

## Statistical learning of Korean phonotactics\*

Hyesun Cho  
(Seoul National University)

**Cho, Hyesun. 2012. Statistical learning of Korean phonotactics.** *Studies in Phonetics, Phonology, and Morphology* 18.2. 339-370. As in many languages, Korean distinguishes phonotactically impossible sound sequences from phonotactically possible, but rare sequences. This paper examines three such cases in Korean: post-obstruent tensing, diphthong restrictions, labial-[i] sequences ([pi] but \*[wi]). Learning simulations of the Korean phonotactics were conducted using the phonotactics learning model by Hayes and Wilson (2008). It turns out that the resulting grammar sometimes fails to distinguish possible but rare sequences from impossible sequences, that is, rare possible forms sometimes scored worse than impossible sequences. Since the Hayes and Wilson model uses markedness constraints only, learning simulations that employs both markedness and faithfulness constraints (Goldwater and Johnson 2003) were also carried out. This seemingly does better, but the success of learning with faithfulness constraints is due to the proper level of generality provided by human linguists. (Seoul National University)

Keywords: Phonotactics, Korean, statistical learning, gradient acceptability, post-obstruent tensing, alternation, diphthong

### 1. Introduction

Phonological grammars differentiate between sequences that are phonotactically impossible (e.g. English onset [bd-]), and sequences that are possible even though they occur with low frequencies (e.g. English onset [θw-]). The term “phonotactically impossible” in this paper refers to sequences that are disallowed by phonotactics and unattested; “phonotactically possible” sequences are ones that are allowed regardless of actual attestation. A difference in grammatical well-formedness is reflected in differences in native speakers’ intuitive judgments of acceptability (for English onsets, [bd-] is unacceptable; [θw-] is acceptable), and the goal of a phonotactic learning model is to find a grammar that conforms to human intuitions.

Distinctions between impossible and possible, but low-frequency, sequences also exist in Korean. For example, the diphthong [wo] is phonotactically impossible whereas vowel sequences [oo] and [uo] are

---

\*I thank Adam Albright and Edward Flemming for their helpful advice. I also thank Michael Kenstowicz, Donca Steriade, Jongho Jun, Minhwa Chung, three anonymous reviewers, and the audience of SICOLI 2009 and the seminar at the Linguistics Department of Seoul National University. All remaining errors are my own.

\*\*The learning data and feature chart files used in this paper are available at the author’s homepage (<http://plaza4.snu.ac.kr/~hyesun/archive.html>)

phonotactically possible, though they occur with lower frequencies than other vowel sequences. That is, in the same context (before [o]), the segments [w], [o], and [u] have very small differences in frequency: zero or near zero. If there are only small differences in frequency, and if constraints are learned based solely on frequency, it is a challenge for a phonotactic learning model to learn a grammar that makes intuitively correct distinctions between possible and impossible sequences. In the Korean example, it is possible that a phonotactic learning model learns a constraint like  $*\{w,u,o\}\{o\}$ , because the segments [w], [u], and [o] have similar frequency in the context of [o]. However, the constraint  $*\{w,u,o\}\{o\}$  is an overly broad generalization; this constraint penalizes both impossible ([wo]) and possible ([uo],[oo]) sequences equally. The current paper presents three such cases in Korean and discusses the problems that arise in designing a statistical learning model that arrives at a grammar which makes distinctions between possible and impossible sequences according to native speakers' intuitions. Through learning simulations with varying conditions, I show that the success of a learning model depends on its ability to learn constraints with the appropriate level of generality.

The problem of constraint learning with an appropriate level of generality, however, has not been an issue in most phonotactic learning models to this point, because in most models, the constraints themselves do not have to be learned. Instead, constraints are formulated by linguists, and are assumed to be available to the learning model *a priori*. Learning involves only ranking given constraints or assigning weights to the given set of constraints. Constraints tailor-made by linguists guarantee a certain level of generality because the constraints are formulated by human intuitions. In contrast, the phonotactic learning model of Hayes and Wilson (2008) attempts to discover the constraints themselves based on surface forms in the training data. Given a large search space of possible constraints, the model adopts a generality heuristic that selects constraints with a larger class if the segments in that class are close enough to each other in frequency in the same context. In other words, the model has a learning bias toward constraints with larger natural classes.

Under this heuristic, the model makes overly broad generalizations when the frequency differences in the training data between possible and impossible sequences are very small. The problem is demonstrated by the results of our simulation of Korean phonotactics learning, using the Hayes and Wilson phonotactics learning model. The Hayes and Wilson model yields a phonotactic grammar consisting of markedness constraints learned based on surface forms in the given data. The problem of overly broad generalizations arises when a group of segments differ in their legality in the same context, but have only small differences in frequency. In the Korean example mentioned above, the model selects the constraint  $*\{w,u,o\}\{o\}$ , instead of the two constraints  $*\{w\}\{o\}$  and  $*\{o,u\}\{o\}$ , due to the generality heuristic. The constraint ( $*\{w,u,o\}\{o\}$ ) ends up penalizing

not just impossible forms ([wo]) but also possible forms ([oo], [uo]), resulting in high penalties for possible forms. Similar problems are found in post-obstruent tensing (POT) and labial-[i] sequences. In the problematic cases, possible sequences are sometimes predicted to be worse than impossible sequences. The counterintuitive predictions are due to constraints that equally penalize both possible and impossible sequences, which native speakers differentiate as unacceptable or acceptable. Thus, the resultant grammar fails to properly model the native speaker's acceptability judgments and lacks descriptive adequacy – a basic requirement of a grammar (Chomsky 1965).

To achieve an appropriate level of generality, I test two possible solutions: restricting possible natural classes by adjusting feature specifications and learning with faithfulness constraints. In the Hayes and Wilson model, the constraints and their weights are learned through the observed surface forms in the training data. The learned constraints are all markedness constraints. I apply a learning algorithm that uses both markedness and faithfulness constraints (Goldwater and Johnson 2003). It turns out that a proper level of generalization is also required for the faithfulness constraints.

## 2. Acceptability judgments in Korean

As in other languages, distinctions between phonotactically impossible and phonotactically possible, but low-frequency sequences exist in Korean. The first example is shown in (1a) and (1b). Korean has a three-way contrast in obstruents: lax, aspirated, and tense. Between vowels, lax-lax obstruent clusters (1a) are impossible because of post-obstruent tensing, whereas lax-aspirated sequences (1b) are possible but occur with low frequencies (lower than lax-tense sequences). Second, some glide-vowel sequences (referred to as “diphthongs”) (1c) are impossible. V-V sequences can freely occur, but some of the V-V combinations (1d) have low frequencies. Third, the labial glide [w] followed by [i] is impossible, whereas labial stops followed by [i] occur with low frequencies, as shown in (1e) and (1f). The sequences in (1a), (1c), and (1e) do not surface in Korean. On the other hand, those in (1b), (1d), and (1f) are among the rarest sequences in Korean (morpheme-internally in particular), though they can appear more frequently through morpheme concatenation or inflection. Native speakers of Korean find that the sequences in (1a), (1c), and (1e) are more unacceptable than those in (1b), (1d), and (1f).

- |                        |   |
|------------------------|---|
| (1) Impossible forms   | Possible forms  |
| (a) tt, pt, pp, ...    | (b) tt <sup>h</sup> , pt <sup>h</sup> , pp <sup>h</sup> , ... |
| (c) wi, wu, wo, ji, ji | (d) uo, oo  |
| (e) wi                 | (f) pi, p <sup>h</sup> i, p'i, mi                             |

In a sequence of two obstruents, if the second obstruent is underlyingly lax, it undergoes tensification, which is known as Post-Obstruent Tensing. This rule is very productive, exceptionless, and post-lexical. The rule is described as in (2).

(2) Post-Obstruent Tensing (POT) (Ahn 1998:111)

$$\begin{bmatrix} \text{-son} \\ \text{-asp} \end{bmatrix} \rightarrow [+tense] / [-son] \text{ \_\_\_\_}$$

POT targets the lax-lax sequences such as in (1a) only, and thus do not surface faithfully; they turn into sequences of lax-tense (tt', pt', pp'). This process changes lax consonants in C2 (the second obstruent in a CC cluster) into tense, for example, /akki/ [akk'i] 'musical instrument', /cap-ta/ [capt'a] 'to grasp'. However, aspirated consonants in C2 are allowed and remain unchanged, for example, /sɪkp<sup>h</sup>um/ [ʃɪkp<sup>h</sup>um, \*ʃɪkp'um] 'food'. The lax-aspirated sequences in (1b) do not undergo any alternation rules; they surface faithfully.

On the other hand, the sequences in (1c), (1d) and (1e), (1f) are not differentiated by the existence of alternations. Instead, static phonotactic restrictions filter out the impossible sequences in (1c) and (1e) while allowing the possible sequences in (1d) and (1f). For the glide-vowel sequences in (1c), no alternation rules have been established in the previous literature, though speakers may repair them, as in loanword adaptation (e.g., ji→i, [isɪt<sup>h</sup>i] 'yeast'). The illegal diphthongs have been rather regarded as static phonotactic restrictions (Lee 1996, Sohn 1987). Lee (1996) describes the restrictions on diphthong formation as follows: /j/ cannot be followed by /i, i/; /w/ cannot be followed by /u, o, i/, which rules out the diphthongs [ji], [ji], [wu], [wo], [wi], as listed in (1c). Sohn (1987) attributes the ban on [ji], [wu], and [wo] to the OCP constraint: adjacent elements under a single nucleus node cannot share the same [-back] or [+round] features. The OCP constraint, however, cannot ban [wi] and [ji]. Rather, the underlying structures of [wi], [ji] do not surface because the vowel [i] is deleted (Sohn 1987). This vowel is more vulnerable to vowel deletion than other vowels. Whereas the glide-vowel sequences in (1c) are totally banned, sequences of two vowels can freely occur, but some vowel sequences, in (1d), occur with low frequencies.

