Linguistic Research 38(1): 123-150
DOI: 10.17250/khisli.38.1.202103.006

A weighted constraint grammar analysis of word-medial syllabification in English*

Sung-Hoon Hong
(Hankuk University of Foreign Studies)

Hong, Sung-Hoon. 2021. A weighted constraint grammar analysis of word-medial syllabification in English. *Linguistic Research* 38(1): 123-150. This paper proposes a weighted constraint grammar analysis that best models English syllabification as reported in Eddington et al. (2013). After we examine the factors governing the syllabification of medial consonants in two-syllable words, a set of constraints are formulated to address these factors. Based on these constraints, an initial grammar is constructed, to which we submit distribution information about medial syllabifications obtained from Eddington et al.'s (2013) syllabification survey data. We then perform learning simulations using Exponential Noisy Harmonic Grammar, a weighted constraint grammar designed especially to deal with constraints with negative weight (Pater 2009). As a result of the learning, an output grammar is produced in which each constraint is assigned a numerical weight. From the output grammar, we generate predicted syllabifications, which are then compared with observed syllabifications to evaluate the success of the analysis. We measure the predictive performance of the proposed analysis in terms of Root Mean Square Error and R^2, both of which provide a positive assessment. **(Hankuk University of Foreign Studies)**

Keywords weighted constraints, Noisy Harmonic Grammar, medial syllabification, phonological legality, sonority, Root Mean Square Error, R^2

1. Introduction

Previous phonological and psycholinguistic studies have shown that several factors or principles dictate the syllabification of word-medial consonants in English. Some of the factors work directly on medial consonants so as to make sure that as many consonants as possible are assigned to the onset of the following syllable, provided that they ascend in sonority and are allowed to occur also at word-initial position (Pulgram 1970; Hooper 1972; Kahn 1976; Selkirk 1982; Clements and Keyser 1983; Clements 1990). Other

* This work was supported by the 2020 Research Grant from Hankuk University of Foreign Studies. I appreciate anonymous reviewers for their helpful comments and suggestions. Special thanks to William Hart for valuable feedback on the final draft of the paper.

factors are associated with the stress and vowel quality of the preceding syllable such that a medial consonant following a stressed and lax vowel is syllabified as the coda of the preceding syllable (Pulgram 1970; Hoard 1971; Hooper 1972; Kahn 1976; Bailey 1978; Selkirk 1982; Treiman et al. 1994; Hammond 1997). Yet another factor related to morphological boundaries ensures that a morphological boundary, if any, coincides with a syllable boundary (Selkirk 1982; Borowsky 1986; Wells 1990; Derwing 1992; Laeufer 1995; Eddington et al. 2013).

Two fundamental aspects of these syllabification factors particularly draw our attention. First, it is often the case that these factors have conflicting effects on syllabification. In *mattress* [mǽtrɪs], for example, onset maximization forces both medial consonants to be syllabified as the onset of the second syllable as in [mǽ.trɪs] (“.” for a syllable boundary), running counter to the observation that the lax vowel of a stressed should be closed by a coda consonant. On the other hand, if we separate the sequence *tr* into two syllables as in [mǽt.rɪs], then the stressed lax vowel is closed by a coda, but a violation of onset maximality is induced. Another aspect of the syllabification factors worth mentioning is that there are no absolutely “strong” factors, and it is highly likely that each factor has its own relative effect realized in some way, so that multiple syllablications are possible for a given input. As for the aforementioned example *mattress*, Eddington et al.’s (2013) syllabification survey reports that its syllabification splits almost evenly such that there are 11 subject responses for [mǽ.trɪs] and 10 for [mǽt.rɪs]. Conflicting, relative factors then result in variable, not categorical, syllabifications.

