Lexical Propensity and Taiwanese Min Tone Sandhi Rules

Ho-hsien Pan[1], *Hsiao-tung Huang*[1]

[1]Department of Foreign Languages and Literatures, National Chiao Tung University, Taiwan
hhpan@faculty.nctu.edu.tw, fernah823@gmail.com

Abstract

Following chain shift cyclic Taiwanese Min tone sandhi rules, 55, 13→ 33→31→51→55 and 5→3→5, sandhi tones surface at the non-final syllables of tone sandhi groups, whereas canonical base tones surface at the final syllables of tone sandhi groups. Pan (2019) investigated the sandhi and base tone (S/B) alternations with spontaneous speech corpus, TaiMin (www.taimin.tw), and found that the base tones increased from low-level weak syllable and word boundaries to high-level strong intermediate phrase (ip) and intonation phrase (IP) boundaries. However, there were base tones produced before weak word boundary and sandhi tones before strong ip and IP boundaries. This study explores the effect of lexical propensity on S/B alternations in TaiMin. Results of data-driven decision tree model revealed that 90.75% of /i 55/ "he/she" were in sandhi forms. In the TaiMin corpus, there are 31 morphemes produced with over 95% sandhi tones and three morphemes produced with over 95% base tones. There is an effect of lexical propensity on the S/B tone alternations. S/B tone alternations are not entirely determined by phonological or morpho-syntactic conditions.

Index Terms: sandhi tones, base tones, decision tree, random forest, lexical propensity, prosodic boundary

1. Introduction

Lexical propensity was found to influence phonological rules application in Slovenian palatalization and French liaison [1], [2], [3], and [4]. In French, when connecting sounds, the end of a Word 1 is pronounced only if following Word 2 has a vocoid initial. For example, the word très beau ("very beautiful") is produced as [trɛ bo] but as [trɛ z ɛ̃tɛliʒɑ̃] in très intelligent ("very intelligent.") Data from a spoken French corpus study revealed that Word 1 of frequency above one hundred could be divided clearly into either liaiser or non-liaisers [5]. Thus, instead of a stochastic binary choice for phonological rule application (e.g. +/- Rule), due to lexical propensity, some morphemes are more likely to undergo a certain phonological process than are other lexicons (e.g. 0.4 Rule). Furthermore, nonce word probe experiment found that French listeners internalize these liaison propensities [6]. When asked whether they prefer the liaised or non-liaised form for Word 1 très, bien, moins, and pas that were followed by nonce word with vowel initial (e.g. très arvant), their preferences reflected the lexical propensity in the corpus.

Wug tests on the Taiwanese Min tone sandhi chain application to real word and novel words found a low level of productivity on novel words [7], [8], and [9]. Recent auditory priming tests also recommend an allomorph listing of existing syllables/ morphemes instead of a phonological underlying and surface form mapping [10], [11], [12], and [13]. It is likely that there is a lexical effect on S/B alternations.

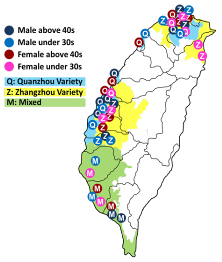
Pan [14] investigated the sandhi and base tone alternations with a spontaneous speech corpus, TaiMin (www.taimin.tw), and found that the syllables before mrophos-syntactic elements such as final particles and modification marker, /ɛ0/, carried predominantly base tones. Moreover, base tones increased from low-level weak syllable and word boundaries to high-level strong intermediate phrase (ip) and intonation phrase (IP) boundaries. Even after excluding checked tones 5 and 3 which were undergoing sound changes, and tones followed by neutral tones due to morpho-syntactic influences, there were 4, 678 cases of base tones before weak word boundaries. These results are in line with the results of Wug tests. As Taiwanese Min tone sandhi rules depend on morpho-syntax, and prosodic hierarchy, it is difficult for native listeners to apply Taiwanese Min tone sandhi rules to isolated novel or real words.