Comparing [wi] (1e) with the sequences in (1f) shows that before /i/, the labial glide [w] and the labial stops [p, p<sup>h</sup>, p', m] behave differently, though they both belong to the labial class. The labial stops are underattested (morpheme internally) but possible, but the labial glide never surfaces (\*wi). The sequence [wi] is banned by static phonotactic restrictions, and it is considered unacceptable. On the other hand, there are no alternation rules or phonotactic restrictions that rule out the sequences in (1f), except in word-initial position. The sequences in (1f) are not allowed in word-

initial position according to descriptions of Korean phonotactics (Lee 1996), except in loanwords ([p<sup>h</sup>iraŋs'i] 'France'). This is due to the historical labial assimilation process which changed /i/ to [u] after labials (\*[mil]>[mul] 'water'). This change applied to initial syllables obligatorily, and only optionally in later syllables, e.g., /kip'i-ta/ [kip'ida] or [kip'uda] 'happy' (Kim-Renaud 1974). This process probably considerably decreased the frequency of surface labial-[i] sequences in the language, but there is no alternation rule that applies the sequences in (1f) in modern Korean. Yet, the labial-[i] sequences are rare morpheme-internally. There are only a few native words that contain the sequences morpheme-internally ([cimin] 'thousand (archaic)', [kirəmiro] 'therefore'), and there are no Sino-Korean words containing the sequences. The labial-[i] sequences are found more frequently in verbal conjugations with labial-final stems followed by [i]-initial suffixes (/nam+inik'a/ [naminik'a] 'because one remains').

In summary, each of the three cases involves a group of segments that differ in their legality in the same context, but have only small differences in frequency. That is, after a lax stop, lax obstruents are impossible, whereas aspirated obstruents are possible, but much rarer than tense obstruents. Before the vowel [o], the glide [w] is impossible, but [u], [o] are possible and occur with low frequencies. Before the vowel [i], [w] is impossible, but labial stops are possible and occur with low frequencies. This poses a problem for a statistical learning model in differentiating the possible and impossible sequences. It turns out that the resulting grammar of our simulation fails to make intuitively correct distinctions between possible and impossible sequences in the cases described in this section.

### 3. Learning Algorithm

For the learning simulation of Korean phonotactics, I used the phonotactics learning model of Hayes and Wilson (2008). This model is different from other grammar learning models (Boersma 1997, Goldwater and Johnson 2003, Coetzee and Pater 2006, 2008) in two respects. First, the Hayes and Wilson model attempts to derive constraints from surface forms in the training data, whereas in the other models, constraint sets are assumed to be universal and available to the model *a priori*. Thus, in those models, linguists provide hand-crafted constraints to the model, and learning mainly involves discovering the rankings or weights of constraints that account for the training data the best. On the other hand, the Hayes and Wilson model does not posit any hand-crafted constraint set in advance. Instead, constraint learning is a part of the learning algorithm. The Hayes and Wilson learning algorithm thus involves finding *both* constraints and their weights. Second, whereas most learning models in OT assume faithfulness constraints as well as markedness constraints, the Hayes and Wilson model does not use faithfulness constraints. Only markedness constraints are learned from the relative sequence frequencies in the

surface forms, without consideration of the underlying forms. Hence, the model is not provided with input-output pairs. Hayes and Wilson intend to assess the phonological forms “simply for their phonological legality, not for their legality as derived from some particular input (Hayes and Wilson 2008: 5).” Thus, faithfulness constraints, which regulate the relation between input and output forms, are not used.

In the Hayes and Wilson model, the constraints are learned in the following manner. The user feeds the model the phoneme inventory and the feature specifications and specifies the number of feature matrices in the constraints. The model generates natural classes from the given feature system. The number of possible constraints grows exponentially depending on the number of natural classes. To effectively search in a large constraint space, two search heuristics are employed: accuracy and generality. The accuracy of a constraint is defined by the  $O[C_i]/E[C_i]$  ratio of constraint violations.  $O[C_i]$  is the number of observed violations of a constraint  $C_i$ , and  $E[C_i]$  is the number of expected violations of a constraint given the current grammar. The generality of a constraint is defined by its length (=the number of feature matrices): the shorter the constraint, the more general. If the lengths are the same, constraints with more general natural classes are considered to be more general. The generality of a class is determined by the number of segments in that class: the more segments, the more general. This generality heuristic is necessary to make simpler generalizations, but it turns out that the heuristic results in overly broad generalizations in the problem cases in Korean.

The weights of the constraints are learned based on the principle of maximum entropy. The weights are the set of vectors that maximize the probability of all the surface forms in the given learning data. The resulting grammar consists of a set of weighted markedness constraints. A “score” of a phonological form is the weighted sum of constraint violations. Thus, the higher the score, the worse the form is. The predicted well-formedness of a form is inversely correlated to the score.

#### 4. Learning simulation

##### 4.1 The method and training data

The UCLA Phonotactic Learner<sup>1</sup> that applies the phonotactics learning algorithm in Hayes and Wilson (2008) was used for the learning simulations I now report. The simulations were run five separate times. Each run resulted in a similar set of constraints. In this section, I discuss the constraints from one of the simulations where the total number of the learned constraints was 109.

---

<sup>1</sup> <http://www.linguistics.ucla.edu/people/hayes/>

#### 4.1.1 The training data

The training data were native and Sino-Korean words from “the word list for Korean learners” (Cho 2003) from the National Academy of the Korean Language. The word list was created based on the results of a study of Korean word frequency (Cho 2002). The corpus includes 3,404 nouns, 1,345 verbs, 376 adjectives, and all the other word classes. It contained 2,399 native words, 2,474 Sino-Korean words, 249 loanwords, and a small number of words composed of the combinations of native, Sino, or loan words.

For the simulation here, loanwords were excluded, and only native and Sino-Korean words and the combination of the two categories were used, a total of 5,702 words.<sup>2</sup> The corpus was transcribed in the Korean alphabet, so first they were romanized using the HCode Hangul Code Conversion software (Lee 1994), and then the graphemes were converted to phonetic transcriptions using a Perl script that applied a set of Korean phonological rules (adapted from a script written by Adam Albright). In this process, several phonological rules were applied such as coda neutralization, post-obstruent tensing, nasalization, and aspiration. Lateralization was consistently applied to all sequences of /ln, nl/ (to [ll]) despite possible variations. Other optional or variable rules (place assimilation, post-sonorant tensing) were not applied. Palatalization, which applies across morphological boundary, was disregarded, since the corpus lacks the morphological boundary information and it is not of our interest.

#### 4.1.2 The feature system

The Learner was fed the feature chart for Korean phonemes and allophones, as in (3). Both contrastive underspecification and privative underspecification were used for some consonantal features to minimize the number of possible natural classes and hence the search space<sup>3</sup>. Different feature specifications result in different natural classes, and it turns out that learning results are sensitive to feature specifications and natural classes. In Section 5.1, I show that feature specifications are crucial to getting the correct result. That is, when there are more than one possibility in specifying features, it can be the case that only one of them yields the correct results. For example, the features [asp] and [tense] are

<sup>2</sup> Similarly, Hayes and Wilson (2008) have conducted a whole-language analysis of Wargamay phonotactics. Their training data were about 6,000 words, consisting of approximately 950 vocabulary items and their nominal, verbal conjugations.

<sup>3</sup> Following an anonymous reviewer’s suggestion, a separate simulation was run with contrastive underspecification for both vowels and consonants. The resulting grammar was tested for the cases where vowels are relevant (diphthong restrictions and labial-[i] sequences). The similar patterns and problems arise as those from the feature system (3), with minor changes in constraint formulations and weights. Only the results from the feature chart given in (3) will be presented in this paper.

used to describe the three-way contrast in Korean stops and affricates (lax, aspirated, and tense). The nature of the three-way distinction is a much-debated issue (Kim and Duanmu 2004, Silva 2006). The relevant laryngeal features include glottal constriction [Glottal constriction] (Chomsky and Halle 1968:326-8) and vocal cords stiffness [stiff vocal cords] (Halle and Stevens 1971). Halle and Stevens assign [+stiff vocal cords] to both aspirated and tense series, so in much of the literature both aspiration and tense consonants are assigned [+tense], as opposed to [-tense] for lax ones (Kim 1965, Kim-Renaud 1974:109, Ahn 1985). However, depending on which laryngeal feature is in consideration, tense and aspirated consonants can be grouped together or separately. In Chomsky and Halle (1968), the feature [+Glottal constriction] is assigned to tense obstruents only, and [-Glottal constriction] is assigned to aspirated and lax obstruents. In the present paper I first adopt the feature system that specifies [+tense] to tense consonants only. Under this analysis the feature [+tense] is the reflex of [+Glottal constriction]. Lax obstruents are [-asp,-tense], aspirated ones are [+asp,-tense], and tense ones are [-asp,+tense], so aspirated and lax obstruents can be grouped together as a natural class through the specification of [-tense]. A problem arises under this grouping, so the feature system where both aspirated and tense are assigned [+tense] will be applied in Section 5.1.

