.*

If the surveyed speaker chooses option (a), it suggests that they have stored the word for ‘ant’ as /ma2 ji3/ in their mental lexicon. Option (b) maps to a /ma3 ji3/ UR. Option (c) “Both forms are fine” also points to a speaker UR of /ma3 ji3/. Recall that the AABB reduplication diagnostic states that it is only a rejection of the form [T2 **T3** T2 T3] that counts as evidence for a /T2 T3/ UR. Given the availability of option (c) “Both forms are fine”, any speaker who chooses option (a) [ma2 **ma2** ji2 ji3] does so as a rejection to the alternative (b) [ma2 **ma3** ji2 ji3]. This means that their stored UR for the lexical item is truly /ma2 ji3/.

If a dictionary *T3 T3* word is shown to have been learned as /T2 T3/ by a speaker, a case of underlearning is identified. The term “underlearning” is named after underapplication, not underperformance. To learn a *T3 T3* target word as /T2 T3/, the speaker is acquiring a grammar in which the sandhi rule is underapplied. It is important to note that whenever a case of underlearning is identified, the speaker is never in the wrong. The /T2 T3/ UR they have acquired is a valid and robust representation of the lexical item.

5.2 Participants
6 native speakers of Mandarin answered the survey, presented on Google Forms.

5.3 Materials
There are 40 disyllabic words with [T2 T3] SR included in this survey. 26 have the citation tones of *T3 T3*. 14 are labeled as *T2 T3*. Among the disyllabic words

across the two tonal categories, some are compositionally transparent, others are opaque. The 14 non-sandhi *T2 T3* words are included, in order to make sure that selecting the AABB tonal SR is not a trivial task for the speakers surveyed. The first 2 lexical items the speakers see in the survey are the compositionally transparent *ɥy3 san3* 'umbrella' and *jaŋ2 san3* 'parasol', serving as a baseline.

5.4 Results and discussion

The speakers surveyed are shown to have robust judgement on the tones of the AABB reduplicated forms, even though they are not attested. This can be seen in their responses to the first two AB bases in the survey, *ɥy3 san3* 'umbrella' and *jaŋ2 san3* 'parasol'. These two words are both compositionally transparent, therefore the Mandarin speaker URs for them are expected to align with the dictionary entries. This is indeed what we observe, as shown in Table 4. The AABB reduplication diagnostic is a reliable measure to gauge speaker UR.

Dictionary tones	[T2 **T2** T2 T3]	[T2 **T3** T2 T3]	Both are fine	Speaker UR
ɥy3 san3 'umbrella'	0% speakers	33.3% speakers	66.7% speakers	/T3 T3/
jaŋ2 san3 'parasol'	100% speakers	0% speakers	0% speakers	/T2 T3/

Table 4: Speakers have robust judgement on AABB tones

Among the 26 sandhi words listed as *T3 T3* in the dictionary, 9 have been underlearned as /T2 T3/ by one or more speakers surveyed. The 9 sandhi words subject to underlearning can be described as (i) compositionally opaque and (ii) containing an initial syllable that has an available tone 2 alternative. These two aspects of UR underlearning are discussed in 5.4.1 and 5.4.2 respectively.

5.4.1 Compositional transparency

Recall that the hypothesis the survey aims to test is a learning asymmetry between sandhi words that are compositionally transparent and those that are opaque. It is largely borne out. 8 out of the 9 sandhi words that have been underlearned by a few speakers are opaque. The words in Table 5 all contain an initial syllable that does not contribute to the overall meaning of the word in a transparent manner.

Sandhi word (dictionary tone)			English	UR: /T3 T3/	UR: /T2 T3/
a.	*law3* 老 'old'	*ʂu3* 鼠 'rodent'	'rat'	33.3% speakers	66.7% speakers
b.	*ma3* 蚂 ?	*ji3* 蚁 'ant, termite'	'ant'	50% speakers	50% speakers
c.	*ma3* 马 'horse'	*tʰʊŋ3* 桶 'bucket'	'toilet'	66.7% speakers	33.3% speakers
d.	*mɤŋ3* 猛 'ferocious'	*ma3* 犸 ?	'mammoth'	66.7% speakers	33.3% speakers

Table 5: Compositionally opaque words are prone to /T2 T3/ underlearning.