This study expanded beyond the prosodic and morpho-syntactic influences on sandhi and base tone alternations, by also investigates the effect of lexical propensity on sandhi and base tone alternations. By using data driven decision tree model lexical effects were explored. Prosodic wise, special attentions was given to base tones produced before weak word boundaries and sandhi tones strong ip and IP boundaries. These cases cannot be explained with prosodic hierarchy

2. Method

2.1. Speakers

Forty-one speakers were recruited from six dialect regions in Taiwan, including Northern Zhangzhou, Northern Quanzhou, Central Zhangzhou, Central Quanzhou, Southern Mixed and Yilan (Figure 1). The terms Zhangzhou and Quanzhou here do not refer to geographic locations in Mainland China. Instead, Zhangzhou and Quanzhou are the heritage dialects of the speakers. Taiwanese Zhangzhou and Quanzhou dialects are not as distinctively different from each other. The dialects spoken in the Southern Taiwan is Southern Mixed which blend the Zhangzhou and Quanzhou dialects together.



2.2. Corpus

TaiMin (www.taimin.tw) contains elicited monologues thirty minutes in length. Forty-one speakers discussed various topics including their hometowns, professions, hobbies, favorite food and favorite vacation destinations. There were

10.21437/SpeechProsody.2020-106

37,589 utterances with 164,782 words and 218,149 syllables transcribed on nine tiers (Figure 1), including orthography, words, underlying phonemic tones, surface tones, syllable, segment, break, miscellaneous and linguistic tiers.

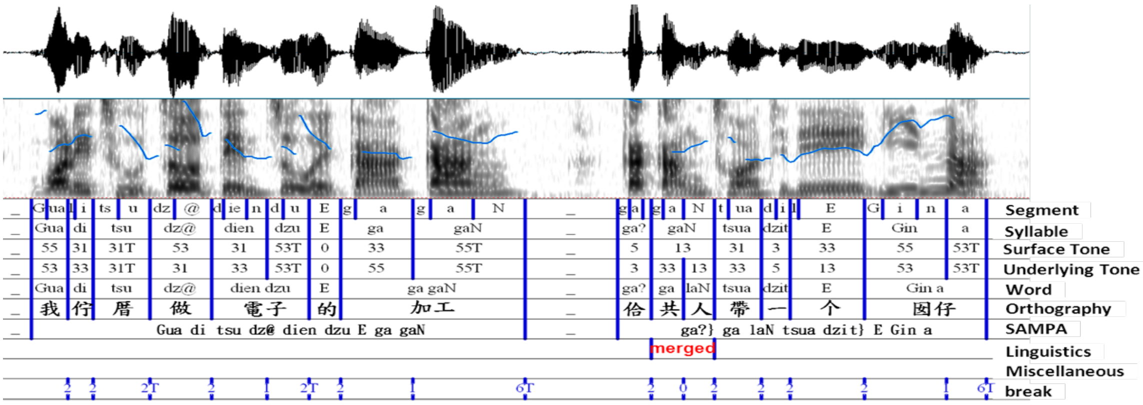

_ G ua di i ts u dz @ d ie n d u E g a g a N _ g a g a N t ua d i l E G i n a
Segment
_ Gua di tsu dz@ dien dzu E ga gaN _ ga? gaN tsua dzit E Gin a
Syllable
_ 55 31 31T 53 31 53T 0 33 55T _ 5 13 31 3 33 55 53T
Surface Tone
_ 53 33 31T 31 33 53T 0 55 55T _ 3 33 13 33 5 13 53 53T
Underlying Tone
_ Gua di tsu dz@ dien dzu E ga gaN _ ga? ga laN tsua dzit E Gin a
Word
_ 我 佇 厝 做 電子 的 加工 _ 佮 共 人 帶 一 个 囡仔
Orthography
_ Gua di tsu dz@ dien dzu E ga gaN _ ga?} ga laN tsua dzit} E Gin a
SAMPA
merged
Linguistics
Miscellaneous
2 2 2T 2 1 2T 2 1 6T 2 0 2 2 2 2 1 6T
break

Figure 1: *The underlying and surface transcription in Taiwanese Min spontaneous speech corpus (TaiMin).*

Each utterance was first transcribed using Chinese characters with spaces inserted between words following the conventions for lexical parsing of the Ministry of Education (MOE) dictionary. The word tier was transcribed using the Speech Assessment Methods Phonetic Alphabet (SAMPA) symbols.