### (3) Feature chart

	p	p <sup>h</sup>	p'	t	t <sup>h</sup>	t'	c	c <sup>h</sup>	c'	k	k <sup>h</sup>	k'	s	ʃ	s'	h	m	n	ŋ	r	l	j	w	i	e	ɨ	ə	a	o	u
syl	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+
cons	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-
son	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+
cont	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+														
asp	-	+	-	-	+	-	-	+	-	-	+	-	-	-	-	-														
tense	-	-	+	-	-	+	-	-	+	-	-	+	-	-	+															
nas																	+	+	+											
lat																														
spread																	+													
lab	+	+	+															+						+						
cor				+	+	+	+	+	+				+	+	+				+		+	+								
ant				+	+	+	-	-	-				+	-	+				+		+	+								
strid				-	-	-	+	+	+				+	+	+				-		-	-								
dors										+	+	+								+										
high																							+	+	+	-	+	-	-	-
low																							-	-	-	-	-	-	+	-
back																							-	+	-	-	+	+	+	+
round																							-	+	-	-	-	-	-	+

For vowels, the seven vowel system of {i, ɨ, u, e, ə, o, a} was used. Since /æ/ is merged to [e] by younger speakers, /æ/ was converted to [e] in our



phonetic transcriptions. Diphthongs were treated as sequences of a glide and a vowel: /j/-diphthongs: /ja, jə, jo, ju, je/, /w/-diphthongs: /wa, wə, we, wi, wæ/, /i/-diphthong: /ij/. Because /wæ/ is pronounced as [we], /wæ/ does not exist in the phonetic transcriptions. Thus, there were four /w/-diphthongs [wa, wə, we, wi], five /j/-diphthongs [ja, jə, jo, ju, je], and one /i/-diphthong [ij].

#### 4.1.3 The learner settings

The length of constraints (the number of feature matrices) is set at 2 so that the algorithm searches constraints that consist of a maximum of two feature matrices. The maximum number of constraints to discover was unspecified, which means that the program stops when it finishes learning all the constraints that meet the highest O/E threshold. The highest O/E threshold is set at 0.30, the value used in most of the simulations in Hayes and Wilson (2008). Complement natural classes, which mean “any segment which is not this”, are set to be allowed in this simulation, and are indicated by “^”. For example, [^+high,-back,+syl] means “any segments other than [i]”.

### 4.2 The results

This section presents the learning results from the simulation described in Section 4.1. In Section 4.2.1, I illustrate the examples that show the general properties of the resulting grammar. In Section 4.2.2, the problem cases are illustrated.

#### 4.2.1 Gradient phonotactic acceptability

The resulting grammar consists of a set of markedness constraints, each assigned a weight. The model captures phonotactic constraints in a very detailed way, reflecting segment frequencies. Most of the known phonotactic restrictions in Korean are captured in the resultant grammar. The model also found phonotactic restrictions that are unknown. For example, in (4), the constraints #1-3 are already known in the literature, but those in #4-10 have not been noticed before.

The constraints #1-3 and #4-10 in (4) show the distinction between phonotactically impossible and possible rare forms. The segments in the constraints #1-3 ([ŋ, l, r]) are impossible to occur in word-initial position (native and Sino Korean). On the other hand, the segments in #4-10 in (4) may occur, but the frequencies are low. No phonological rules absolutely ban the segments in the constraints #4-10 to surface in word-initial position. The constraints only reflect the fact that these segments occur with low frequencies in word-initial position.

The model captured other underrepresented patterns that have not been

noticed in the previous literature. In (5), only the constraint #3 has been previously noticed in word-initial position, but there are other sequences that are predicted to be worse than the labial-[i] sequences (#1,2). The [c<sup>h</sup>i] sequence is underrepresented (#4), which seems to reflect the result of the historical process of vowel fronting (Middle Korean \*[ac<sup>h</sup>im]>[ac<sup>h</sup>im] ‘morning’).

(4) Constraints in word-initial position

	Constraint	Weight	Meaning
1.	*[+word_boundary][+lat]	4.774	*#l
2.	*[+word_boundary][+son,+dors]	4.68	*#ŋ
3.	*[+word_boundary][-lat]	4.01	*#r
4.	*[+word_boundary][-high,-back]	3.526	*#e
5.	*[+word_boundary][+high,+back,+syl]	2.146	*#u
6.	*[+word_boundary][-ant,+tense]	1.954	*#c’
7.	*[+word_boundary][-high,+round]	1.928	*#o
8.	*[+word_boundary][+tense,+lab]	1.9	*#p’
9.	*[+word_boundary][+cont,+tense]	1.767	*#s’
10.	*[+word_boundary][+asp,+dors]	1.584	*#k <sup>h</sup>

(5) Underrepresented sequences in Korean

	Constraint	Weight	Meaning
1.	*[-high,-back][+lat]	3.772	*el
2.	*[+asp,+dors][+high,+round,+syl]	2.81	*k <sup>h</sup> u
3.	*[+lab][+high,+back,-round]	2.641	*{p,p’,p <sup>h</sup> ,m,w}i
4.	*[-ant,+asp][+high,+back,-round]	1.994	*c <sup>h</sup> i
5.	*[+asp,+dors][-high,-low]	2.146	*k <sup>h</sup> {e,o,ə}
6.	*[+asp,+lab][-high,-low,-round]	2.987	*p <sup>h</sup> ə
7.	*[-cont,+tense,+cor][+high,+round,+syl]	1.899	*t’u

Some of the sequences in (5) are phonetically motivated. For example, the rarity of [k<sup>h</sup>u] in #2 might be explained as arising from the fact that it can be easily misperceived as [p<sup>h</sup>u], so [k<sup>h</sup>u] may have merged with [p<sup>h</sup>u] (cf. [kw] > [p], e.g. [akwa] (Latin) > [apə] (Romanian) ‘water’ (Bourciez 1967)). The labialized velar consonant [kw] became a plain labial [p] because the two sounds were similar (Flemming 2002:138). Confusability may also have caused the historical change in Korean where the unrounded vowel following a labial consonant becomes rounded in word initial position (\*pīl > pul, ‘grass’). However, in other cases, such as [el],

phonetic explanations are not available. The sequence [el] is found in loanwords, but rarely appears in native or Sino Korean words. The underrepresentation of such sequences seems to be rather a language-particular, arbitrary gap.

#### 4.2.2 The problem cases

All the 109 constraints learned by the phonotactic model were scrutinized for constraints that would penalize possible and impossible sequences together. This section presents the problem cases where the resulting grammar makes counterintuitive predictions, which the rest of the paper aims to solve.

##### 4.2.2.1 Post-obstruent tensing (POT)

An OT-style constraint for POT is \*[-son][tense,-asp]. However, this is not the exact constraint form that the model learned. The model is sensitive to segment frequency, so the effect of one process is often split among multiple constraints in a manner that reflects the frequency of individual segments. Thus, instead of learning one constraint as above, nine constraints were learned that in concert express the effects of POT. This is an important characteristic of the grammar learned by the Hayes and Wilson model. The constraints may look complicated and too specific from a phonologist's point of view. However, "ganging-up" or summation of such constraints, based on their assigned weights, leads to penalty scores for given test words. Ill-formed sequences can be penalized by one or two high-weighted constraints; but sometimes a sequence can be penalized by multiple constraints whose individual weights are not very high.

To test whether the resulting grammar makes the correct predictions about the sequences relevant to POT, I created test words and applied the resultant grammar to the test words to get the model's predicted acceptability of the words. The test words were created with all possible combinations of  $C1=\{p,t,k\}$  and  $C2=\{p,t,k,c,s,p^h,t^h,k^h,c^h,p',t',k',c',s'\}$ , in the form of "aC1C2a". There were 16 constraints that were violated by at least one of the test words. Among them, nine constraints reflected the effect of POT, listed in (6).

## (6) POT constraints

	Constraint	Weight	C1	C2
1.	*[-son][-cont,-asp,-tense]	2.958	p,t,k	p,t,k,c
2.	*[-son][-tense,+dors]	2.579	p,t,k	k,k <sup>h</sup>
3.	*[-son][+ant,-tense]	2.458	p,t,k	t,t <sup>h</sup> ,s
4.	*[-son,+lab][-cont,-tense]	2.277	p	p,t,k,c,p <sup>h</sup> ,t <sup>h</sup> ,k <sup>h</sup> ,c <sup>h</sup>
5.	*[-son,+cor][-tense]	1.889	t	p,t,k,c,p <sup>h</sup> ,t <sup>h</sup> ,k <sup>h</sup> ,c <sup>h</sup> ,s,f
6.	*[-son,+dors][-tense,+lab]	1.791	k	p,p <sup>h</sup>
7.	*[-son,+dors][-cont,-ant,-tense]	1.729	k	c,c <sup>h</sup>
8.	*[-son][+ant,-asp,-tense]	1.688	p,t,k	t,s
9.	*[-son,+lab][-tense,+lab]	0.86	p	p,p <sup>h</sup>

In (6), most sound groups in C2 in the learned constraints contain lax and aspirated obstruents together, except for constraints #1,8. As said, in C2 position, lax obstruents are impossible, while aspirated obstruents are possible but much less frequent than tense obstruents (as shown below in (7)). With such a frequency distribution, due to the generality heuristic, the model selects constraints with general classes that contain both lax and aspirated obstruents together. Note that this happens when the natural classes that include both lax and aspirated obstruents are available to the model. The current feature system specifies lax obstruents as [-tense, -aspirated] and aspirated obstruents as [-tense, +aspirated]. Under this feature system, [-tense] includes lax and aspirated obstruents. Since generality is defined by the number of segments in the class (the more segments, the more general), {t,t<sup>h</sup>,s} (in #3) is more general than {t,s} or {t<sup>h</sup>}. Given the bias for more general constraints, the model selects the constraint with {t,t<sup>h</sup>,s} instead of individual constraints with {t,s} and {t<sup>h</sup>} separately. The constraints that contain both lax and aspirated penalize possible and impossible forms equally. Thus, the penalty scores of some of the C2=aspirated sequences are sometimes overestimated.