The purpose of the current paper is to develop an explanatory model that effectively deals with the variability of medial syllabification. The data for variable syllabifications are obtained from Eddington et al. (2013), for which this paper provides an analysis using a weighted constraint grammar, a version of Optimality Theory (Prince and Smolensky 1993/2004) based on numerical constraints. This analytic framework is chosen because the basic architecture of OT is suited for dealing with conflicting factors, and numerical constraints are primarily motivated to account for non-categorical phenomena (Boersma and Hayes 2001; Hayes and Londe 2006; Jesney 2007; Pater 2009; Coetzee and Pater 2008, 2011; Coetzee and Kawahara 2013; Coetzee 2016; Hayes 2017).

The remainder of this paper is organized as follows. Section 2 introduces the factors that affect medial syllabification and proposes OT-based constraints to handle these factors. In section 3, we will see how the current paper approaches variable syllabifications with Exponential Noisy Harmonic Grammar, a weighted constraint grammar which has been developed to cope with negative constraints (Pater 2009; Boersma and Pater 2016). We will then evaluate the predictive accuracy of the proposed analysis by comparing the observed and predicted syllabifications. Section 4 provides the discussion and conclusions of this paper.

2. Word-medial syllabification in English

2.1 Factors affecting English medial syllabification

Among the factors that affect medial syllabification, the following three well-known principles are first to be noted: Onset Maximization, the Sonority Contour Principle, and Phonological Legality. Onset Maximization (OM) ensures that as many consonants are placed as possible at the beginning of a syllable (Pulgram 1970; Hooper 1972; Kahn 1976; Selkirk 1982; Clements and Keyser 1983). OM is regulated by the Sonority Contour Principle (SCP), which asserts that sonority rises from the beginning of a syllable to the nucleus and falls from the nucleus to the end of a syllable (Hooper 1972; Selkirk 1982; Clements 1990). The third principle, Phonological Legality (PL), places a further restriction on OM by stipulating that a consonant or a sequence of consonants which occurs at the edge of a syllable must also occur at the same respective edge of a word[1] (Pulgram 1970; Hooper 1972; Selkirk 1982).

What should be noted here is that these principles are often in conflict with one another. Consider the medial sequences, -*tr*-, -*tl*-, and -*st*-, for example. The consonant cluster *tr* has a rising sonority slope and can occur at the beginning of a word. Thus, an onset can be maximized to include both *t* and *r* without violating PL or the SCP. However, the sequence *tl*, in which sonority also rises, does not occur in word-initial position, and thus maximizing the onset violates PL while observing the SCP. Lastly, the sequence *st* occurs as a word-initial consonant cluster, so maximizing the onset to include both *s* and *t* follows PL while violating the SCP, since the second consonant *t* is lower

[1] Phonologically *illegal* consonant clusters (i.e. clusters that are *not* allowed in word-initial or final position) fall into two types, depending on whether they conform to or violate the SCP: (i) clusters that also violate the SCP (e.g. initial clusters such as *nt*-, *lt*-, and *rt*-); and (ii) clusters that conform to the SCP (e.g. initial clusters such as *pm*- and *tn*-). The clusters of the first type are prohibited universally while the permissibility of the second type is determined on a language-particular basis. See Appendix for the lists of the phonologically legal and illegal two consonant clusters assumed in this paper.

in sonority than the first, *s*.

(1) Onset Maximization, Sonority Contour Principle, and Phonological Legality

CC onset	OM	SCP	PL
.tr	yes	yes	yes
.tl	yes	yes	no
.st	yes	no	yes

In addition to the above three principles, the stress and the vowel quality of the preceding syllable also affect medial syllabification. Most notably, stressed lax vowels tend to attract a following consonant into their syllables (Pulgram 1970; Treiman et al. 1994), yet others have observed that the stress and the vowel quality act separately, and that either being lax (Pulgram 1970) or stressed (Hoard 1971; Hooper 1972; Kahn 1976; Bailey 1978; Selkirk 1982; Hammond 1997) is sufficient to draw a following consonant.[2] Medial syllabification is also influenced by the presence of a morphological boundary such that the edges of a morpheme coincide with the edges of a syllable (Selkirk 1982; Borowsky 1986; Wells 1990; Derwing 1992; Laeufer 1995; Eddington et al. 2013). The effects of these factors on medial syllabification, along with those of the three principles introduced earlier, are illustrated in the following table.

