Opacity, Phonetics, and Frequency in Taiwanese Tone Sandhi

Jie Zhang, Yuwen Lai, and Craig Sailor

The University of Kansas
1541 Lilac Lane, Blake Hall, Room 427
Lawrence, KS 66044, USA
zhang@ku.edu

Abstract

A wug test study of tone sandhi patterns in Taiwanese indicates that sandhi productivity is affected by phonological opacity as well as the durational property and lexical frequency of the sandhi. Opacity outweighs phonetics and frequency as a global effect; frequency effects are evident for everyday users of the language; phonetic effects only surface for occasional users for whom the frequency effects have been weakened due to the lack of use of the language. The simultaneous underlearning of exceptionless opaque patterns, overlearning of phonetic effects, and proper learning of lexical statistics by Taiwanese speakers can be modeled by a Maximum Entropy grammar that encodes learning biases against lexical listing constraints and phonetically unmotivated patterns.

Keywords: tone sandhi, opacity, frequency, productivity, Maximum Entropy.

1. Introduction

The productivity of a linguistic process refers to its ability to apply to new items (Bybee 2001). The understanding of productivity is important to theoretical linguistics as it provides crucial evidence about the generalizations and cognitive abstractions that speakers make (Bybee 2001, Pierrehumbert 2003).

The productivity of a phonological process crucially depends on various properties of the process. Type frequency, which refers to the dictionary frequency of a pattern, and token frequency, which refers to the frequency of occurrence of a pattern in a corpus, have long been established to influence productivity (Bybee 1985, 2001, Moder 1992, Pierrehumbert 2003, 2006, among others). Recent research has also shown that both the phonetic grounding (Wilson 2003, 2006, Zhang and Lai 2006, Zuraw 2007) and phonological opacity (Hsieh 1970, Sanders 2001, Zhang and Lai 2008) of a pattern may have an effect on productivity. The goal of the this research is to investigate the interaction of the effects of frequency, phonetics, and opacity on the productivity of Taiwanese tone sandhi, in which all three effects are operative. Empirically, we report the results of a wug test that taps into the productivity of the tone sandhi; theoretically, we formally model the gradient productivity results as a reflection of the speakers' phonological knowledge using the Maximum Entropy (MaxEnt) model (Goldwater and Johnson 2003, Wilson 2006, Hayes and Wilson 2008, Jäger, to appear) as a blueprint.

2. Taiwanese tone sandhi

Taiwanese tone sandhi is a positionally conditioned tonal alternation pattern that is characterized by circular opacity: tones in nonfinal positions of syntactic phrases undergo sandhi, and four out of five tones in the tonal inventory are involved in a "tone circle," as in (1) (Chen 1987, Lin 1994). The other tone 24 is involved in a phonotactically transparent sandhi 24 → 33, motivated by the generalization that 24 cannot occur in nonfinal positions.

(1) Taiwanese tone sandhi in non-XP-final positions:

$$\begin{tikzcd}
51 \arrow[r] & 55 \arrow[r] & 33 \arrow[dl] & 24 \arrow[l] \\
& 21 \arrow[ul] & &
\end{tikzcd}$$

The different sandhi patterns in Taiwanese have different phonetic bases. Phonetic studies by Lin (1988) and Peng (1997) showed that the two falling tones 51 and 21 have considerably shorter intrinsic durations than 55, 33, and 24. Given that the sandhis occur on nonfinal syllables, which are known to be shorter than final syllables due to the lack of final lengthening (Oller 1973, Klatt 1975, Wightman *et al.* 1992, among others), the 33 → 21 sandhi has a durational basis, as it turns a longer tone into a shorter tone; the 51 → 55 sandhi is an anti-duration change, as it turns a shorter tone into a longer tone; the other two sandhis are durationally neutral.

Different base tones in Taiwanese have different type and token frequencies, as estimated from a spoken corpus by Tsay and Myers (2005). In (2), syllable type frequency refers to the number of different syllables in the Taiwanese syllabary represented in the corpus that can carry a tone; morpheme type frequency refers to the number of different monosyllabic morphemes represented in the corpus that can carry a tone; and token frequency is the number of occurrence of a tone in the entire corpus.

Interestingly, these frequency counts do not put the tones in the same order, which provides us with an opportunity to study which frequency count has the greatest effect on productivity. Furthermore, the effects of frequency may also be manifested differently for speakers in different environments; for example, speakers who speak the language daily may exhibit a stronger frequency effect than those who only speak it occasionally.

(2) Tone frequency counts in Taiwanese:
Syllable type frequency: 55 > 51 > 24 >21 >33
Morpheme type frequency: 55 > 24 > 51 > 33 > 21
Token frequency: 55 > 24 > 33 > 51 > 21

Previous research has shown that opacity, phonetics, and frequency all have an effect on the productivity of the sandhi pattern. Hsieh (1970) showed that the application rate of the entire Taiwanese tone sandhi pattern was as low as 10-30% in novel disyllabic words. Hsieh (1975), Wang (1993), and Zhang *et al.* (2006) all showed that the phonotactically transparent sandhi 24 → 33 had a higher application rate to novel words than opaque sandhis in the tone circle. Hsieh (1975) indicated that there might be a frequency effect, as the 55 → 33 sandhi, which has the highest counts for both type and token frequencies, had a higher application rate in novel words than other opaque sandhis in the tone circle. Finally, Zhang *et al.* (2006) reported that Taiwanese speakers tested at the University of Kansas showed a higher productivity for the duration reducing sandhi 33 → 21 than for the duration increasing sandhi 51 → 55. Our current study further quantifies the interaction among these effects. We also explicitly address the question of how these effects may differ for different speaker populations by comparing two groups of speakers — one from Taiwan, who uses the language in their daily functions; one from Kansas, US, who only uses the language occasionally.