Test results in (7) shows the scores and frequencies of sequences, categorized according to C2 (lax, asp, tense). Impossible forms (C2=lax) are all zero-frequency. The “score” of a form is the weighted sum of the form’s constraint violations. This is a penalty score that is inversely related to the form’s predicted acceptability (the higher the score, the less acceptable). The frequencies of possible but infrequent forms (C2=asp) range from 0 to 7, and those of frequent forms (C2=tense) range from 14 to 111. The optional place assimilation is not considered.

(7) Test results (Sorted by C2 category)

C2=lax	Score	Freq	C2=asp	Score	Freq	C2=tense	Score	Freq
<b>pt</b>	9.381	0	<b>pt<sup>h</sup></b>	6.96	0	<b>pt'</b>	0	111
<b>tt</b>	8.993	0	<b>tt<sup>h</sup></b>	4.348	0	<b>tt'</b>	0	83
<b>kk</b>	5.536	0	<b>kk<sup>h</sup></b>	2.579	0	<b>kk'</b>	0	100
<b>pp</b>	6.095	0	<b>pp<sup>h</sup></b>	3.137	1	<b>pp'</b>	0	16
<b>kt</b>	7.104	0	<b>kt<sup>h</sup></b>	2.458	3	<b>kt'</b>	0	68
<b>pc</b>	5.234	0	<b>pc<sup>h</sup></b>	2.277	4	<b>pc'</b>	0	19
<b>kp</b>	4.749	0	<b>kp<sup>h</sup></b>	1.791	5	<b>kp'</b>	0	17
<b>tc</b>	4.847	0	<b>tc<sup>h</sup></b>	1.889	7	<b>tc'</b>	0	23
<b>kc</b>	4.687	0	<b>kc<sup>h</sup></b>	1.729	7	<b>kc'</b>	0	63
<b>ts</b>	6.035	0				<b>ts'</b>	0	14
<b>ks</b>	4.146	0				<b>ks'</b>	0	72
<b>ps</b>	4.146	0				<b>ps'</b>	0	32

The prediction that the model will fail to distinguish the acceptability of sequences where C2 is tense or aspirated is borne out. This is demonstrated in (8). The sequences were listed in the order of predicted badness (decreasing order of score). The “Score” column shows the penalty scores, the model’s predicted acceptability value. The “Acceptability” column shows categorical judgments of acceptability by a Korean speaker (Y=acceptable, N=unacceptable). The Y’s are shaded.

(8) Test results (Sorted by penalty score)

C1C2	Score	Frequency	Acceptability	C1C2	Score	Frequency	Acceptability
pk	11.748	0	N	tt <sup>h</sup>	4.348	0	Y
pt	9.381	0	N	tp'	4.33	0	N
tk	9.038	0	N	ks	4.146	0	N
tt	8.993	0	N	ps	4.146	0	N
pk <sup>h</sup>	8.79	0	N	pk'	3.934	0	N
tk <sup>h</sup>	7.285	0	N	tp <sup>h</sup>	3.759	0	N
kt	7.104	0	N	tk'	3.672	0	N
pt <sup>h</sup>	6.96	0	Y	pp <sup>h</sup>	3.137	1	Y
tp	6.717	0	N	kk <sup>h</sup>	2.579	0	Y
pp	6.095	0	N	kt <sup>h</sup>	2.458	3	Y
ts	6.035	0	N	pc <sup>h</sup>	2.277	4	Y
kk	5.536	0	N	tc <sup>h</sup>	1.889	7	Y
pc	5.234	0	N	kp <sup>h</sup>	1.791	5	Y
tc	4.847	0	N	kc <sup>h</sup>	1.729	7	Y
kp	4.749	0	N	kt'	0	68	Y
kc	4.687	0	N	kk'*	0	100	Y

(\*The sequences under here are omitted. They are all lax-tense sequences with frequency higher than 14, Score=0)

The model predicts that [pt<sup>h</sup>] (6.96), [tt<sup>h</sup>] (4.348) are worse than [ks] (4.146), [ps] (4.146). However, this is counterintuitive because lax-aspirated sequences are possible and lax-lax sequences are impossible. For Korean speakers, all the lax-aspirated sequences, including [pt<sup>h</sup>] and [tt<sup>h</sup>], are more acceptable than all the lax-lax sequences such as [ks], [ps], because lax-aspirated sequences can surface through morpheme concatenation even though their morpheme-internal frequencies are low, whereas lax-lax sequences cannot occur at all. Although the sequences [pt<sup>h</sup>], [tt<sup>h</sup>] do exist in morpheme-internally but they are rare (they surface faithfully across morpheme boundaries, e.g. [[pap]<sub>N</sub>[t<sup>h</sup>ucəŋ]<sub>N</sub> 'complaining about meals'). Since it turns out that the training data did not contain any [pt<sup>h</sup>], [tt<sup>h</sup>] sequences, one may think that larger training data that contain these sequences would prevent this problem. An additional simulation was run where the sequences [pt<sup>h</sup>], [tt<sup>h</sup>] were added to the training data, which will be discussed in Section 5.1.2.

#### 4.2.2.2 Diphthong Restrictions

In Korean, the diphthongs (GV sequences) [wi], [wu], [wo], [ji], and [ji] are not allowed (Lee 1996, Sohn 1987). On the other hand, sequences of two vowels (VV) surface faithfully, except in a few diphthongization or

vowel deletion environments. Thus, VV sequences are phonotactically possible while some GV sequences are impossible. As in the POT case, the grammar fails to make intuitively correct distinctions between impossible GV sequences and possible but low frequency VV sequences.

The resultant grammar was tested by applying it to the test words. The test words were created by combining Seg1={w,u,o,i,j} and Seg2={i,i,u,o}, in the form of “Seg1Seg2n” strings (/n/ was added to avoid the irrelevant violations of word-final (non-absolute) restrictions found in the simulation, e.g. \*i#). In (9) were listed the constraints violated by any of the test words at least once. The constraints #1-7 penalize GV sequences as well as VV sequences, and those in #8-12 penalize VV sequences.

Note that the constraints in #1-7 penalize not just the diphthongs, but also sequences of two vowels. For example, in the constraint #5, \*[+round][-high,+round], the class [+round] contains {w,u,o} and the class [-high, +round] defines {o}. This constraint penalizes the union of these two classes, that is, [wo], [uo], and [oo]. Among them, the diphthong [wo] is phonotactically impossible, but [uo] and [oo] sequences are possible, and in fact, attested in existing words (e.g. [cuok] ‘jewelry’, [koon] ‘high temperature’<sup>4</sup>). However, it turns out that words containing [uo] and [oo] were not present in the learning data, reflecting the fact that [uo] and [oo] are the rarest among vowel sequences. The three segments [w], [u], and [o] thus accidentally have the same zero frequency before [o] which led the model to select the general class [+round] that contains all of them. In this way, the constraints #5-7 penalize impossible diphthongs and possible but rare vowel sequences equally, and thus, some rare vowel sequences have higher penalty scores than impossible diphthongs. For example, the vowel sequences [uo] (6.275), [oo] (6.196) have higher penalty scores than the impossible diphthongs [wo] (6.178), [ji] (5.692).

---

<sup>4</sup>In a larger corpus of 1.5 million words (Cho 2002), the frequencies of these words are 2 and 8 respectively.

## (9) Constraints for diphthongs and vowel-vowel sequences

	Constraint	Weight	Seg1	Seg2
	<i>GV, VV sequences</i>			
1.	*[-back,-syl][+high,-round,+syl]	2.992	j	i
2.	*[-cons,+lab][^round,+syl]	2.904	w	o,u
3.	*[+lab][+high,+back,-round]	2.641	p,p <sup>h</sup> ,p',m,w	i
4.	*[-cons,-syl][+high,+back,-round]	1.54	w,j	i
5.	*[+round][-high,+round]	1.36	w,u,o	o
6.	*[+high,-round][+high,-round,+syl]	1.16	i,j,i	i,i
7.	*[+high,+round][+high,+back,+syl]	0.473	w,u	u,i
	<i>VV sequences</i>			
8.	*[+syl][+syl]	2.341	vowel	vowel
9.	*[+high,+back][+round,+syl]	1.914	u,i	u,o
10.	*[-back,+syl][-high,+round]	1.729	i,e	o
11.	*[-high,+round][-high,-low]	1.567	o	e,ə,o
12.	*[+high,+syl][-high,+round]	0.66	i,u, i	o

## (10) Test results

	Score	Frequency	Acceptability		Score	Frequency	Acceptability
uo	6.275	0	Y	ii	3.501	7	Y
oo	6.196	0	Y	ui	2.814	6	Y
wo	6.178	0	N	oi	2.341	2	Y
ji	5.692	0	N	ou	2.341	4	Y
wu	5.291	0	N	ui	2.341	16	Y
io	4.73	1	Y	oi	2.341	16	Y
uu	4.728	1	Y	iu	2.341	18	Y
wi	4.654	0	N	ju	0	83	Y
ji	4.152	0	N	wi	0	111	Y
ii	3.501	1	Y	jo	0	191	Y

## 4.2.2.3 Labial-[i] sequences

Another problem case is found in the labial-[i] sequences. The constraint learned for the labial-[i] restriction is \*[+lab][+high,+back,-round], that is, \*{p,p',p<sup>h</sup>,m,w}{i}. The combinations of the segments in the constraints yield both possible and impossible sequences. The sequences of labial stops and [i] (pi, p'i,p<sup>h</sup>i, mi) are phonotactically possible, while the sequence of the labial glide and [i] (wi) is not possible. The model makes correct predictions on the relative well-formedness of these sequences,



when in isolation (e.g. #pi# is predicted to be better than #wi#). The model assigns a higher penalty score for [wi] than for any other sequences with labial stops [pi, p'i, p<sup>h</sup>i, mi], because [wi] is multiply penalized by other diphthong constraints such as \*[+high,+round][+high,+back,+syl] (= \*{wu}{ui}), which do not penalize labial stops. However, because the constraint \*{p,p',p<sup>h</sup>,m,w}{i} contains natural classes whose members are heterogeneous, there may be cases where the model's predictions diverge from speakers' intuitions. Such cases are found when the labial sequences are embedded in a string of more than two segments. The test words were "ak{p,p',p<sup>h</sup>,m,w}in" and "{a,i}{p,p',p<sup>h</sup>,m,w}in". The constraints violated by the test words are shown in (11).