The surface tone values were checked manually while listening to the sound files. The locations of the base tones were manually marked first on the underlying and surface tone tiers.

In addition to tagging segmental and supra-segmental information, the prosodic break indices were labeled for each syllable. To discriminate between different breaks, acoustical cues before and after boundaries were used. Specifically, duration (final lengthening) and f0 (final lowering) were used before boundaries, and pausing, f0 reset, and increased speed were used after the boundary. There were seven types of break indices, numbered 0-6, each of which is discussed below.

Break index “0” was used to represent syllable contraction. For example, the syllable index between the first and the second syllables of /ga33 laŋ13/ [gaŋ13] “for” was 0, as shown in Figure 1. Break indices “1” and “2” were used to stand for syllable and word boundaries, respectively. Finally, break indices “5” and “6” were used to stand for ip and IP, respectively. For break index “5”, final lengthening and f0 final lowering were observed before the boundary and f0 reset after the boundary. However, there was no pause after the ip boundary. Break index “6” was used when there were duration final lengthening and f0 final lowering before the boundary. There were also pauses and f0 resets after the IP boundaries.

The inter-transcriber agreement rates were checked for the surface tone tier and break tiers. Within the surface tone tier, 65.47% of the tones carried by syllables were transcribed by two transcribers and the inter-transcriber agreement rate was 89.81%. Within the break tier, 52.52% of the breaks were labelled by 2 transcribers, and the inter-transcriber agreement rate was 95.66%.

The prosodic boundary indicies in the break tier were used along with surface sandhi and base tones in the surface tone tier, and lexical information in the word tier.

2.3. Data Analysis

A Classification and Regression Tree (CART) model analysis of the 129,273 sandhi or base tones in all syllable structures occurring in the domain-final position is presented in this study. Since it is impossible to include each of the thousands of lexical items as an individual factor in the model, syllable structures were used as a surrogate factor. The lexical items of any syllable structures found to be effective in determining sandhi and base tone alternations were analyzed further.

The “rpart” package in R was used for the CART modelling. Within the decision tree model, data were randomly divided into training data (70%) and test data (30%). For each node, the decision tree algorithms selected the best attribute and divided the learning data so that the best prediction values for the right classification could be obtained. In this process, each individual feature/attribute was taken in turn and a tree consisting of nodes containing that feature/attribute was built. The single best node was kept and the remaining features were further analyzed in turn and added hierarchically to the tree model. This procedure was repeated for each of the features until no significant gain in accuracy was obtained by the inclusion of additional features.

The features included age (< 30, > 40), dialect regions (Northern Zhangzhou, Northern Quanzhou, Central Zhanzhou, Central Quanzhou, Southern Mixed, and Yilan) and gender (male, female), syllable duration (a potential focus index), all syllable structures, underlying tone identity (55, 13, 51, 31, and 33), following elements (boundary, modification marker and final particles) and prosodic boundary (syllable, word, ip, and IP). Each of these features was in turn included in the data-driven self-learning approach model to sort the tone identity (sandhi and base tones). This model is particularly suitable for Taiwanese Min, an under-investigated language, of which the relevant features determining sandhi and base tone alteration remain elusive.