3. Experimental methods

3.1. Stimuli and participants

The basic method of our experiment was to present the subjects with two monosyllables and ask them to pronounce the syllables together as a true disyllabic word in Taiwanese. Our analyses focused on the tone on the first syllable of the subjects' responses.

There are two within-subjects factors in the experiment. The first is Word Type. Following Hsieh (1970)'s experimental design, we constructed five types of disyllabic words in Taiwanese. The first type is real words, denoted by AO-AO (AO = actual occurring morpheme). These words served as the control for the experiment. The other four types 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 the Taiwanese 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 hand-picked by the second author, who is a native speaker of Taiwanese. In each AG, both the segmental composition and the tone of the syllable are legal, but the combination happens to be missing in Taiwanese. The second within-subject factor is Sandhi Type, which is determined by the tone on the first syllable of the disyllables. There are five different sandhi types, represented by the five tones in the tonal inventory on non-checked syllables — 24, 55, 33, 21, and 51. The tone on the second syllable was kept to a constant 33. Necessary frequency controls across the sandhi types were exercised using Tsay and Myers (2005). Eight words for each Word Type × Sandhi Type combination were used, making a total of 200 test words (8×5×5). We also used 160 filler words, which had tones other than 33 on the second syllable.

The experiment also has a between-subject factor, which is Speaker Group. We recruited 20 speakers from Taiwan (6 male, 14 female), who had an average age of 51.6 and all used Taiwanese in their daily functions, and 16 speakers from Kansas, US (6 males, 10 females), who had an average age of 31.3 and had been in the US for an average of 3.8 years — these speakers only spoke Taiwanese for an average of 45 minutes a week, usually in phone calls to their families in Taiwan. The Kansas speakers are also considerably younger than the Taiwan speakers, and sociolinguistic studies have shown that younger speakers use Taiwanese less often due to the increasing influence of Mandarin (Sandel 2003, Scott and Tiun 2007).

3.2. Experimental procedure

The experiment was conducted with SuperLab (Cedrus) in the Phonetics and Psycholinguistics Laboratory at the University of Kansas for the Kansas speakers and in a quiet room for the Taiwan speakers. The stimuli were played through a headphone worn by the subjects. Each stimulus consisted of two monosyllabic utterances read by the second author, separated by an 800ms interval. The subjects were instructed to put the two syllables together and pronounce them as a true disyllabic word in Taiwanese. Their response was collected by either a Sony PCM-M1 DAT recorder (in Kansas) or a Marantz solid state recorder PMD 671 (in Taiwan).

3.3. Data analyses

The sandhi tones on the first syllable of the test words were transcribed by the three authors — a native speaker of Taiwanese (Lai), a native speaker of Beijing Mandarin (Zhang), and a native speaker of American English (Sailor), all phonetically trained — using a 1-5 scale with the help of pitch tracks in Praat (Boersma and Weenink 2005). There was clear agreement for nearly all cases among the authors. In cases of disagreement, the judgment of the native Taiwanese author was taken. Based on the transcriptions, the correct response rates for the tone sandhis in each Word Type × Sandhi Type combination for each speaker were then calculated.

Mixed-design ANOVAs with Word Type and Sandhi Type as within-subject factors and Speaker Group as a between-subject factor were conducted for the correct response rates. Regression analyses with the durational property of the sandhis and the type and token frequencies of base tones as predictors for the correct application rates of the opaque sandhis in wug words (*AO-AO, AO-AG, AG-AO, AG-AG) were also conduced for the two groups of speakers. All statistics were carried out in SPSS.

4. Results and discussion

The main effects of Speaker Group, Word Type, and Sandhi Type on the correct response rate from the ANOVA are plotted in Figure 1. The effect of Speaker Group is not significant: F(1, 34)=0.296, p=0.590. The effect of Word Type is significant: F(3.552, 120.766)=462.281, p<0.001. So is the effect of Sandhi Type: F(3.000, 102.012)=37.140, p<0.001). Post-hoc analyses with Bonferroni adjustments showed that real words AO-AO have a significantly higher correct response rate than all wug groups (p<0.001 for all paired comparisons). The average correct response rate is 70.1% for real words, but only 16.9% for wug words. Post-hoc analyses also showed that the phonotactically transparent sandhi 24 → 33 has a significantly higher correct response rate than all opaque sandhis (p<0.001 for all paired comparisons). The average correct response rate is 44.2% for 24 → 33, but only 23.3% for the opaque sandhis.