[2] Some researchers note that a stressed vowel may even attract a preceding consonant as its onset (Fallow 1981; Treiman et al. 1994; Eddington et al. 2013). Since similar observations have not been reported in other studies, this effect of stress will not be considered further in this paper.

(2) Effects of factors/principles on medial syllabification

	Examples	Syllabifications[3]	OM	SCP	PL	Stress	Vowel quality	Morpheme boundary
a.	lemon /lɛmən/	[lɛ́.mən]	yes	yes	yes	**no**	**no**	NA[4]
		[lɛ́m.ən]	**no**	yes	yes	yes	yes	NA
b.	weapon /wɛpən/	[wɛ́.pən]	yes	yes	yes	**no**	**no**	NA
		[wɛ́p.ən]	**no**	yes	yes	yes	yes	NA
c.	mattress /mætrɪs/	[mǽ.trɪs]	yes	yes	yes	**no**	**no**	NA
		[mǽt.rɪs]	**no**	yes	yes	yes	yes	NA
		[mǽtr.ɪs]	**no**	**no**	**no**	yes	yes	NA
d.	atlas /ætləs/	[ǽ.tləs]	yes	yes	**no**	**no**	**no**	NA
		[ǽt.ləs]	**no**	yes	yes	yes	yes	NA
		[ǽtl.əs]	**no**	**no**	**no**	yes	yes	NA
e.	pack-ing /pæk-ɪŋ/	[pǽ.kɪŋ]	yes	yes	yes	**no**	**no**	**no**
		[pǽk.ɪŋ]	**no**	yes	yes	yes	yes	yes

As the final factor for medial syllabification, another aspect of sonority, what is called "the preferred sonority slope," must be considered. According to Clements (1990: 301), "the preferred syllable type shows a sonority profile that *rises maximally toward the peak and falls minimally towards the end*, proceeding from left to right" (cf. Hooper 1972, 1976; Murray and Vennemann 1983; Vennemann 1988). Since the sonority difference from onset to peak is more radical for obstruents than for sonorants, the preferred sonority slope stipulates that obstruents are preferred to sonorants for syllable-initial consonants, and sonorants are preferred to obstruents for syllable-final consonants.[5] Taking examples from (2) above, the preferred sonority slope opts for the syllabification [wɛ.pən] in (2b), with the obstruent *p* being the onset of the second syllable, and [lɛm.ən] in (2a), with the sonorant *m* being the coda of the first syllable.

Before we close this section, I would like readers to note two important aspects of medial syllabification. As we have seen above, it is most likely that syllabification factors conflict with one another, and if they do, that syllabification will eventually be

[3] Following Eddington et al. (2013), from which the data for this paper come, we do not consider ambisyllabicity as a syllabification option here. We take the stance that ambisyllabicity can be represented alternatively in terms of resyllabification (Selkirk 1982; Myers 1987) or foot structure (Kiparsky 1979; Hammond 1997).

[4] We mark 'not applicable (NA)' where the conditional clause is not met (i.e. the cells are marked NA because there is no morphological boundary in the examples). Later in the constraint-based analysis, cases like this will be treated as vacuous satisfaction.

[5] Treiman (1984) and Treiman and Danis (1988b) make the further observation that liquids, which are considered more sonorous than nasals among sonorants, are preferred to nasals in the coda.

determined by the more influential factors. In *lemon*, for example, OM conflicts with the factors regarding stress and vowel quality. If OM exerted a greater influence than the latter two, the syllabification would be [lɛ́.mən]; but if the stress or the vowel quality were more dominant than OM, the syllabification would be [lɛ́m.ən].