In Table 5, (a) *law3 ʂu3* ‘rat’ is an animal name. It is ostensibly composed of *law3* ‘old’ and *ʂu3* ‘rodent’. However, there is nothing ‘old’ about rats. The initial syllable *law3* is not contributing transparently to the overall meaning of the word, which makes it prone to underlearning. In (b) *ma3 ji3* ‘ant’, the initial syllable *ma3* does not have a meaning other than ‘ant’. The word formation of *ma3 ji3* is akin to that of *cranberry* in English, where *berry* clearly denotes a category, while *cran* does not have an independent meaning. The syllable *ma3* appears again in (c) *ma3 tʰʊŋ3* ‘toilet’, albeit transcribed with a different character. This time, it is written with the famous ‘horse’ character. If the meanings of the character components of *ma3 tʰʊŋ3* were used for word composition, one would arrive at ‘horse bucket’, not ‘toilet’. The comical result could be due to *ma3 tʰʊŋ3* being a historical compound, where the etymology is unclear. For Mandarin learners, the initial syllable in ‘toilet’ does not have the meaning of ‘horse’. Therefore, the surface [ma2] does not necessarily map onto /ma3/, resulting in underlearning. (d) *mɤŋ3 ma3* ‘mammoth’ is a loanword, which is naturally compositionally opaque.

In contrast, the compositionally transparent sandhi words in Table 6 are never underlearned. (a) *ɥy3 san3* ‘umbrella’ begins with the morpheme *ɥy3* ‘rain’, which transparently points to the context in which the physical object is used. (b) *tʰu3 ɻaŋ3* ‘soil’ is a coordinative compound made out of two nouns describing very similar concepts, *tʰu3* ‘earth’ and *ɻaŋ3* ‘soft soil’. In (c) *ʂwej3 mu3* ‘jellyfish’ is an animal that dwells in *ʂwej3* ‘water’. As to (d) *li3 xaj3* ‘Caspian Sea’, one only need to refer to a map to see that it is a landlocked sea ‘inside’ a continent.

Sandhi word (dictionary tone)			English	UR: /T3 T3/	UR: /T2 T3/
a.	*ɥy3* 雨 ‘rain’	*san3* 伞 ‘umbrella’	‘umbrella’	100% speakers	0% speakers
b.	*tʰu3* 土 ‘earth’	*ɻaŋ3* 壤 ‘soft soil’	‘soil’	100% speakers	0% speakers
c.	*ʂwej3* 水 ‘water’	*mu3* 母 ‘mother’	‘jellyfish’	100% speakers	0% speakers
d.	*li3* 里 ‘inside’	*xaj3* 海 ‘sea’	‘Caspian Sea’	100% speakers	0% speakers

Table 6: Compositionally transparent words are not underlearned.

5.4.2 Availability of tone 2 alternative

Sandhi word (dictionary tone)			English	UR: /T3 T3/	UR: /T2 T3/
a.	*kan3* 橄 ?	*lan3* 榄 ?	‘olive’	100% speakers	0% speakers
b.	*kow3* 枸 ?	*tɕʰi3* 杞 ?	‘goji berry’	100% speakers	0% speakers

Table 7: Exceptional opaque words that are not underlearned.

Unexpectedly, a few sandhi words that are compositionally opaque are not underlearned by any of the surveyed speakers, as shown in Table 7. (a) *kan3 lan3* 'olive' and (b) *kow3 tɕʰi3* 'goji berry' are monomorphemic plant names.

I argue that the compositionally opaque sandhi words in Table 7 all lack a crucial ingredient to trigger underlearning — an available tone 2 alternative for the initial syllable. In other words, their initial syllables have a TONAL GAP at tone 2. A Mandarin syllable is said to have a tonal gap when one of the 4 available lexical tonal configurations is unattested in the lexicon. This can be observed in the comparison between *ma*, the linguists' favorite syllable for Mandarin tone demonstrations, and *pʰan*. In Table 8, *ma* enjoys the full paradigm, appearing in tone 1, 2, 3, and 4, as 'mother', 'hemp', 'horse', and 'scold'. In contrast, *pʰan* can only appear in tone 1, 2, and 4. There is no **pʰan3* anywhere in the lexicon.