## (11) Relevant constraints

	Constraint	Weight	Meaning
1.	*[-son][+cons,+son]	4.997	Nasalization
2.	*[-son][-cont,-asp,-tense]	2.958	POT
3.	*[+lab][+high,+back,-round]	2.641	*[+lab][i]
4.	*[-low][+tense]	2.511	
5.	*[+back][+tense,+lab]	2.399	
6.	*[-son,+dors][-tense,+lab]	1.791	POT
7.	*[-cons,-syl][+high,+back,-round]	1.54	*{w,j}{i}
8.	*[+high,+round][+high,+back,+syl]	0.473	*{w,u}{u}

## (12) Test results

				*[+lab] [+high,+back,-round]	*[-low] [+tense]	*[+back] [+tense,+lab]	*[-cons,-syl] [+high,+back,-round]	*[+high,+round] [+high,+back,+syl]
	Word	Score		2.641	2.511	2.399	1.54	0.473
1.	ip'in	5.152	Y	1	1			
2.	ap'in	5.04	Y	1		1		
3.	akwin	4.654	N	1			1	1
4.	awin	4.654	N	1			1	1
5.	iwin	4.654	N	1			1	1
6.	akp'in	2.641	Y	1				

The results in (12) show the results for some of the test words that are relevant to our discussion. As predicted, some of the possible words are

assigned penalty scores higher than impossible words. Possible surface forms (1,2) are scored worse than impossible forms (3,4,5). If the constraint  $*[\text{lab}][\text{i}]$  had treated the labial stops and the labial glide differently, this mismatch would have been avoided. The frequencies of the sequences of interest in the training data are shown in (13). The constraint with the more general class  $\{p, p', p^h, m, w\}$  has been selected instead of the two separate constraints  $\{p, p', p^h, m\}$  and  $\{w\}$ , because the frequency differences of the labial stops and the labial glide are very small and the model has a bias toward larger classes.

(13) Frequency of labial-[i] sequences

	Frequency	Acceptability
$p^i$	0	Y
$p^h i$	6	Y
$p' i$	5	Y
$m i$	5	Y
$w i$	0	N

To improve the result, the constraints that penalize possible and impossible forms together should be avoided. If the current constraint,  $*[+\text{lab}][+\text{high}, +\text{back}, -\text{round}] (= \{p, p', p^h, m, w\})$ , can be replaced by two constraints (i.e.  $\{p, p', p^h, m\}$  and  $\{w\}$ ), the results will become better. Since the frequencies of the labial stops before [i] range from 0 to 6, whereas the frequency of [wi] is zero, the two constraints ( $\{p, p', p^h, m\}$  and  $\{w\}$ ) may have different weights.

In sum, in all the three problem cases, the model selects the general class that contains segments some of which are allowed and others of which are disallowed (bold-faced) in the same context, as shown in (14). This is an undesirable effect of the generality heuristic employed by Hayes and Wilson (2008).

(14) Constraints with overly general classes

	Constraint	Impossible	Possible
Post-Obstruent Tensing	$\{p, t, k\} \{t, s, t^h\}$	pt, tt, kt, ps, ts, ks	pt <sup>h</sup> , tt <sup>h</sup> , kt <sup>h</sup>
Diphthongs vs. Vowels	$\{w, o, u\} \{o\}$	wo	oo, uo
Labial-[i] sequences	$\{p, p^h, p', m, w\}$	wi	

## 5. The solutions

### 5.1 Adjusting feature specifications

I have suggested some possible solutions to the problem in the previous sections. One is to provide more data, i.e. to provide missing but existing sequences. We can also adjust feature specifications if an alternative specification is available, such as the case of POT. In order to compare the

effects of the two (feature adjustment and additional data), three simulations were conducted with the conditions varied as in (15).

- (15)a. Data: More data, with the feature specifications (16a)  
 b. Classes: The feature specifications (16b), without adding more data  
 c. Data and classes: More data, and the feature specifications (16b)

In (15a), only the data were varied. In (15b), only the classes were varied. In (15c), both the data and the classes were varied. The added data were the words that contain [pt<sup>h</sup>] and [tt<sup>h</sup>] sequences: [apt<sup>h</sup>a], [att<sup>h</sup>a], the frequency of each was 1, reflecting their rarity. The alternative feature specification can be tried for POT, which is shown in (16).

(16) Alternative feature specifications

	(a) Previous		(b) Alternative	
	[asp]	[tense]	[asp]	[tense]
Lax	-	-	-	-
Aspirated	+	-	+	+
Tense	-	+	-	+

By the feature specification in (16a), lax and aspirated obstruents are grouped together as [-tense], which allows classes such as {t,t<sup>h</sup>}. The alternative feature system in (16b) does not allow such grouping of lax and aspirated obstruents. Under (16b), aspirated and tense obstruents are grouped together by [+tense] while separating lax obstruents, so classes such as {t,t<sup>h</sup>} are not available to the model.

The alternative feature specification in (16b) has an independent motivation as follows. It has been suggested that both aspirated obstruents and tense obstruents are specified as [+tense] (Kim 1965, Kim-Renaud 1974:109, Ahn 1998). According to this view, lax obstruents are not grouped together with aspirated or tense obstruents. This is phonetically motivated. First, producing tense and aspirated obstruents both involves stiff vocal cords. Second, the vowels following aspirated or tense obstruents are produced with higher pitch than those following lax obstruents (Jun 2000, Kenstowicz and Park 2006). Third, intervocalic voicing applies to lax obstruents, but not to tense or aspirated obstruents.

## (17) Simulation results

Previous results		(a) More data		(b) Proper classes		(c) More data & proper classes	
<b>pt<sup>h</sup></b>	6.96	<b>pt<sup>h</sup></b>	7.304	tt	9.155	tc	5.703
pp	6.095	pp	6.429	kt	8.054	ts	4.607
ts	6.035	kt	6.367	pt	8.054	kk	4.587
kk	5.536	kk	5.907	tc	5.676	kp	4.587
pc	5.234	pc	5.565	ts	4.58	kc	4.587
tc	4.847	tc	5.217	kk	4.575	pp	4.587
kp	4.749	kp	5.101	kp	4.575	pc	4.587
kc	4.687	<b>tt<sup>h</sup></b>	4.772	kc	4.575	ks	3.49
<b>tt<sup>h</sup></b>	4.348	ts	4.772	pp	4.575	ps	3.49
ks	4.146	kc	3.406	pc	4.575	kk <sup>h</sup>	3.31
ps	4.146	<b>pp<sup>h</sup></b>	3.023	<b>pt<sup>h</sup></b>	3.822	pp <sup>h</sup>	2.905
pp <sup>h</sup>	3.137	<b>kt<sup>h</sup></b>	2.961	ks	3.479	kt <sup>h</sup>	2.751
kt <sup>h</sup>	2.458	ks	2.961	ps	3.479	<b>tt<sup>h</sup></b>	2.751
pc <sup>h</sup>	2.277	ps	2.961	<b>tt<sup>h</sup></b>	3.383	<b>pt<sup>h</sup></b>	2.751

The table in (17) shows the results from the simulations under different conditions. The best results are obtained under the condition (15c), where both classes and data were changed. The overall results suggest that more data would not necessarily improve the previous results. Instead, using proper natural classes is a prerequisite. The result under the condition (15a) is shown in (17a). Comparing with the previous result, we can see that just adding more data does not improve the results. [pt<sup>h</sup>] and [tt<sup>h</sup>] have even higher scores (by about 0.5) than those in the previous results, despite the fact that the sequences each now have a frequency of 1. Moreover, [pp<sup>h</sup>] and [kt<sup>h</sup>] are also predicted to be worse than [ks], [ps]. However, the worse result in (17b) should not be interpreted to mean that providing more data worsens the result. Rather, the worse result is due to the model's stochastic nature of the constraint selection. The set of constraints that the model learns is slightly different at each run. That is, the difference between the previous result and (17a) is only accidental. On a careful examination, the worse result in (17a) is because a constraint that penalizes [ks] and [ps], \*[-son][+ant,-asp,-tense], was not learned in that run, and thus [ks] and [ps] had low penalty scores. To confirm this, three more same simulations were run. The constraint \*[-son][+ant,-asp,-tense] was learned only once out of 4 separate runs, and only when this constraint was learned, we had the almost identical result as the previous result. Under the previous condition

(without more data), this constraint was learned once out of 5 separate runs. Therefore, we cannot say that the addition of the sequences makes the learning results worse, and of course, it does not improve the results, either.