Another notable aspect of medial syllabification is that the resulting syllabifications are not categorical but variable, as revealed by a number of psycholinguistic studies (Treiman and Danis 1988a; Treiman and Zukowski 1990; Derwing 1992; Treiman et al. 1994; Smith and Pitt 1999; Redford and Randall 2005). According to Treiman and Danis (1988a: Experiment 2), for example, a medial consonant that is nasal and preceded by a stressed syllable, as in *lémon*, is syllabified as the onset of the second syllable (V.CV) in 32% of the responses. When the medial consonant is an obstruent and the stress in on the first syllable, as in *wéapon*, almost half of the responses, 49%, choose the V.CV syllabification. If the stress is on the second syllable, as in *guitár*, the V.CV syllabification predominates so that the medial obstruent is syllabified as the onset of the second syllable in an overwhelming 93% of the responses.

Any study dealing with medial syllabification should address these issues, and in what follows, we will pursue them with a weighted constraint grammar. The factor interactions will be approached with ranked constraints, which represent the fundamental tenet of OT grammar, and the variability of medial syllabification will be accounted for using numerical or weighted constraints.

2.2 Constraints for English medial syllabification

In this section we will see how the syllabification factors discussed above are formulated into OT constraints.[6] First of all, the effects of PL can be captured by the following two constraints:

(3) a. Legal(C-initial): A consonant or a sequence of consonants that occurs at the beginning of a syllable must also be licit at the beginning of a word.
b. Legal(C-final): A consonant or a sequence of consonants that occurs at the end of a syllable must also be licit at the end of a word.

[6] Since the complete constraint ranking will be determined by performing learning simulations in section 3, we will focus here on motivating the constraints, limiting discussion about ranking only where necessary.

PL is not just a condition on permissible consonants at each margin of a syllable, but can also apply to vowels in syllable-final position. Since the word-final position is where vowel lengthening typically takes place to result in tensing (Féry 2003), legality demands that a vowel in syllable-final position also be tense, giving rise to the following legality constraint:

(4) Legal(Vowel): Syllables may not end with a lax vowel.[7]

The next factor we will consider is stress. As we saw earlier, a stressed vowel tends to attract a following consonant as the coda, and this attraction is stronger when the stressed vowel is lax than when it is tense (Pulgram 1970; Treiman et al. 1994). On the assumption that lax vowels are monomoraic but tense vowels bimoraic (Hammond 1997), such effects can be explained by the interaction of the following three constraints.

(5) a. Stress-to-Weight: If stressed, then heavy. (Myers 1987; Kager 1999: 268)
b. *Superheavy: No trimoraic syllables are allowed.[8] (Kager 1999: 268)
c. Weight-to-Stress: Heavy syllables are stressed. (Prince and Smolensky 2004: 63)

Stress-to-Weight (STW) ensures that a stressed lax vowel is followed by a coda consonant, while *Superheavy inhibits coda attraction when the preceding vowel is tense. The last constraint, Weight-to-Stress (WTS), penalizes a stressed lax vowel without a coda consonant. The following table illustrates how the three constraints, together with Legal(Vowel), assess syllabifications.

[7] As for the tense-lax distinction in English, this paper follows Ladefoged (2006: 96) in which /ɪ, ɛ, æ, ʊ, ə, ʌ/ are distinguished from /i, eɪ, ɑ, ɔ, oʊ, u, aɪ, aʊ, ɔɪ, ju/ as lax and tense vowels, respectively (cf. Halle and Mohanan 1985; Giegerich 1992; Hammond 1997, 1999). For British English, Ladefoged (2006: 97) posits an additional lax vowel [ɒ] (as in *cod*, *common*, *con*). This classification is primarily motivated from distributional considerations: (i) only tense vowels are found in stressed open syllables; and (ii) only lax vowels are found in syllables ending in /ʃ/ or /ŋ/ (Ladefoged 2006; Yavaş 2011).

[8] This constraint bans a lax vowel followed by two coda consonants, as well as a tense vowel or a diphthong followed by a coda consonant.