Tone 1	Tone 2	Tone 3	Tone 4
ma1'mother'	ma2 'hemp'	ma3 'horse'	ma4 'scold'
pʰan1 'to climb'	pʰan2 'plate'	*pʰan3 **Tonal Gap**	pʰan4 'to judge'

Table 8: Tonal gap

Tonal gap offers a clue to speakers' divergent learning pattern for the two sets of compositionally opaque sandhi words in Table 5 and 7. In a frequently underlearned opaque sandhi word, the initial syllable, which is the site of tonal UR ambiguity, does not have a tonal gap at tone 2. This is shown in Table 9. In other words, the underlearned sandhi word's initial syllable has an available tone 2 alternative. In contrast, the initial syllable in an opaque sandhi word that is never underlearned has a tonal gap at tone 2, or that it does not have an available tone 2 alternative, as seen in Table 10.

Sandhi Word (Dictionary Tone)		Initial Syllable		Rate of /T2 T3/ Underlearning
		Tone 2 alternative	Tone 3 alternative	
a.	*law3 ʂu3* 'rat'	*law2* 'labor'	*law3* 'old'	66.7% speakers
b.	*ma3 ji3* 'ant'	*ma2* 'hemp'	*ma3* 'horse'	50% speakers

Table 9: Available tone 2 alternative for initial syllable leads to underlearning.

Sandhi Word (Dictionary Tone)		Initial Syllable		Rate of /T2 T3/ Underlearning
		Tone 2 alternative	Tone 3 alternative	
a.	*kan3 lan3* 'olive'	**kan2* **Tonal Gap**	*kan3* 'to dare'	66.7% speakers
b.	*kow3 tɕʰi3* 'goji berry'	**kow2* **Tonal Gap**	*kow3* 'dog'	50% speakers

Table 10: Unavailable tone 2 alternative for initial syllable blocks underlearning.

The Mandarin learner hears the initial syllable of a sandhi word in its surface [T2]. Their task is to map it onto an underlying tone. When the syllable can appear in tone 2 or tone 3 in the wider lexicon, the learner needs to select either /T2/ or /T3/ as its UR. Sometimes they make a decision that disagrees with the dictionary entry, which results in underlearning of the sandhi word. However, when the initial syllable has a tonal gap at tone 2, as is the case for the opaque sandhi words in Table 10. the Mandarin learner knows that the syllable cannot possibly be /T2/ in UR. Therefore, they always map the initial syllable to /T3/, landing on a /T3 T3/ UR for the entire word. No underlearning takes place.

I have demonstrated that a Mandarin sandhi word is prone to underlearning if (i) it is compositionally opaque; (ii) its initial syllable has an available tone 2 alternative. On the flip side, a sandhi word is almost never underlearned if it is (iii) compositionally transparent; or (iv) compositionally opaque but has an initial syllable with a tonal gap at tone 2.

5.4.3 Overlearning

In the previous two subsections, I have discussed the factors contributing to underlearning of sandhi words transcribed as *T3 T3* in the dictionary. But what about overlearning? Do words listed as *T2 T3* ever get learned as having a /T3 T3/ UR, in which the Mandarin learner's mental lexicon includes an overapplication of tone 3 sandhi for these words? The answer is almost never, with exceptions. 14 of the 40 words surveyed are listed as *T2 T3* in the dictionary. 12 are learned consistently by the surveyed speakers as having /T2 T3/ in UR. The 2 non-sandhi items in Table 11 show some degree of overlearning.