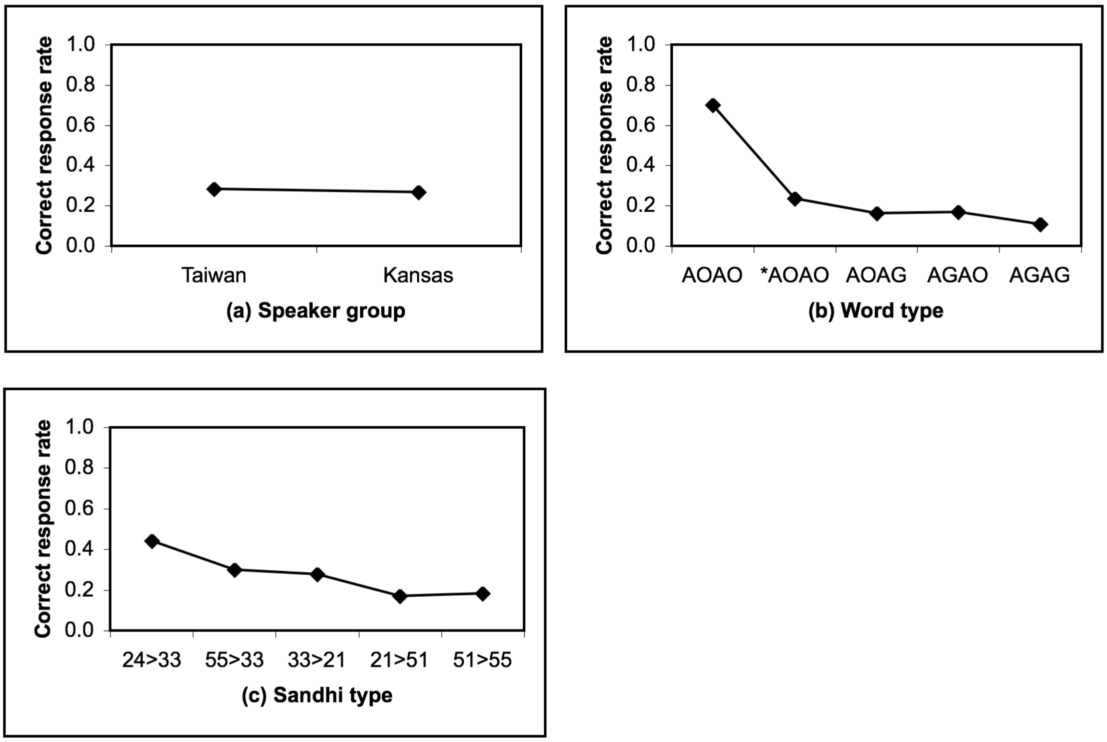


Figure 1: Effects of (a) Speaker Group, (b) Word Type, and (c) Sandhi Type on the correct response rates for tone sandhis in Taiwanese.

The interactions between the within-subject factors and the between-subject factor are plotted in Figure 2. The Word Type × Speaker Group interaction just reached significance: F(3.552, 120.766)=2.626, p=0.044. The Sandhi Type × Speaker Group interaction is also significant: F(3.000, 102.012)=3.022, p=0.033.

Two separate ANOVAs and subsequent post-hoc analyses on the two subject groups showed that for both Taiwan and Kansas speakers, real words AO-AO have a significantly higher correct response rate than all wug groups at the p<0.001 level; for Taiwan speakers, the phonotactically transparent sandhi 24 → 33 has a significantly higher correct response rate than all the opaque sandhis in the tone circle at the p<0.001 level, and for Kansas speakers, 24 → 33 has a significantly higher correct response rate than all opaque sandhis at the p<0.05 level except for 33 → 21. These results indicate that both groups of Taiwanese speakers had considerably more trouble in the application of tone sandhis to wug words and the application of opaque tone sandhis.

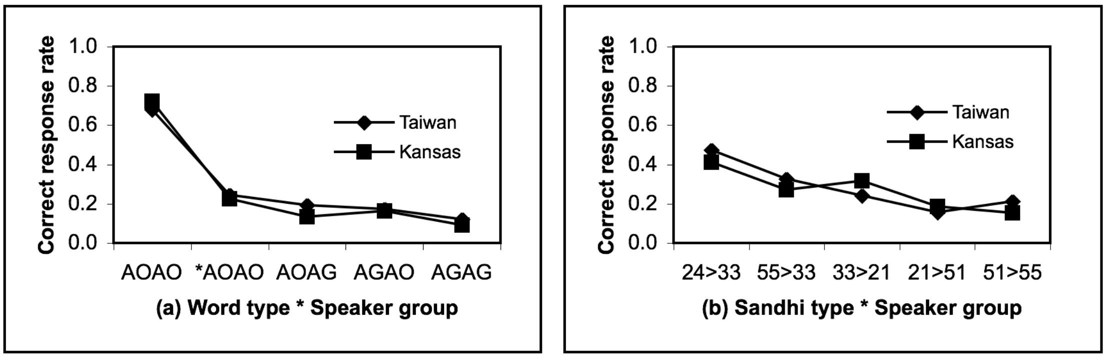


Figure 2: Interactions (a) between Word Type and Speaker Group and (b) between Sandhi Type and Speaker Group on the correct response rates for tone sandhis in Taiwanese.

To further compare the two groups of speakers on the potential effects of phonetics and frequency on sandhi productivity, we conducted an additional mixed-design ANOVA on only the correct response rates for opaque sandhis in wug words. The main effect of Word Type is still significant: F(3.000, 102.000)=23.896, p<0.001. So is the main effect of Sandhi Type: F(2.239, 76.132)=15.229, p<0.001. The interactions between the within-subject factors and the between-subject factor are graphed in Figure 3. The Word Type × Speaker Group interaction is not significant: F(3.000, 102.000)=1.795, p=0.153. But the Sandhi Type × Speaker Group interaction is significant: F(2.239, 76.132)=4.480, p=0.012).

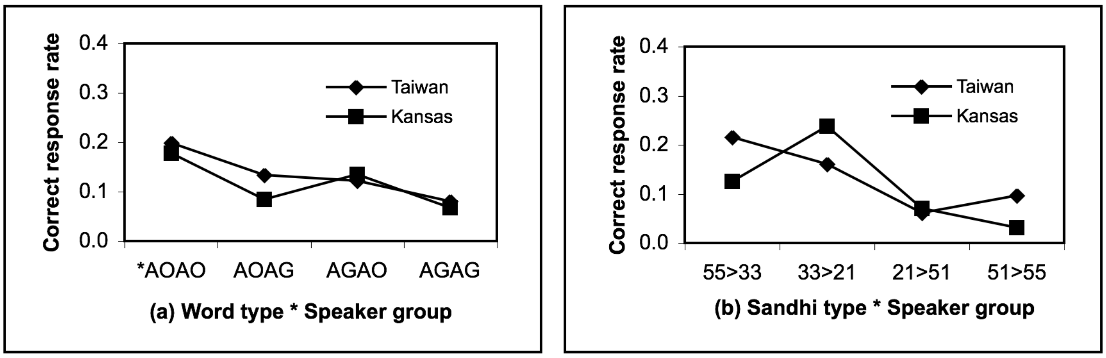


Figure 3: Interactions (a) between Word Type and Speaker Group and (b) between Sandhi Type and Speaker Group on the correct response rates for opaque tone sandhis in Taiwanese wug words.