On the other hand, under the condition (15b), the results get noticeably better. The result is shown in (17b). Only the sequence [pt<sup>h</sup>] remains worse than impossible [ks] and [ps], though all four sequences in (17c) still have zero frequencies. Even without any additional data, the scores of both [pt<sup>h</sup>] and [tt<sup>h</sup>] are much lowered (6.96 to 3.822, 4.348 to 3.383), and also the score of [tt<sup>h</sup>] is lowered below those of [ks], [ps]. Finally, under the condition (15c) the results become consistent with their grammatical status, as shown in (17c).

Comparing the results in (17b) and (17c) shows that the problems can be improved by providing more data to the model. However, proper natural classes are a prerequisite. Otherwise, adding more data does not necessarily improve the result, as shown in the results in (17a). Adding too many low-frequency sequences would distort the whole frequency profile of the language. Alternatively, using a bigger corpus can increase the total size of the training data. However, with larger data, rare sequences will remain relatively rare, since the grammar is learned from the relative frequencies. Instead, it is crucial to block improper natural classes from being learned before providing more data.

## 5.2 Learning with faithfulness constraints

Knowledge of alternations may improve the counterintuitive predictions about some of the rare sequences in the Korean case. Section 4 showed that the sequences [pt<sup>h</sup>], [tt<sup>h</sup>] had higher penalty scores than the sequences [ks], [ps], which is counterintuitive. If an alternation grammar allows [pt<sup>h</sup>], [tt<sup>h</sup>] to surface faithfully, and disallows [ks], [ps] to surface, it would mean that the alternation grammar makes better predictions than the phonotactic grammar. However, the Hayes and Wilson model does not account for alternations because it uses markedness constraints only. In their model, grammar learning is based on surface forms only, and no underlying forms are given. To examine the influence of knowledge of alternations, a model is applied that incorporates faithfulness constraints as well as markedness constraints, the Maximum Entropy model of Goldwater and Johnson (2003).

This model is chosen for two reasons. First, the model finds constraint weights in the principle of maximum entropy, as the Hayes and Wilson model does. Second, the model makes gradient predictions on well-formedness which other categorical learning algorithms cannot (such as Biased Constraint Demotion (Prince and Tesar 1999)). The Goldwater and Johnson model uses input-output mappings and both faithfulness and markedness constraints. The markedness constraints learned by the Hayes and Wilson model were imported, instead of using hand-crafted

markedness constraints. This is one difference from the simulations run in Goldwater and Johnson (2003) as well as in other works (Boersma 1997, Coetzee and Pater 2008). The markedness constraints of the Hayes and Wilson model are learned from surface forms, so the constraints are different from those created by linguists in that the machine-learned constraints minimize the influence of human intuitions in linguistic analyses. This offers an indirect way of examining the influence of human intuitions in the analyses.

### 5.2.1 Applying the Goldwater and Johnson model

#### 5.2.1.1 Post-Obstruent Tensing

To apply the Goldwater and Johnson algorithm, OTSoft (Hayes et al. 2003) was used. The markedness constraints were those obtained from the Hayes and Wilson model (all of them in (6)), and the faithfulness constraints ID(asp) and ID(tense) were added. Each sequence in (8) corresponds to an input-output mapping, a few examples are illustrated in (18). For each underlying sequence, there were three candidate surface forms where C2 varied among tense, aspirated, and lax obstruents. The frequency of each mapping was indicated in the “Frequency” column. The numbers of violations of the faithfulness and markedness constraints were marked under relevant constraints.

(18) A part of the input-output mappings file

Underlying	Surface	Frequency	ID(asp)	ID(tns)	*[-son] [-cont,-asp,-tense]	*[-son,+cor] [-tense]	*[-son] [+ant,-tense]	*[-son,+lab] [-cont,-tense]	*[-son] [-tense,+dors]	*[-son] [+ant,-asp,-tense]	*[-son,+lab] [-tense,+lab]	*[-son,+dors] [-tense,+lab]	*[-son,+dors] [-cont,-ant,-tense]
kt	kt'	65		1									
	kt <sup>h</sup>	0	1			1	1						
	kt	0			1	1	1			1			
kt'	kt	3											
	kt <sup>h</sup>	0	1	1		1	1						
	kt	0		1	1	1	1			1			

The grammar learned with faithfulness constraints under the Goldwater and Johnson learning algorithm selects the correct surface forms, still making gradient distinctions among the sequences [(19),(20)]. The model can make gradient predictions because each constraint in the resulting

grammar has a weight (a non-negative real number). In the tableaux negative integers are used to mark constraint violations (Keller 2006, Coetzee and Pater 2008).

(19)

Acc.	Weight	.113	.112	.110	.043	.029	.000	H
	/pt <sup>h</sup> /	ID (asp)	*[-son] [-cont,-asp, -tense]	*[-son] [+ant,-asp, -tense]	*[-son] [+ant,- tense]	*[-son,+lab] [-cont,-tense]	ID (tns)	
.041	pt <sup>h</sup>				-1	-1		-.072
-.041	pt'	-1					-1	-.113
-.335	pt	-1	-1	-1	-1	-1		-.407

'Acc.': Acceptability score (cf. Coetzee and Pater 2008), 'H': Harmony score

(20)

Acc.	Weight	.110	.043	.000	H
	/ks/	*[-son][+ant,-asp,-tense]	*[-son][+ant,-tense]	ID(tense)	
.110	ks'			-1	.000
-.110	ks	-1	-1		-.110

The harmony score of a candidate is the weighted sum of its constraint violations (the "H" column). Because constraint weights are non-negative and constraint violations are negative, the harmony score decreases as the form's constraint violations are more serious. Thus, the winner is the one that has the highest harmony score. Further, harmony scores are translated into acceptability scores, following the method used in Coetzee and Pater (2008) and Pater (2008) (the "Acc." column). The acceptability score of a form is the difference between the harmony score of that form and that of the best candidate among the rest, as defined in (21).

(21)  $\text{Acceptability}(x) = H(x) - H(y)$

Where y is the most harmonic candidate for the same input as x, and  $y \neq x$ .

According to this definition, the acceptability score of the winner is the difference between the harmony score of the winner and that of the best candidate among the losers. The acceptability score of a loser is the difference between the harmony score of that form and that of the winner. Therefore, the acceptability value of the winner is always positive, and that of a loser is always negative. In the Hayes and Wilson model, the predicted acceptability is directly equated with the penalty score, which is the weighted sum of the constraint violations of a surface form. In contrast, in the Goldwater and Johnson model, acceptability scores should be calculated from harmony scores.

The grammar learned by the Hayes and Wilson model wrongly predicted that [pt<sup>h</sup>] is worse than [ks]. In contrast, according to the grammar learned with the faithfulness constraints, [pt<sup>h</sup>] surfaces but [ks] does not, which reverses the incorrect predictions previously observed. As shown in (19) and (20), the grammar selects the correct surface forms.

#### 5.2.1.2 Diphthong restrictions

Although diphthong restrictions have been considered as static phonotactic restrictions in the previous literature, the possibility of learning them through alternations will be investigated. Once the alternation grammar is learned, it is examined whether the alternation grammar makes correct predictions where the grammar learned by the Hayes and Wilson model fails, more specifically, whether it allows [uo], [ou], [io], [uu] (the vowel sequences that have higher penalty scores than impossible diphthongs) to surface while disallowing impossible diphthongs to surface. In the following, I consider several possible sources of the input-output mappings in learning diphthongs restrictions through alternations. In Korean, a sequence of two vowels (VV) is diphthongized (GV) in the following environments:

(22) Diphthongization environment

- a. Stem-suffix context: Vowel-final verbal/adjectival stem + vowel-initial inflectional affix
- b. Vowel coalescence results in monophthongs or diphthongs. (Lexically conditioned)

Let us first consider (22a): the stem-suffix context. Any vowel can appear in verbal/adjectival stem final position, but only three vowels appear in the initial position of inflectional affixes: /i, ə, a/. Thus, the inflection process may provide us evidence with regard to diphthongs where the second vowel is /i, ə, a/ (i.e. the combinations of /w, j/ and /i, ə, a/. Among these, the learner may learn that [wa], [wə], and [jə] are well-formed but [wi], [ji] are not, because the former surfaces whereas the latter does not, as shown in (23).

(23) Evidence for \*wi, \*ji

a. /po + a/ → poa, pwa	(o+a → wa)	‘see + Imperative’
b. /tu + ə/ → tuə, twə	(u+ə → wə)	‘put + Imperative’
c. /ki + ə/ → kiə, kyə	(i+ə → yə)	‘crawl + Imperative’
d. /cu + imjə/ → cumjə, *cwimjə	(u+i → *wi)	‘give + Progressive’
e. /ki + imjə/ → kimjə, *kjimjə	(i+i → *ji)	‘crawl + Progressive’

Diphthongization in (23a)-(23c) is optional but does occur if the final syllable of the stem has an onset, e.g. /katu+ə/ → [kaduə], [kadwə] ‘lock +



Imperative'. [wi] and [ji] do not surface. Instead of diphthongization /i/ gets deleted [(23d),(23e)]. This can be evidence for learning the constraints \*wi, \*ji through alternations.