Non-Sandhi Word (Dictionary Tone)			English	UR: /T3 T3/	UR: /T2 T3/
a.	*li2* 厘 'ancient unit'	*mi3* 米 'meter'	'centimeter'	33.3% speakers	66.7% speakers
b.	*xu2* 湖 'lake'	*pej3* 北 'north'	'Hubei'	16.7% speakers	83.3% speakers

Table 11: Non-sandhi T2 T3 words overlearned as /T3 T3/

These results are unexpected, because we usually expect a learning bias towards faithful mapping between the UR and the SR, as a result of lexicon optimization (see Prince & Smolensky 1993). When a Mandarin-learning child hears a word with [T2 T3] SR, they are expected to map it to an identical UR /T2 T3/, unless given evidence that suggests otherwise. As we have seen in the discussion of sandhi words underlearning, evidence can come in the form of transparent morphological alternation involving the initial syllable or a tonal gap at tone 2. The 2 non-sandhi words in Table 11 do not provide such evidence. They include an initial syllable that is transcribed as tone 2 in the dictionary, therefore they have no reason to participate in morphological alternation with any lexical item that contains the initial syllable in tone 3, at least prescriptively speaking.

It is possible that some learners have reanalyzed the initial syllable as a different morpheme. For example, *li2* ‘ancient unit’ in (a) *li2 mi3* ‘centimeter’ could have been reanalyzed as *li3* ‘inside’ or *li3* ‘mile’. Similarly, *xu2* ‘lake’ in (b) *xu2 pej3* ‘Hubei’ could be taken to mean *xu3* ‘tiger’ or *xu3* ‘amber’.

An alternative explanation for the unexpected overlearning is a confound that results from the question sequence in the survey. Both items in Table 11 are located at the very end of the survey (item number 39 and 40). A certain level of speaker fatigue is expected. For future studies, a randomized stimulus sequence for each participant is necessary to identify real cases of overlearning.

6 Implications for phonological learning

The investigation on the UR learning of Mandarin sandhi words aims to answer the research question: can speakers learn non-identical UR-SR mapping in the absence of morphological alternation? The answer appears to be no. Sandhi words listed as *T3 T3* in the dictionary are a case of non-identical UR-SR mapping, at least prescriptively speaking. In Mandarin, morphological alternation comes in the form of transparent compounding. If a sandhi word is compositionally transparent, it offers evidence of morphological alternation with individual morphemes that appear elsewhere in the lexicon. However, if a sandhi word is compositionally opaque, its individual syllables are not recurring morphemes in the lexicon, therefore no morphological alternation can be found.

The results from the AABB reduplication survey show that the morphological alternations provided by compositionally transparent sandhi words make the UR learning task straightforward for the Mandarin speaker. For these words, the speakers are almost guaranteed to acquire the non-identical mapping between /T3 T3/ and [T2 T3] consistently. When there is a true absence of morphological alternation, as in compositionally opaque sandhi words, Mandarin speakers do not always acquire the prescriptive non-identical UR-SR mapping. Instead, they frequently acquire identical UR-SR mappings for these opaque sandhi words.

The widespread speaker variation we observe in the UR learning of opaque sandhi words raises the possibility that compositional transparency, and therefore what count as morphology alternations, might differ from speaker to speaker. In the borrowed animal name *mɤŋ3 ma3* ‘mammoth’, the initial syllable *mɤŋ3* might have been reanalyzed as the morpheme ‘ferocious’ by some Mandarin learners. This can account for the 66.7% of surveyed speakers who have acquired the dictionary UR /mɤŋ3 ma3/, as opposed to underlearning the lexical item. It can be said that these speakers have identified a morphological alternation between [mɤŋ2 ma3] ‘mammoth’ and [mɤŋ3] ‘ferocious’, that the other 33.3% have not. Other opaque sandhi words might be subject to the same type of reanalysis, or learner-initiated search for morphological alternations. This might account for

why there is no opaque sandhi word in the survey that is underlearned by 100% of the speakers. For example, *ma3 ji3* ‘ant’ might be reanalyzed as *ma3* ‘horse’ *ji3* ‘ant’, given that both ants and horses are animals that carry weight.

The results from the survey on Mandarin speakers’ tonal UR show that the phonological learning of non-identical UR-SR mapping in the absence of morphological alternation is, if not impossible, difficult. In addition, it is the child language learner who decides on what counts as a morphological alternation, not the lexicographer or the adult literate speaker. This is likely true for the UR learning of other languages as well. For a child acquiring German final devoicing, the task of determining what counts as morphological alternation is a trivial one. It is evidently not trivial for a child acquiring Mandarin tonal UR.