The significant interaction between Sandhi Type and Speaker Group indicates that the two groups of speakers behaved differently in the application of opaque sandhis to wug words. Figure 3b shows that for Taiwan speakers, the highest productivity is found in the sandhi whose base tone has the highest frequency counts (55 → 33), and the lowest productivity is found in the sandhi whose base tone has the lowest morpheme type and token frequencies (21 → 51); for Kansas speakers, however, the sandhi with the greatest duration reduction (33 → 21) has the highest productivity, while the sandhi with the greatest duration increase (51 → 55) has the lowest productivity. These results indicate that Taiwan speakers' behavior may be closely related to frequencies, while Kansas speakers' behavior may be closely related to phonetic duration.

To further investigate the effects of frequencies and duration on sandhi productivity in the two speaker groups, we conducted a forward stepwise regression analysis for each speaker group with syllable type frequency, morpheme type frequency, and token frequency of the base tone and the amount of duration reduction of the sandhi as potential predictors for the productivity of opaque sandhis in wug words. The values for these predictors are given in Table 1. The frequency counts are from Tsay and Myers (2005) and have been *log*-transformed. The duration reduction data were deduced from the durations of the tones read in isolation reported in Lin (1988). A positive value indicates a duration decrease from the base tone to the sandhi tone, and a negative value indicates a duration increase.

	55 → 33	33 → 21	21 → 51	51 → 55
Syllable type frequency	2.499	2.391	2.407	2.467
Morpheme type frequency	2.872	2.715	2.687	2.736
Token frequency	5.085	5.058	4.972	5.020
Duration reduction (ms)	1	69	-9	-59

Table 1: Predictors for correct response rates for opaque tones sandhis in Taiwanese wug words used in regression analyses.

The regression model for the Taiwan speakers only included token frequency as a significant predictor (R^2=.502, adjusted R^2=.466, β=.708, p=.002; syllable type frequency: p=.952; morpheme type frequency: p=.664; duration reduction: p=.843). The model for the Kansas speakers, on the other hand, only included duration reduction as a significant predictor (R^2=.616, adjusted R^2=.589, β=.785, p<.001; syllable type frequency: p=.491; morpheme type frequency: p=.558; token frequency: p=.348).

In summary, our wug-test results showed a speaker-group independent opacity effect in support of our hypothesis. The opaque tone circle is largely unproductive, as indicated by its low correct response rates in wug words (12.6%). In addition, an analysis of the sandhi errors indicates that the vast majority of the errors are non-application errors (78.7%), which lends further support to the unproductive nature of these sandhis. The phonotactically transparent sandhi 24 → 33, on the other hand, is significantly more productive, as indicated by a higher correct response rate in wug words (34.5%). Moreover, the non-application rate of the sandhi in wug words is considerably lower (31.3%) than those of the opaque sandhis, indicating that even though the speakers might not know the exact outcome of the sandhi, they have effective knowledge of the phonotactic generation that "a rising tone 24 cannot occur in nonfinal positions."

Our hypothesis that the effects of phonetics and frequency may be speaker-group dependent, in that Kansas speakers will be more strongly affected by duration, while Taiwan speakers will be more strongly affected by frequency counts, is also borne out by the regression analyses. To interpret the speaker-group dependent frequency effect, we assume that the Taiwan speakers exhibit a strong frequency effect due to their everyday usage of the language, and the Kansas speakers have stopped using the language regularly, causing the attrition of lexical strength of all words, which consequently causes the frequency effects to fall below a certain threshold. The age difference between our two groups of speakers, which also translates into usage differences, may have contributed to their difference in frequency effects as well. The difference in phonetic effects between the two speaker groups is correlated with their difference in frequency effects. Due to the incompatibility of the two effects (for instance, phonetic effects predict a high productivity of 33 → 21 and a low productivity of 51 → 55; but frequency effects predict a high productivity for 55 → 33 and a low productivity for 21 → 51), for Taiwan speakers, the phonetic effects may have simply been overridden by the strong frequency effects. But when the frequency effects weaken due to the lack of usage, as for our Kansas speakers, the phonetic effects surface. Therefore, failing to detect the phonetic effects for the Taiwan speakers does not necessarily entail that they do not exist.

We recognize that our experimental result on the speaker group effect may have two opposing interpretations. One interpretation is that it demonstrates the usage-based nature of phonology *à la* Bybee (1985, 2001), as it shows that the speakers' behavior is determined by the context of usage. But conversely, it could also be interpreted as demonstrating that the only true linguistic effects are those of opacity and phonetics, as they are shared by all speakers; the effects of frequency are metalinguistic. Without committing ourselves to either of these extreme positions, in the following section, we propose an alterative that incorporates both processing and phonological factors to predict the observed behavior of both Taiwan and Kansas speakers. This echoes the position espoused in works such as Hayes and Londe (2006), Pierrehumbert (2003, 2006), Wilson (2006), Zuraw (2000, 2007), and Zhang and Lai (2008).

5. A theoretical model

To model Taiwanese speakers' tone sandhi behavior in the wug test, we are faced with the following challenges. First, the model needs to account for the fact that the speakers are exposed to exceptionless opaque patterns in the lexicon, but somehow do not internalize the patterns productively. We term this the underlearning challenge. Second, the model needs to ensure that the speakers have phonetic knowledge that affects productivity, even though the lexical statistics does not inform them of such knowledge. We term this the overlearning challenge. Lastly, the model also needs to capture the speakers' detailed knowledge gleaned from lexical statistics, which we term the proper-learning challenge.