A problem of this analysis is that (23d)-(23e) can also be interpreted as more general /i/-deletion, instead of a ban on particular diphthongs. /i/-deletion is not limited to occur just after the vowel /u/ or /i/, but productive with all other vowels, too. /i/ is obligatorily deleted after any vowel-final stems, as shown in (24).

(24) /i/-deletion in suffix-initial position

se + imjən [semjən]	ei→e	'count + Subjunctive'
sə + imjən [səmjən]	əi→ə	'stand + Subjunctive'
pə + imjən [pəmjən]	oi→o	'see + Subjunctive'
t <sup>h</sup> a + imjən [t <sup>h</sup> amjən]	ai→a	'burn + Subjunctive'
camki + imjən [camkimjən]	ii→i	'lock + Subjunctive'

The deletion is also a mirror image process (Kim-Renaud 1977, Ahn 1998:191). The position of [i] does not matter. When a stem and a suffix create a vowel hiatus, it is always [i] that gets deleted, regardless of whether it belongs to the stem or the suffix, e.g. camki + a [camka] 'lock + Imperative'. This position-independent vowel deletion is also observed in other languages. In Afar, iV and Vi becomes V (Casali, 1998). Casali (1998) proposes a class of PARSE(F) constraints to account for this deletion pattern. High vowels are typically deleted in favor of non-high vowels, i.e. PARSE(-high) is ranked higher than PARSE(+high). Following this analysis, Korean [i]-deletion can be analyzed as in (12). Max is used instead of PARSE.

(25) \*VV, Max(-high), Max(-back), Max(+round) » Max(i)

As for \*wu, consider the following conjugations of /p/-irregular verbs. Underlying /p/ in /p/-irregular verbs becomes [w] after underlyingly long vowels, a process called /p/-extreme weakening (Ahn 1998, Kim-Renaud 1974). [wi] and [wu] do not surface as in (26). The diphthongs [wi] and [wu] undergo labial assimilation and glide deletion, showing the mappings wi→u (\*wi) and wu→u (\*wu).

(26) Evidence for \*wi and \*wu

/ko:p + ina/ 'beautiful but' (Kim-Renaud 1974:37, Ahn 1998)	
kow <u>i</u> na	/p/-extreme weakening
kow <u>u</u> na	Labial Assimilation (i→u/w__)
kou <u>u</u> na	Glide Deletion (w→ø/ __u)
[kouna] <sup>5</sup>	

<sup>5</sup>Phonetic realization of [ou], whether it is pronounced as [ou] or [owu], is not obvious.

Let us now consider the vowel coalescence. The processes in (27) include several vowel coalescence patterns, such as  $a+i \rightarrow e$  (/sai/ [se] ‘gap’) (Ahn 1998). There are two diphthongs created by vowel coalescence: [wi] and [we]. (ü and ö are pronounced as [wi] and [we] in younger generations). Vowel coalescence is optional and lexical. It occurs in only some of the native Korean words ( $oi \rightarrow \ddot{o}$  ( $\rightarrow$ [we]) ‘cucumber’), but not in borrowed words (‘boy’ [poi], \*[pwe]) and compounds.

- (27)  $u + i \rightarrow \ddot{u}$  (wi)       $o + i \rightarrow \ddot{o}$  (we)  
        $i + i \rightarrow i$                  $\ddot{a} + i \rightarrow e$   
        $a + i \rightarrow e$

In (28), all the input-output mappings found so far are listed. These mappings are fed to the Goldwater and Johnson model to obtain the alternation grammar. The markedness constraints are from the Hayes and Wilson model (shown in (9)). The faithfulness constraints were those in (28), in addition to ID(round), ID(syl), and Max(Glide).

(28) Input-output mappings

(a) Possible diphthongs	(b) Impossible diphthongs	(c) [i]-deletion
$u\ddot{a} \rightarrow w\ddot{a}$ $oa \rightarrow wa$	$ui \rightarrow u$ (*wi)	$i\ddot{i} \rightarrow i$ $e\ddot{i} \rightarrow e$
$ui \rightarrow wi$ $oi \rightarrow we$	$i\ddot{i} \rightarrow i$ (*ji)	$\ddot{a}i \rightarrow \ddot{a}$ $oi \rightarrow o$
$i\ddot{a} \rightarrow j\ddot{a}$ $io \rightarrow jo$	$wi \rightarrow u$ (*wi)	$\ddot{a}i \rightarrow \ddot{a}$ $i\ddot{a} \rightarrow \ddot{a}$
	$wu \rightarrow u$ (*wu)	$ia \rightarrow a$

The grammar is obtained as in the constraint ranking in (29). The resulting grammar makes correct predictions for all the given input-output pairs in (29). In the tableau in (29), the candidate [wi] is ruled out because it violates highly weighted markedness constraints. In addition, Max(i) is lower than the markedness constraints, so [i]-deletion takes place instead of diphthongization. The candidates [wi] and [wu] both are ruled out due to the multiple violations of the highly weighted markedness constraints.

However, the resultant grammar cannot capture the cases where sequences of two vowels surface faithfully. Morpheme-internally, VV sequences surface faithfully: e.g. [cuok], \*[cwok] ‘jewelry’. Learners need to have knowledge of grammatical categories because the surface realization of vowel sequences depends on the grammatical contexts. If the resultant grammar is applied to VV sequences, the grammar selects /uo/ for [u], /oo/ for [o], and /uu/ for [u]. This is shown in (30). Under this grammar, the sequences /uo/, /oo/, and /uu/ cannot surface faithfully because it would

---

However, the focus here is the existence of phonological contrasts, rather than the actual phonetic realization. There is an underlying contrast between /ua/ ‘elegant’ and /uwa/ (exclamation), but there is no minimal pair showing a contrast between /ou/ and /owu/ (with an illegal diphthong underlying).

violate the highly weighted constraints (e.g. \*VV and \*{ui}{uo} in (30)).

(29)  $wi \rightarrow u$

Acc.	Weight	.101	.085	.078	.076	.075	.074	.071	.000	.000	.000	.000	Harmony
	/w <sub>1</sub> i <sub>2</sub> /	*VV	*{wu}{ui}	*{wj}{i}	*{w}{ou}	*wi	*{ui}{uo}	Max(+rd)	Max(i)	ID(Syl)	Max(G)	ID(rd)	
.071	$\textcircled{u}_1$								-1	-1			.000
.071	$\textcircled{u}_2$										-1	-1	.000
-.185	ui	-1	-1							-1			-.185
-.238	wi		-1	-1		-1							-.238
-.235	wu		-1		-1		-1					-1	-.235
-.071	i							-1			-1		-.071

\*VV = [+syl][+syl]

\*{wu}{ui} = [+high,+round][+high,+back,+syl]

\*{wj}{i} = [-cons,-syl][+high,+back,-round]

\*wi = [+lab][+high,+back,-round]

(In all the subsequent tableaux, the constraints will be written in terms of segments rather than natural classes for legibility.)

(30)

Acc.	Weight	.101	.076	.074	.071	.069	.068	.000	H
	/uo/	*VV	*{w}{ou}	*{ui}{uo}	Max(+rd)	*{wuo}{o}	*{iui}{o}	ID(syl)	
-.241	$\textcircled{uo}$	-1		-1		-1	-1		-.312
.148	$\textcircled{u}$				-1				-.071
-.148	wo		-1	-1		-1		-1	-.219

Thus, with this alternation grammar, the counterintuitive predictions on the vowel sequences such as /uo/, /oo/ and /uu/ cannot be reversed, because the alternation grammar applies only in the limited context, and cannot be extended to outside the domain of application. This is, however, not a failure of faithfulness constraints *per se*. The incorrect result in (30) can be improved by implementing knowledge of where to apply the alternation grammar. For example, \*VV is weighted high in the diphthongization environment, but should have little weight in other contexts where VV sequences surface faithfully. Knowledge of grammatical categories is thus necessary for better results.

## 5.2.1.3 Labial-[i] sequences

For the labial-[i] sequences, the following words are used for input-output mappings.

- (31) Faithful mappings: ip'in, ipin, ip'un, ipun, ap'in, apin, ap'un, apun  
 Unfaithful mappings: akwin, awin, iwin (wi→u alternation)

The Hayes and Wilson model gave a higher penalty score for [ip'in] than for [akwin], which is counterintuitive ((12) in Section 4.2.2.3). The grammar learned with faithfulness constraints selects the correct surface forms for /ip'in/ and /akwin/. That is, the word /ip'in/ surfaces faithfully, but /akwin/ does not, as shown in (32) and (33) respectively. Thus, the grammar learned by the Goldwater and Johnson model makes correct predictions where the Hayes and Wilson model fails.