In order to confirm the key role of speaker-specific morphological alternations in UR learning, some concrete evidence of such speaker variation is needed. A future study will benefit from a lexical association experiment, which can firmly link a speaker’s UR to their morphological representation of sandhi words.

7 Implication for Chinese orthography

Throughout the discussion above, I have assumed that phonological learning generally precedes the learning of the writing system. Therefore, Chinese orthography is not expected to affect the learning of Mandarin tonal URs. In this section, I consider the possibility that Mandarin speaker’s URs for sandhi words are influenced by the written Chinese characters.

Help from written Chinese is only expected to be required in sandhi words that are not considered to be compositionally transparent by all speakers. In such a scenario, the orthography might be able to aid tonal UR learning directly or indirectly. If the target word contains common Chinese characters, the learner can directly identify the tonal UR. *ma3 tʰʊŋ3* 马桶 ‘toilet’ is such an example. The first syllable is written with a common character 马. In isolation, 马 means ‘horse’ and is pronounced as [ma3]. The child learner can make use of their knowledge of the character to posit /ma3 tʰʊŋ3/ as the UR.

Sometimes, the target word is written with specialized characters that rarely appear in the environment. Nevertheless, many of these specialized characters often contain a common phonetic component, which offers an indirect clue for the learner. Most Chinese characters are made up of two parts: a SEMANTIC COMPONENT and a PHONETIC COMPONENT. The semantic component, or the radical, indicates the semantic category of the character. The phonetic component provides some rough clues to its pronunciation. For example, the character 橄 in *kan3 lan3* 橄榄 ‘olive’ has a phonetic component 敢, which is a common

character pronounced as [kan3]. A Mandarin-speaking child encountering the word 橄榄 'olive' for the first time, might be able to infer that 橄 ought to be pronounced as 敢 [kan3].

However, the pronunciation of the phonetic component has been subject to sound change over time. A Mandarin speaker might find the phonetic component as helpful as the English spelling *ough*, which is used in *rough, ought, through, drought*, each with a different pronunciation. For instance, the character 枸 in *kow3 tɕʰi3* 枸杞 'goji berry' has the phonetic component 句, which is pronounced as [tɕy4] on its own. It shows up as [tɕy1] in 驹, [ɕy4] in 煦, [kow3] in 狗, [kow4] in 够, to name a few common characters. The tone varies, as well as the segments of the syllable. A child learning the word 枸杞 'goji berry' will not be able to pinpoint the UR of 枸 with accuracy. Despite the one-to-many mapping between the phonetic component 句 and its pronunciation in related characters, the child might still be able to locate 狗 [kow3] as the indicator to the sound of 枸. A probabilistic learning process is likely to be involved, in which the learner searches through all the Chinese characters sharing the same phonetic component in their inventory, making a prediction on the most likely pronunciation of the target character. A computational model of the character learning process can be an interesting extension of this project.

If it is indeed the case that knowledge on Chinese orthography has helped Mandarin learners in acquiring tonal URs, then we should expect an age of acquisition effect. Specifically, sandhi words that are acquired before literacy ought to not be affected by orthography, therefore displaying a higher rate of underlearning. In contrast, sandhi words acquired after the child has learned the writing system are expected to show a lower rate of underlearning. For any word, its age of acquisition surely differs from speaker to speaker, but its lexical frequency might serve as a good enough approximation of the age of acquisition within a population. Words with higher lexical frequency are more likely to be acquired at an earlier age. This is yet another avenue for future research.

Regardless of the role that Chinese orthography plays in phonological learning, the results from the speaker survey challenge the long-held view that Chinese characters are logographic, namely that it transcribes meanings, as opposed to sounds in an alphabetical writing system. Instead, there are multiple examples of meaningless characters in the survey. There are characters that only appear in certain animal and plant names like 蚂 in 蚂蚁 'ant' and 橄 in 橄榄 'olive', with no independent meaning on their own. Even characters that enjoy a high token frequency in the written language can appear meaningless in a compositionally opaque word. The best example is the character 马. On its own, it means 'horse', but when it is found inside 马桶 'toilet', it has no meaning. In the synchronic

Mandarin grammar, its only function is to transcribe the sound of [ma2] in [ma2 tʰʊŋ] ‘toilet’. The Chinese writing system includes many more characters like 马, used to transcribe sound, not meaning.