5.1. The Maximum Entropy model

The Maximum Entropy model of phonology (Goldwater and Johnson 2003, Wilson 2006, Hayes and Wilson 2008, Jäger, to appear) is a model that is equipped to handle the challenges laid out above. As a variant of Optimality Theory, it is inspired by conditional random fields in information theory (Della Pietra *et al.* 1997, Lafferty et al. 2001) and closely related to Harmonic Grammar championed by Smolensky (1986) and Smolensky and Legendre (2006).

In lieu of ranking the constraints on a linear scale, the MaxEnt model assumes that each constraint C_i has a weight of w_i. For a candidate y of a given input x, if $C_i(y|x)$ is the number of times y violates the constraint C_i, then the Harmonic Score of y given x is defined as e to the negative power of the weighted sum of the numbers of violations of all constraints, as shown in (3).

(3) $H(y \mid x) = \exp(-\sum_i w_i C_i(y \mid x))$

Given the Harmonic Score of *y*, the probability of *y* as the output to *x* is then the proportion of its Harmonic Score to the sum of all Harmonic Scores of *x*'s candidates, as shown in (4).

(4) $$p(y \mid x) = \frac{H(y \mid x)}{\sum_{y \in \Omega} H(y \mid x)},\ \Omega \text{ is the set of candidates for } x.$$

Given the constraint set, learning in a MaxEnt model from a set of training data *D* composed of input-output pairs is to determine the constraint weights that maximize the log probability of *D*, which equals to the log of the product of the probabilities of all input-output pairs in *D*, as in (5).

(5) $$\log(p(D)) = \log(\prod_{y \mid x \in D} p(y \mid x))$$

To prevent overfitting the training data, each weight is often considered to be associated with a regularizing Gaussian prior (Martin et al. 1999, among others). The prior specifies μ as the default weight for a constraint, and σ^2 as the determinant of how severe the penalty is if the weight deviates from the default — the smaller the σ^2, the greater the penalty. Learning, then, is to find the constraint weights that maximize the function combining log(*p*(*D*)) and the penalty, as shown in (6). Crucially, learning *biases* can be encoded as different σ^2s for different constraints, as we will see in §5.3.

(6) $$\log(\prod_{y \mid x \in D} p(y \mid x)) - \sum_i \frac{(w_i - \mu_i)^2}{2\sigma_i^2}$$

5.2. Constraints

We turn to the necessary constraints for our analysis in this section. Further statistical analyses of the correct response rates reported above showed that for both Taiwan and Kansas speakers, there is a gradation in sandhi productivity from real words to wug words in which the first syllable is an actual occurring syllable (*AO-AO, AO-AG) and to wug words in which the first syllable is an accidental gap (AG-AO, AG-AG). We hence represent the three types of words as AO, *AO, and AG. All pairwise comparisons are significant at the p<.001 level except for the difference between *AO and AG for Kansas speakers, which is at p<.05. This gradation in sandhi productivity from AO to *AO to AG indicates that the theory needs to encode three different levels of listedness. First, the higher productivity of AO than *AO requires the listing of disyllabic words in the lexicon, and we posit a group of USELISTED constraints (see Zuraw 2000) on disyllables that force the listed disyllables to be used, as in (7a). Second, *AO's higher productivity than AG indicates that sandhi *allomorphs* of existing syllables are also listed, and we posit a second group of USELISTED constraints that forces the listed syllable allomorphs to be used in nonfinal sandhi positions, as in (7b). Finally, the modest productivity of the sandhi in AG words indicates that the *tonal* allomorphs independent of segmental contents are also listed in the grammar, and we posit a third group of USELISTED to force the use of such tonal allomorphs, as in (7c).

(7) a. USELISTED(σ55-σ): Use the listed /σ33-σ/ for /σ55/+/σ/.
Mutatis mutandis for USELISTED(σ33-σ), USELISTED(σ21-σ), USELISTED(σ51-σ), and USELISTED(σ24-σ)
b. USELISTED(σ55): Use the listed allomorph /σ33/ for /σ55/ non-XP-finally.
Mutatis mutandis for USELISTED(σ33), USELISTED(σ21), USELISTED(σ51), and USELISTED(σ24)
c. USELISTED(55): Use the listed tonal allomorph /33/ for /55/ non-XP-finally.
Mutatis mutandis for USELISTED(33), USELISTED(21), USELISTED(51), and USELISTED(24)

The analysis also needs two other constraints, shown in (8). The markedness constraint *24-NONFINAL expresses a phonotactically true generalization in Taiwanese. IDENT(Tone) discourages tone sandhi.

(8) a. *24-NONFINAL: Tone 24 cannot occur on non-XP-final syllables.
b. IDENT(Tone): The tones of the corresponding syllables in the input and the output must be identical.

5.3. Learning biases as σ^2 values

Wilson (2006) argued that learning biases can be encoded as different σ^2 values of the Gaussian prior in a MaxEnt model. We capitalize on Wilson's claim to capture the underlearning and overlearning aspects of our Taiwanese speakers' behaviors here.

We assume that the default weight μ for all constraints to be zero. The σ^2 for *24-NONFINAL and IDENT(Tone) is set to 10^{-3}; the σ^2 values for USELISTED constraints, however, are $10^{-3}\times B_{Listed}\times B_{Duration}$, where B_{Listed} and $B_{Duration}$ are two coefficients representing the biases based on listedness and phonetic duration, respectively.