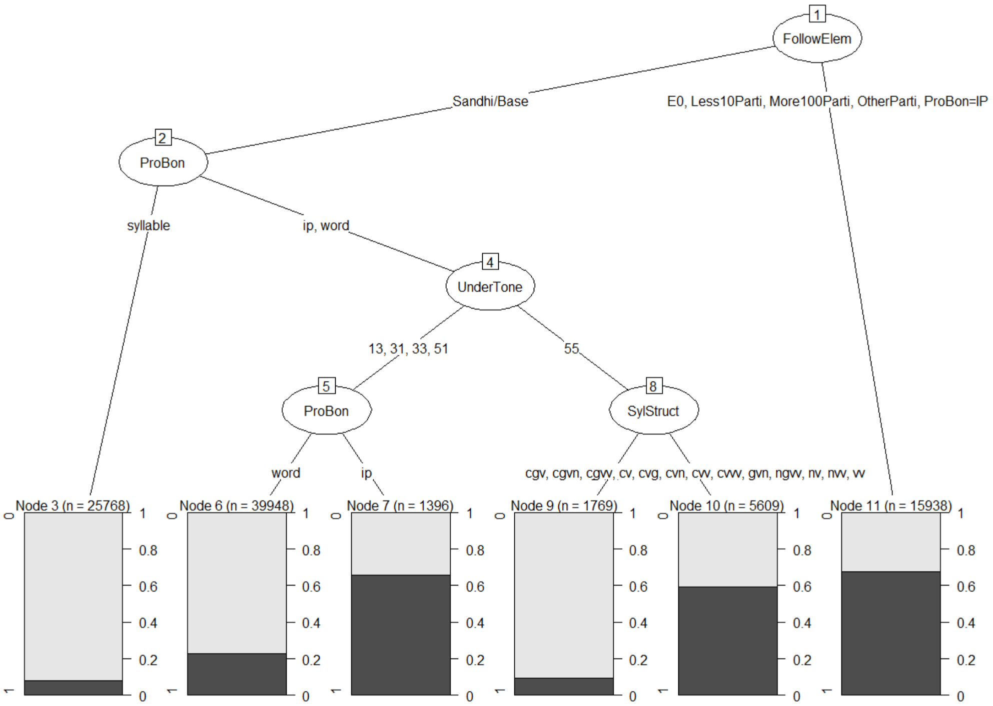


Figure 2: *The decision tree model presenting the sandhi and base tones at the syllable, word, ip and IP domain-final syllables. FollowElem: Following elements, ProBon: Prosodic boundary, UnderTone: Underlying tones, SylStrcut: Syllable structure.*

It should be noted that both tone 3 and 5 were excluded from analysis, since they were undergoing sound changes and were unstable in tonal values.

In addition to using decision tree models to analyze the S/B alternation in the corpus, this study also focused on canonical base tones in low level weak word-final positions and non-canonical sandhi tones in high level strong ip-final and IP-final positions. The reason for these analyses is that the occurrence of canonical base tones before weak boundaries and non-canonical sandhi tones before strong prosodic boundaries cannot be explained by prosodic hierarchical structure of domain-final strengthening alone.

3. Results

3.1. Decision Tree Model

As shown in Figure 2, the most dominant factor for the root node #1 was following elements, including final particles or IP boundary. Node #1 branched into a leaf node #11 for syllables followed by either final particles, or IP boundary, and a parent node #2 for syllables followed by either sandhi or base tones. The parent node #2 further branched into a leaf node #3 for tones in syllable-final positions, and a parent node #4 for syllables in word or ip-final positions. The parent node #4 further branched into two parent nodes, including parent node #5 for tones 13, 51, 31 and 33 in word-final (leaf node #6) or ip-final (leaf node #7) positions, and another parent node #8 for tones 55. The parent node #8 further divided into according to syllable structures, with leaf node #9 for GV (glide + vowel), V (vowel) and VN (vowel + nasal) syllable structures, and another leaf node #10 for other syllable structures.

leaf node # 9 were consisted of 2547 morphemes, including 1, 804 repetitions for the morpheme /i^{55}/ "he /she" 伊 produced with 91.46% of sandhi tones, and 681 repetitions for /in^{55}/ "They / Them" that were produced with 92.2 % sandhi tones in ip-final and word-final positions. It is proposed that there was an effect of morpheme / lexical propensity on the sandhi and base tone alternations.

3.2. Morphemes Produced with Predominantly Base Tones

To determine the lexical propensity effect, the morphemes carrying non-checked tones with over lexical frequency were investigated. Following is a list of seven morphemes produced with over 90% base tones.

Table 1: *Morphemes produced with over 90% base tones.*

Hanzi	Gloss	IPA	Frqncy	Base %
區	District	k^{h}u	65	0.969
囝	Child	kĩã	55	0.964
兜	Home	tau	121	0.959
舍	House	sia	66	0.939
課	class	k^{w}e	71	0.930
箍	gather	ko	70	0.929
漢	man	han	136	0.927

Figure 3 shows the numbers of base and sandhi tones for the 26 morphemes produced with over 80 % base tones in syllable-final, word-final, ip-final and IP-final positions. Regardless of prosodic position, these morphemes were produced more base tones than sandhi tones. In other words, prosodic boundaries has little or no effects on the high percentage of base tones.