8 Conclusion

This project investigates the tonal UR learning in Mandarin words that are purported to have undergone tone 3 sandhi. It poses two questions: what UR have the Mandarin speakers learned for sandhi words? What does Mandarin speakers’ behavior in UR learning tell us about phonological learning in general? To address the first question, I have designed a novel AABB reduplication diagnostic, which is shown to be an effective tool at gauging individual speaker’s UR for sandhi words. For the second question, I incorporated the diagnostic in a speaker judgement survey, in order to see if morphological alternations, found only in compositionally transparent compounds, are a requisite for establishing non-identical UR-SR mappings. A key finding of the survey is that Mandarin speakers frequently learn identical UR-SR mappings for compositionally opaque sandhi words that have non-identical UR-SR mappings prescriptively, in a process termed underlearning. A few compositionally opaque sandhi words that are never underlearned appear to be exceptions at first sight, but closer inspection shows that tonal gaps can serve as a supplementary tool for the learning of non-identical UR-SR mapping. The survey results confirm the hypothesis that there is a learning asymmetry between compositionally transparent and opaque sandhi words. They also strongly support the claim that morphological alternations are crucial to the learning of non-identical UR-SR mappings. At the same time, the between-speaker variation in UR learning points to the possibility that morphological alternations are speaker-specific. For a morphological alternation to be useful, it needs to be identified by the child language learner as such.

References

Feng, G. B. 2003. Lexical category specific constraints: Mandarin verb versus adjective reduplication. *USC Working Papers in Linguistics* 1. 1-12.

Lin, M., L. Lin, G. Xia, & Y. Cao. 1980. Putonghua erzici biandiao de shiyan yanjiu. [Experimental study of Mandarin disyllabic word tone sandhi.] *Zhongguo Yuwen* [*Studies of the Chinese Language*] 1980(1).

Liu, X. 2013. Shangsheng biandiao de shengxue yu ganzhi shiyan yanjiu. [Acoustic and perceptual experimental study of third tone sandhi.] *Wenjiao Ziliao* [*Data of Culture and Education*] 625. 139-142.

McCarthy, J. J. 2005. Taking a free ride in morphophonemic learning. *Catalan Journal of Linguistics* 4(19). 19-55.

Port, R., F. Mitleb, & M. O’Dell. 1981. Neutralization of obstruent voicing in German is incomplete. *J. Acoust. Soc. Am.* 70(S1). S13-S13.

Prince, A., & P. Smolensky. 1993. *Optimality Theory: Constraint Interaction in Generative Grammar*.

Rasin, E., I. Shefi, & R. Katzir. 2020. A unified approach to several learning challenges in phonology. In *Proc. NELS 50*, 73-87.
Richter, C. L. 2021. *Alternation-sensitive Phoneme Learning: Implications for Children's Development and Language Change.* University of Pennsylvania dissertation.
Wang, W. S., & K. P. Li. 1967. Tone 3 in Pekinese. *Journal of Speech and Hearing Research.* 10(3). 629-636.
Yip, M. 1996. *Lexicon Optimization in Languages without Alternations.*
Yuan, J., & Y. Chen. 2014. 3rd tone sandhi in Standard Chinese: a corpus approach. *Journal of Chinese Linguistics.* 42(1). 218-237.
Zee, E. 1980. A spectrographic investigation of Mandarin tone sandhi. *UCLA WPP* 49. 98-116.
Zhang, C., & G. Peng. 2013. Productivity of Mandarin third tone sandhi: a wug test. *Eastward Flows the Great River: Festschrift in Honor of Prof. William S-Y. Wang on his 80th Birthday*. 256-282.
Zhang, J., & Y. Lai. 2010. Testing the role of phonetic knowledge in Mandarin tone sandhi. *Phonology* 27(1). 153-201.
Zhang, N. N. 2015. The morphological expression of plurality and pluractionality in Mandarin. *Lingua* 165. 1-27.