For each type of USELISTED constraints (disyllabic words, monosyllabic allomorphs, tonal allomorphs), we posit B_{Listed} to be 10 to the negative power of a logistic function, where x represents the number of morphemes that the type of USELISTED constraints covers on average, as in (9a). As estimated from Tsay and Myers's corpus, the average number of monosyllabic homophones is 2.1, and the average number of morphemes that a lexical tone may denote is 585.4. These values represent the x values for the USELISTED constraints for monosyllabic allomorphs and tonal allomorphs, respectively. The x value for the USELISTED constraints for disyllabic words is naturally 1. We can then calculate the B_{Listed} values for these constraints accordingly, as in (9b). The intuition behind this bias coefficient is that if USELISTED is the learner's strategy to cope with exceptional patterns that cannot be captured by regular means, such as the MARKEDNESS » FAITHFULNESS ranking, then the learner is first of all cautious about positing exceptions, expressed in the model by assigning USELISTED constraints greater penalties if they deviate from the default ranking of 0; secondly, the learner is unwilling to treat massive amounts of data as exceptions, expressed in the model as greater penalties for USELISTED constraints that make generalizations.

(9) a. $B_{Listed} = 10^{-\frac{1}{1+e^{2-2x}}}$, x = the number of morphemes that the type of USELISTED constraints covers on average

b.

	x	B_{Listed}
USELISTED(σσ)	1	$10^{-0.5}$=.316
USELISTED(σ)	2.1	$10^{-0.9}$=.126
USELISTED(T)	585.4	10^{-1}=.1

The $B_{Duration}$ coefficient expresses a substantive bias *à la* Wilson (2006). It biases against the duration increasing sandhi 51 → 55 by having a value of .95, but encourages the duration reducing sandhi 33 → 21 by having a value of 1.05. Duration neutral sandhis 55 → 33, 24 → 33, and 21 → 51 all have a $B_{Duration}$ of 1.

The σ^2 values for all constraints are summarized in Table 2.

Constraint	σ^2	Constraint	σ^2
UsLstd(σ55-σ)	.000316	UsLstd(σ21)	.000126
UsLstd(σ24-σ)	.000316	UsLstd(55)	.0001
UsLstd(σ33-σ)	.000332	UsLstd(24)	.0001
UsLstd(σ51-σ)	.000300	UsLstd(33)	.000105
UsLstd(σ21-σ)	.000316	UsLstd(51)	.000095
UsLstd(σ55)	.000126	UsLstd(21)	.0001
UsLstd(σ24)	.000126	*24-Nonfinal	.001
UsLstd(σ33)	.000132	Ident(Tone)	.001
UsLstd(σ51)	.000120		

Table 2: σ^2 values for all constraints.

5.4. Modeling the speakers

We turn to the modeling of learning of our two groups of speakers in this section. Our modeling was conducted with the MaxEnt learner in OTSoft 2.2 (Hayes et al. 2005). To model the Taiwan speakers, we fed the MaxEnt learner a training dataset of AO-AO input-output pairs that approximates the token frequency distribution of the different sandhi patterns: 13,000 /55-33/→[33-33], 12,000 /24-33/→[33-33], 11,000 /33-33/→[21-33], 10,000 /51-33/→[55-33], and 9,000 /21-33/→[51-33]. To model the Kansas speakers, we simulated the lack of frequency effects by feeding the learner an equal number of AO-AO input-output pairs (10,000) for each sandhi pattern. For each input, we considered 25 candidates that represent all possible tonal combinations. The two grammars learned by the MaxEnt model are given in Table 3.

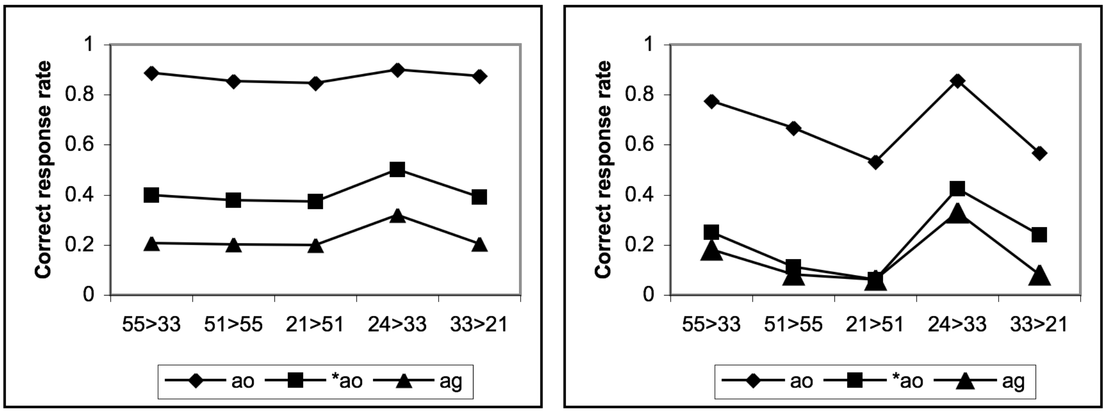
We can observe that for both groups of speakers, the UseListed constraints for real disyllables have the greatest weights, followed by *24-Nonfinal and Ident(Tone). UseListed for monosyllabic allomorphs and tonal allomorphs have lower weights, *even though they are never violated by any of the outputs in the training data.* Within each group of UseListed constraints, for Taiwan speakers, the sandhi that has a higher token frequency has a greater weight, while for Kansas speakers, the sandhi that has a greater duration reduction has a greater weight; the only exception is that the constraints governing 24 → 33 have the lowest weights within the group, as *24-Nonfinal is able to share some of the burden of enforcing the tone sandhi.

Constraint	**Weight (Taiwan)**	**Weight (Kansas)**	**Constraint**	**Weight (Taiwan)**	**Weight (Kansas)**
UsLstd(σ55-σ)	2.958	2.782	UsLstd(σ51)	.875	.873
UsLstd(σ33-σ)	2.878	2.815	UsLstd(σ21)	.866	.882
UsLstd(σ51-σ)	2.745	2.747	UsLstd(σ24)	.755	.730
UsLstd(σ21-σ)	2.708	2.782	UsLstd(55)	.736	.701
UsLstd(σ24-σ)	2.675	2.556	UsLstd(33)	.721	.707
*24-Nonfinal	1.892	1.868	UsLstd(51)	.696	.694
Ident(Tone)	1.755	1.740	UsLstd(21)	.688	.701
UsLstd(σ55)	.926	.882	UsLstd(24)	.600	.580
UsLstd(σ33)	.907	.890			

Table 3: Constraint weights for Taiwan and Kansas speakers learned by the MaxEnt model.