(32) The tableau for /ip'in/

Acc.	Weight	.098	.096	.003	.0000	H
	/ip'in/	ID (round)	ID (tense)	*[+lab] [+high,+back,- round]	*[-low] [+tense]	
.101	ip'in			-1	-1	.003
-.189	ipun	-1	-1			-.186
-.101	ipin		-1	-1		-.098
-.101	ip'un	-1			-1	-.098

(33) The tableau for /akwin/

Acc.	Weight	.098	.080	.080	.003	.000	H
	/akw <sub>1</sub> i <sub>2</sub> n/	ID (round)	*{wj} {i}	*{wu} {ui}	*[+lab] [+high,+back,- round]	Max (Glide)	
.098	aku <sub>1</sub> n					-1	.000
-.098	aku <sub>2</sub> n	-1				-1	-.098
-.163	akwin		-1	-1	-1		-.163

## 5.2.2 Faithfulness constraints, with overly specific natural classes

The learning model that uses both faithfulness constraints and markedness constraints has been successful where the faithfulness-free Hayes and Wilson model makes incorrect predictions. However, this section shows that the faithfulness model also fails when the faithfulness constraints are learned with improper natural classes. In the previous section, the markedness constraints came from the Hayes and Wilson model, but faithfulness constraints were hand-crafted. The hand-crafted constraints

already reflect human intuitions in linguistic analyses. For example, instead of specific constraints such as \*[-son][-tense,+dors], \*[-son][+ant,-tense] for POT, a linguist would make broader generalizations such as \*[-son][-tense]. Likewise, linguists use [asp] as a feature to enforce a faithfulness constraint such as ID(asp), instead of a more specific constraint, say, ID([asp],t) (“Do not change the [asp] value of [t]”). If faithfulness constraints are learned with overly specific natural classes, the grammar is not guaranteed to select a correct output form. Suppose a hypothetical condition where faithfulness constraints are learned with overly narrow natural classes. The hypothetical constraints are shown in (34).

- (34) “ID(asp)/p\_” learned with natural classes divided into  $\{t^h\}$  and  $\{^ht\}$
- ID([asp], $t^h$ )/p\_ : “Do not change the value of feature [asp] on  $t^h$  after [p]”
  - ID([asp], $^ht$ )/p\_ : “Do not change the value of feature [asp] on  $^ht$  after [p]”
- (“^X”: Segments other than X)

The constraint (34a) is a version of the positional faithfulness constraint ID(asp)/p\_ learned specific to the class  $\{t^h\}$ . The constraint (34b) is another positional faithfulness constraint that applies to the class complement of the class  $\{t^h\}$ , i.e.  $\{^ht\}$ . The input-output mapping for /pt/ is given as in (35). The mapping  $pt \rightarrow pt'$  has a frequency of 111, and the mappings  $pt \rightarrow pt^h$  and  $pt \rightarrow pt$  have frequency zero. The constraint ID([asp], $^ht$ )/p\_ is violated by the mapping  $pt \rightarrow pt^h$ . The constraint ID([asp], $t^h$ )/p\_ is vacuously satisfied because the UR is not / $pt^h$ /.

(35) Specifications for /pt/ in the input-output mappings file

UR	SR	frequency	ID([asp], $^ht$ )/p_	ID([asp], $t^h$ )/p_	ID(tns)
pt	pt'	111			1
	pt <sup>h</sup>	0	1		
	pt	0			

In the tableau in (36), the resultant grammar is applied to / $pt^h$ /. The grammar predicts the incorrect output form [pt'] for the input / $pt^h$ /. The highest-weighted constraint ID([asp], $^ht$ )/p\_ is not applied to / $pt^h$ /. The constraint ID([asp], $t^h$ )/p\_ has a zero weight, and cannot effectively penalize the incorrect output form [pt'].

The result in (36) verifies that the correct predictions previously made by the Goldwater and Johnson model is in fact due to the natural classes with a proper level of generalization in the provided faithfulness constraints. Otherwise, the model with faithfulness constraints cannot be more successful than the model without faithfulness constraints. It is not the

inclusion of faithfulness constraints, but the proper level of generalization offered by the faithfulness constraints that contributes to the success of the resultant grammar.

(36)

Acc.	weight	.113	.112	.110	.029	.043	.000	.000	H
	/pt <sup>h</sup> /	ID([asp], ^ <sup>h</sup> )/p _	*[-son] [-cont,-asp,-tense]	*[-son] [+ant,-asp,-tense]	*[-son,+lab] [-cont,-tense]	*[-son][+ant,-tense]	ID (tense)	ID([asp],t <sup>h</sup> )/p _	
-.072	⊗pt <sup>h</sup>				-1	-1			-.072
.072	⊙pt'						-1	-1	.000
-.216	pt		-1	-1	-1	-1		-1	-.288

## 6. Conclusion

It has been a long-standing assumption that constraints in OT are universal. The grammar learning involves ranking the given constraints or finding weights of the given constraints. The Hayes and Wilson model, on the other hand, is distinct from other models because it attempts to discover the constraints themselves. In statistical learning of the constraints themselves, it is not trivial to learn constraints that have the proper level of generality. The learned grammar fails to separate possible forms from impossible forms in some cases where the frequency differences are very small. Correct phonological representations are prerequisites for a phonotactic learning model that learns constraints from the training data. It is also shown that the success of grammar learning largely depends upon the proper level of generality achieved in the natural classes.

## REFERENCES

- AHN, SANG-CHEOL. 1998. *Korean Phonology*. Seoul: Hanshin Publishing.  
 BOERSMA, PAUL. 1997. How we learn variation, optionality, and probability. *IFA Proceedings* 21, 43-58.  
 BOURCIEZ, ÉDOUARD. 1967. *Éléments de Linguistique Romane*, Paris: Klincksieck.  
 CASALI, RODERIC. 1998. *Resolving Hiatus*. New York and London: Garland Publishing.  
 CHO, NAMHO. 2003. *Hangwukeo Hakseupyong Eohwi Seonceong Kjeolkwa Pogoseo*. The National Academy of the Korean Language.

- [<http://www.korean.go.kr/>]
- \_\_\_\_\_. 2002. *Hyeontae Kwukeo Sayong Pinto Cosa*. The National Academy of the Korean Language.
- CHOMSKY, NOAM. 1965. *Aspects of the Theory of Syntax*. Cambridge: MIT Press.
- CHOMSKY, NOAM and MORRIS HALLE. 1968. *The Sound Pattern of English*. New York: Harper and Row.
- COETZEE, ANDRIES W. and JOE PATER. 2006. Lexically ranked OCP-Place constraints in Muna. Ms. University of Michigan and University of Massachusetts, Amherst. ROA-141.
- \_\_\_\_\_. 2008. Weighted constraints and gradient restrictions on place co-occurrence in Muna and Arabic. Ms. (A slightly revised version to appear in *NLLT*)
- FLEMMING, EDWARD. 2002. *Auditory Representations in Phonology*. PhD Dissertation. UCLA. New York and London: Routledge.
- GOLDWATER, SHARON and MARK JOHNSON. 2003. Learning OT constraint rankings using a maximum entropy model. *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, 111-120.
- HAYES, BRUCE, BRUCE TESAR, and KIE ZURAW. 2003. "OTSoft 2.1," 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, 379-440.
- HALLE, MORRIS and KENNETH STEVENS. 1971. A note on laryngeal features. *Quarterly Progress Report* 101, 198-213.
- JUN, SUN-AH. 2000. K-ToBI (Korean ToBI) Labelling Conventions (version 3.1, in November 2000). UCLA. (<http://www.linguistics.ucla.edu/people/jun/ktobi/K-tobi.html>)
- KELLER, FRANK. 2006. Linear optimality theory as a model of gradience in grammar. In Gisbert Fanselow, Caroline Féry, Ralph Vogel, and Matthias Schlesewsky (eds.). *Gradience in Grammar: Generative Perspectives*. Oxford: Oxford University Press.
- KIM, CHIN-WOO. 1965. On the autonomy of the tensivity feature in stop classification (with special reference to Korean stops). *Word* 21, 339-359.
- KIM, MI-RYOUNG and SAN DUANMU. 2004. "Tense" and "lax" stops in Korean. *Journal of East Asian Linguistics* 13, 59-104.
- KIM-RENAUD, YOUNG-KEY. 1978. The syllable in Korean phonology. In C.-W. Kim (ed.). *Papers in Korean Linguistics*, 85-98. Columbia, SC: Hornbeam Press.
- \_\_\_\_\_. 1974. Korean Consonantal Phonology. PhD Dissertation. University of Hawaii.
- KIM-RENAUD, YOUNG-KEY. 1982. *i*-deletion in Korean. In the Linguistic Society of Korea (ed.). *Linguistics in the Morning Calm*. Seoul: Hanshin Publishing Co., 473-488.
- KENSTOWICZ, MICHAEL J. and CHIYOUN PARK. 2006. Laryngeal features

- andtone in Kyungsang Korean: a phonetic study. *Studies in Phonetics, Phonology and Morphology* 12.2, 247-264.
- LEE, HO-YOUNG. 1996. *Kwukeyo Eumseonghak*. Seoul:Taehaksa.
- LEE, JUNE-YUB. 1994. Hcode: Hangul code conversion program, version 2.1.
- LEHISTE, ILSE and GORDON PETERSON. 1961. Transitions, Glides, and Diphthongs. *The Journal of the Acoustical Society of America* 33.3, 268-77.
- PATER, JOE. 2008. Gradient phonotactics in Harmonic Grammar and Optimality Theory. University of Massachusetts, Amherst, ms.
- PRINCE, ALAN and BRUCE TESAR. 1999 Learning phonotactic distributions. Technical Report RuCCS-TR-54, Rutgers Center for Cognitive Science, Rutgers University, New Brunswick. ROA-353.
- SILVA, DAVID. 2006. Acoustic evidence for the emergence of tonal contrast in contemporary Korean. *Phonology* 23.2, 287-308.
- SOHN, HYANG-SOOK. 1987. On the representation of vowels and diphthongs and their coalescence in Korean. *Proceedings from the Annual Meeting of the Chicago Linguistic Society* 23, 307-323.

Hyesun Cho  
Department of Linguistics  
Seoul National University  
1 Gwanak-ro, Gwanak-gu, Seoul 151-742, Korea  
E-mail: ruby1004@snu.ac.kr

received: April 12, 2012  
revised: July 7, 2012  
accepted: August 17, 2012