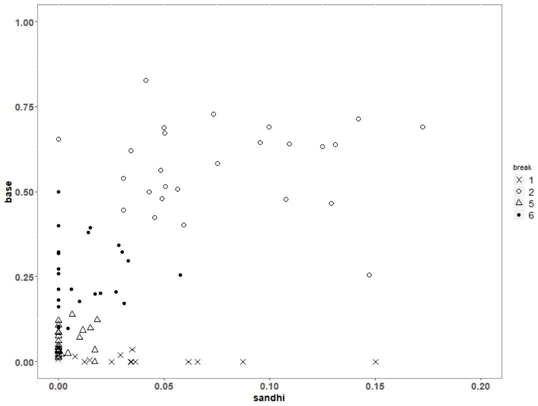


Figure 3: *Percentages of sandhi and base tones for morphemes produced with over 80% of base tones (top panel) distributed in syllable (x), word (open circle), ip (open triangle) and IP (filled circle) domain-final positions.*

3.3. Morphemes Produced with Predominantly Sandhi Tones

Table 2 shows the morphemes produced with over 95% sandhi tones. The fact that there are morphemes produced with sandhi tones than base tones did not come as a surprise, since syllables in non-final positions of tone sandhi groups were all produced with sandhi tones.

Table 2: *Mrophemes produced with over 95% sandhi tones.*

Hanzi	Gloss	IPA	frqncy	Sndh %
蓋	very	kai	120	1.0
工	day	kaŋ	91s	1.0
囡	child	gin	283	0.995
進	enter	dzin	120	0.992
研	study	gien	108	0.991
禮	manner	lɛ	76	0.987
敢	dare	kã	207	0.986
嘉	well	ka	62	0.984
幫	help	paŋ	57	0.983
其	it	ki	473	0.979
愛	love	ai	985	0.978
今	now	kin	89	0.978
四	four	si	205	0.976
盡	exhaust	dzin	81	0.976
傷	wound	sĩũ	107	0.972
鬥	fight	tau	70	0.971
按	according	an	1743	0.971
查	check	ts^{h}a	196	0.969
感	feel	kam	555	0.969
高	high	kə	276	0.967
應	should	iŋ	271	0.967
對	correct	tui	88	0.966
阿	diminutive mark	a	259	0.966
愈	More	lu	101	0.960
寡	Alone	kua	320	0.959
主	master	dzu	120	0.958
嘛	final particle	mã	1682	0.955
的	modification mark	ɛ	579	0.952
新	new	sin	289	0.952
攏	all	lɔŋ	2187	0.952
爾	final particle	nĩ	61	0.951

Figure 4 shows the percentages of sandhi and base tones for morphemes produced with over 80% sandhi tones. In IP final positions, the number of

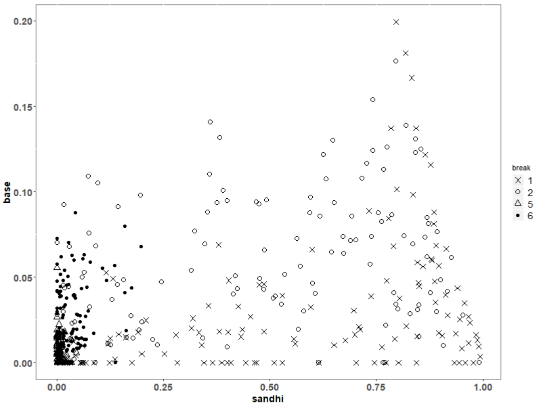


Figure 4: *MPercentages of sandhi and base tones for morphemes produced with over 80% of sandhi tones (top panel) distributed in syllable (x), word (open circle), ip (open triangle) and IP (filled circle) domain-final positions.*

The fact that some morphemes tended to be produced with base tones whereas other tended to be produced with sandhi tones, suggesting that there may be some mental lexicons are dominated with high frequency exemplars. These high frequency examplars may or may not be in canonical base form. Speakers can memorize these high frequency forms and produce the frequent forms without referring to tone sandhi chain or prosodic positions.