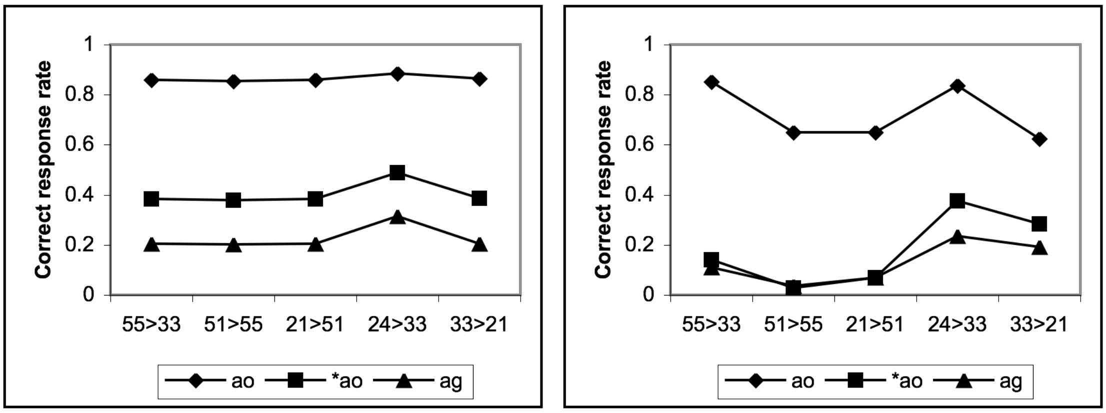
To compare the grammars' predictions with the speakers' wug test behavior, we calculated the correct response rates of AO, *AO, and AG inputs based on each grammar and juxtaposed the predictions with the outcomes of the wug test for the two groups of speakers, as in Figure 4 and Figure 5.

For both groups of speakers, the model successfully captures the overall low productivity of the sandhis in wug words despite the exceptionless input to the learner; it also captures the higher productivity of 24 → 33 than the opaque sandhis in the tone circle. Among the opaque sandhis, the predictions for both groups of speakers are also well matched with the wug test results: for the Taiwan speakers, the predicted order of productivity mirrors the order of token frequency, with 55 → 33 being the highest and 21 → 51 being the lowest; for the Kansas speakers, however, the predicted order of productivity follows a durational pattern, with the duration reduction sandhi 33 → 21 being the most productive and the duration increasing sandhi 51 → 55 being the least productive. The only prediction that the grammars fail to make is the magnitudes of the frequency and duration effects in Taiwan and Kansas speakers, respectively: the predicted effects are considerably smaller than the attested effects. However, regression analyses indicate that the grammars' predictions are significantly correlated with the experimental results. Taiwan speakers: R^2=.866, adjusted R^2=.855, β=.930, p<.001; Kansas speakers: R^2=.879, adjusted R^2=.869, β=.937, p<.001.



(a) (b)

Figure 4: The grammar's predictions (a) and the wug test results (b) for the Taiwan speakers.



(a) (b)

Figure 5: The grammar's predictions (a) and the wug test results (b) for the Kansas speakers.

5.5. Model summary

Our model sets out to take on the underlearning, overlearning, and proper learning challenges posed by the wug test data. The underlearning of exceptionless opaque sandhi patterns in the lexicon is achieved by deriving the opaque patterns via USELISTED constraints that directly state the input-output mappings of the sandhis and positing a learning bias against these constraints in the form of lower σ^2s. The magnitude of the effect predicted by the model is well matched with the experimental results. The overlearning of the phonetic knowledge despite the lack of information from lexical statistics is modeled by a substantive learning bias against phonetically unmotivated patterns, again in the form of lower σ^2s, *à la* Wilson (2006). The model succeeds in predicting the existence of this effect, but the predicted effect size is small compared to the experimental results. The proper learning of lexical statistics happens naturally for the learner, as the MaxEnt model inherently encodes the frequency effects by letting more frequent patterns to have greater pulls on constraint weights. But the magnitude of the effect predicted by the model is again small compared to the experimental results.

The small size the phonetic effect in the model is a direct result of the small frequency effect. Since the model needs to ensure that for Taiwan speakers, the phonetic effect is overshadowed by the frequency effect, it is not able to assign greater substantive biases according the durational property of the sandhis; in particular, the substantive bias cannot be greater than the frequency effect derivable by the model from token frequency exposure. What the model crucially needs, then, is a mechanism that predicts not only the existence, but also the correct magnitude of the frequency effect observed from the wug test.

The lack of frequency effect for the Kansas speakers is captured in the model by feeding it learning data that provide no frequency differentiations among the different sandhis. We recognize that a more comprehensive model should also have an explicit mechanism that captures the attrition of the frequency effect due to lack of usage and the corresponding constraint weight adjustments in the grammar that reflect the usage change. We take the last two points of discussion here as directions of future effort to improve our model.

6. Conclusion

Our wug test study on the tone sandhi patterns in Taiwanese showed that gradient factors such as phonetic duration and lexical frequency interact with formal factors such as opacity in meaningful ways in influencing the productivity of phonological processes. For Taiwanese tone sandhi *per se*, opacity outweighs phonetics and frequency as a global effect, while frequency and phonetic effects are dependent on the speakers' usage of the language in that frequency effects are evident for everyday users of the language, while phonetic effects only surface for occasional users for whom the frequency effects have been weakened due to the lack of use of the language. We have taken the position that a more fruitful approach to phonological grammar is to incorporate both processing and phonological factors to predict the observed behaviors of speakers with different usage backgrounds, and we have shown that a Maximum Entropy grammar that encodes learning biases against lexical listing constraints and phonetically unmotivated patterns can model the simultaneous underlearning of exceptionless opaque patterns, overlearning of phonetic effects, and proper learning of lexical statistics by the two groups of Taiwanese speakers. Finally, an important methodological point also emerged from the study: in experimental phonology research, *where* the speaker comes from can have a significant impact on both the experimental result and its theoretical modeling, and we as researchers need to be aware of this potential effect and incorporate it in our analyses and modeling of phonology when appropriate.

References

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

Bybee, Joan L. 1985. Morphology: a study of the relation between meaning and form. Philadelphia, PA: Benjamins.

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

Chen, Matthew Y. 1987. The syntax of Xiamen tone sandhi. *Phonology Yearbook* 4: 109-150.

Della Pietra, Stephen, Vincent J. Della Pietra, and John D. Lafferty. 1997. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19: 180–191.

Goldwater, Sharon and Mark Johnson. 2003. Learning OT constraint ranking using a maximum entropy model. In Jennifer Spenader, Anders Eriksson, and Osten Dahl (eds.), *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*. 111-120.

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

Hayes, Bruce, Bruce Tesar, Colin Wilson, and Kie Zuraw. 2005. *OTSoft 2.2*, software package. http://www.linguistics.ucla.edu/people/hayes/otsoft/.

Hayes, Bruce and Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39.3: 379-440.

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 Linguistic Society, Chicago. 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.

Jäger, Gerhard. To appear. Maximum entropy models and stochastic Optimality Theory. In Jane Grimshaw, Joan Maling, Chris Manning, Jane Simpson, and Annie Zaenen (eds.), *Architectures, rules and preferences: a festschrift for Joan Bresnan.* Stanford, CA: CSLI Publications. Also ROA-625: http://roa.rutgers.edu/.

Klatt, Dennis H. 1975. Vowel lengthening is syntactically determined in connected discourse. *Journal of Phonetics* 3: 129-140.

Lafferty, John, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. *International Conference on Machine Learning. ICML'01.* San Francisco, CA: Morgan Kaufmann. 282-289.

Lin, Hwei-Bing. 1988. *Contextual stability of Taiwanese tones*. Ph.D. dissertation, University of Connecticut.

Lin, Jo-wan. 1994. Lexical government and tone group formation in Xiamen Chinese. *Phonology* 11: 237-275.

Martin, S. C., H. Ney, and J. Zaplo. 1999. Smoothing methods in maximum entropy language modeling. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1: 545-548.

Moder, Carol Lynn. 1992. *Productivity and categorization in morphological classes*. Ph.D. dissertation, SUNY Buffalo.

Oller, D. Kimbrough. 1973. The effect of position in utterance on speech segment duration in English. *Journal of the Acoustical Society of America* 54: 1235-1247.

Peng, Shu-Hui. 1997. Production and perception of Taiwanese tones in different tonal and prosodic contexts. *Journal of Phonetics* 25: 371-400.

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 L. Goldstein, D. H. Whalen, and C. Best (eds.), *Laboratory Phonology VIII, Varieties of Phonological Competence*. Berlin: Mouton de Gruyter. 81-107.

Sandel, Todd L. 2003. Linguistic capital in Taiwan: the KMT's Mandarin language policy and its perceived impact on language practices of bilingual Mandarin and Tai-gi speakers. *Language in Society* 32.4: 523-551.

Sanders, Nathan. 2001. Preserving synchronic parallelism: Diachrony and opacity in

Polish. In *Papers from the 37th Meeting of the Chicago Linguistic Society*. Chicago Linguistic Society, Chicago. 501-515.

Scott, Mandy and Hak-Khiam Tiun. 2007. Mandarin only to Mandarin-plus: Taiwan. *Language Policy* 6: 53-72.

Smolensky, Paul. 1986. Information processing in dynamical systems: foundations of Harmony Theory. In David E. Rumelhart and James L. McClelland (eds.), Parallel Distributed Processing: explorations in the microstructure of cognition. Cambridge, MA: MIT Press/Bradford Books.

Smolensky, Paul, and Géraldine Legendre. 2006. The harmonic mind: from neural computation to Optimality-Theoretic grammar. Cambridge, MA: MIT Press.

Tsay, Jane and James Myers. 2005. *Taiwanese spoken corpus*. National Chung Cheng University, Taiwan.

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.

Wightman, Collin W., Stefanie Shattuck-Hufnagel, Mari Ostendorf, and Patti J. Price. 1992. Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America* 91: 1707-1717.

Wilson, Colin. 2003. Experimental investigation of phonological naturalness. In *Proceedings of the 22nd West Coast Conference on Formal Linguistics*. Somerville, MA: Cascadilla Press. 533-546.

Wilson, Colin. 2006. Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science* 30.5: 945-982.

Zhang, Jie and Yuwen Lai. 2006. Testing the role of phonetic naturalness in Mandarin tone sandhi. In Jing Wang, Sabri Al-Shboul, and Pedro Mateo (eds.), *Kansas Working Papers in Linguistics* 28: 65-126.

Zhang, Jie and Yuwen Lai. 2008. Phonological knowledge beyond the lexicon in Taiwanese double reduplication. In Yuchau E. Hsiao, Hui-Chuan Hsu, and Lian-Hee Wee (eds.), *Interfaces in Chinese Phonology*. Academia Sinica, Taiwan. 183-222.

Zhang, Jie, Yuwen Lai, and Craig Turnbull-Sailor. 2006. Wug-testing the "tone circle" in Taiwanese. In *Proceedings of the 25th West Coast Conference on Formal Linguistics*. Somerville, MA: Cascadilla Proceedings Project. 453-461.

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: 277-316.