4. Discussions

The results of this study show that the alternations between sandhi and base tones are not purely phonological in nature. Factors such as morpho-syntactic elements, prosodic boundary and lexical propensity also affect the production of sandhi and base tones. That might explain the results of previous perceptual tests or Wug tests, which found that native speakers experienced difficulties in the application of tone sandhi rules to isolated nonce words.

This study is not proposing the abandonment of the Taiwanese Min tone sandhi chain. Instead, it is proposing additional non-phonological factors that may influence the alternations between sandhi and base forms. It is proposed that native speakers do have an understanding of the tone sandhi chain. However, the rules are not applied to every syllable during the online speech production process. Instead, there are certain lexical forms that are predominantly produced with either sandhi tones or base tones. When the lexical items do not have the propensity to be produced with either sandhi or base forms, then the chain-shift tone sandhi rules are proposed to be applied.

Future studies can explore whether native listeners' preference of sandhi or base forms reflect lexical propensity, just as the application of French liaison rules was affected by lexical propensity.

5. References

[1] K. Zuraw, “Frequency influences on rule application within and across words,” *Proceedings of Chicago Linguistic Society*, vol. 43, 2009.

[2] K. Zuraw, “A model of lexical variation and the grammar with application to Tagalog nasal substitution,” *Natural Language and Linguistic Theory*, vol. 28, no. 2, pp. 417–472, 2010.

[3] P. Smolensky and M. Goldrick, *Gradient symbolic representations in grammar: The case of French liaison*. ROA, 1286, 2016. http://roa.rutgers.edu/content/article/files/1552_smolensky_1.pdf (accessed 29th December 2019).

[4] J. Zymet, *Lexical propensities in phonology: corpus and experimental evidence, grammar and learning.* Ph.D. Dissertation, UCLA, 2018.

[5] J. Durand and C. Lyche, “French liaison in the light of corpus data, Journal of French Language studies, vol. 18, pp. 33–66, 2008.

[6] G. Mallet, *La liaison en français: Descriptions et analyses dans le corpus PFC.* Ph.D. dissertation, Université Paris Ouest, Nanterre La Défense. 2008.

[7] H.-I. Hsieh, “The psychological reality of tone sandhi rul in Taiwanese,” *In Papers from the 6th Meeting of the Chicago Linguistic Society*, pp. 489–503, 1970.

[8] H.-I. Hsieh, “How generative is phonology,” In Ernst Frideryk Konrad Koerner (ed.) The Transformational-Generative Paradigm and Modern Linguistic Theory, John Benjamin, Amsterdam, The Netherlands, pp. 109–144, 1975.

[9] H.-I. Hsieh, “On the unreality of some phonological rules,” *Lingua*, vol. 38, pp. 1–19, 1976.

[10] J. Tsay and M. James “Taiwanese tone sandhi as allomorph selection,” *Proceedings of 22nd Meeting of the Berkeley Linguistics Society*, pp. 394–405. Berkeley, CA, 1996.

[11] S.-h. Peng, “Production and perception of Taiwanese tones in different tonal and prosodic contexts,” *Journal of Phonetics*, vol. 25, no. 3, pp. 371–400, 1997.

[12] J. Zhang, “A directional asymmetry in Chinese tone sandhi systems,” *Journal of East Asian Linguistics*, vol. 16, no. 4, pp. 259–302, 2007.

[13] Y.-f. Chien, J. Sereno and J. Zhang, “Priming the representation of Mandarin tone 3 sandhi words,” *Language, cognition and Neuroscience*, vol. 31, no. 2, pp. 179–189, 2016.

[14] H.-h. Pan, H.-t. Huang and S.-r. Lyu, “The occurrence of Taiwanese Min juncture tones before prosodic boundaries and modification marker.” In Sasha Calhoun, Paola Escudero, Marija Tabain & Paul Warren (eds.) *Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, Australia 2019 (pp. 3423-3427) Canberra, Australia: Australasian Speech Science and Technology Association Inc, 2019. https://assta.org/proceedings/ICPhS2019/papers/ICPhS_3472.pdf