(6)

		Legal(V)	STW	*Superheavy	WTS[9]
lemon	[lɛ́.mən]	*(lɛ́)	*(lɛ́)		*(mən)
/lɛmən/	[lɛ́m.ən]				*(ən)
demon	[dí.mən]				*(mən)
/dimən/	[dím.ən]			*(dím)	*(ən)

The first syllable of [lɛ́.mən] is stressed but ends in a lax vowel, violating Legal(Vowel) and STW. Both these constraints are satisfied in [dí.mən], whose first syllable with a tense vowel is stressed and heavy without a coda consonant.

Next, we will examine sonority-related factors. The first of these factors, the Sonority Contour Principle, is directly converted to the following constraint:

(7) Sonority Contour (SonContour): In any syllable, sonority must increase from the left-hand margin of the onset to the syllable peak, and decrease from the syllable peak to the right-hand margin of the coda (Selkirk 1982, Clements 1990,).

Here, we assume with Clements (1990) that sonority values for nonsyllabic segments are distinguished among the four major classes of obstruents (O), nasals (N), liquids (L), and glides (G), with a progressively higher sonority value assigned from O to L to N to G.[10]

Another sonority factor, the preferred sonority slope, suggests that there are markedness harmony scales for onset and coda. That is, at the left margin of a syllable, obstruents are most preferred, nasals the second, liquids the third, and glides are least preferred. At the right margin, the preference order is the opposite; liquids are most preferred, followed by nasals, with obstruents being least preferred.[11] Both of these preference scales are illustrated below in (8).

[9] The weight of the second syllable is evaluated for expository purposes only. In the remainder of the paper, we will consider only the part that is relevant for medial syllabification, that is, VCV, VCCV, or VCCCV (without considering the onset of the first syllable or the coda of the second syllable).

[10] Clements (1990: 292) proposes that the sonority scale is derived by taking the sum of the *plus*-specifications for each major class feature, with major class features being [syllabic], [vocoid], [approximant], and [sonorant]. A glide, for example, has three *plus*-specifications, [+vocoid, +approximant, +sonorant], and hence the sonority value 3 is assigned. A liquid has two, [+approximant, +sonorant], and thus its sonority value is 2. A nasal has one, [+sonorant], and hence its sonority value is 1. An obstruent has no *plus*-specifications, and thus 0 is assigned.

[11] Glides are not possible codas in English, and hence are not considered in the coda markedness scale shown in (8b).

(8) a. Onset: O > N > L > G
b. Coda: O < N < L

Following Prince and Smolensky (1993/2004), I assume that markedness harmony scales are mapped onto markedness constraint hierarchies such that the least-preferred structure is penalized by the highest-ranked constraint. The onset and coda markedness harmony scales in (8) thus yield the two ranked sets of markedness constraints holding at each syllable margin shown below (cf. Gouskova 2004: 209).

(9) Markedness constraint hierarchy at the left margin of a syllable:
$*_{\text{onset}}[\text{G} \gg *_{\text{onset}}[\text{L} \gg *_{\text{onset}}[\text{N} \gg *_{\text{onset}}[\text{O}$
(10) Markedness constraint hierarchy at the right margin of a syllable:
$*\text{O}]_{\text{coda}} \gg *\text{N}]_{\text{coda}} \gg *\text{L}]_{\text{coda}}$

It should be noted that the coda markedness constraints in (10) assign a finely tuned penalty to a syllable with a coda consonant, superseding NoCoda (Prince and Smolensky 1993/2004), which issues a uniform violation to all syllables with codas regardless of their type.

The next factor is OM, for which we need three sets of constraints. The first set of constraints are those which assign medial consonants to the onset of the following syllable rather than to the coda of the preceding syllable. A well-known syllable markedness constraint, Onset (Prince and Smolensky 1993/2004), and the coda and onset markedness constraints provided in (9, 10) will serve this purpose. As long as Onset dominates the onset markedness constraints in (9), these constraints prefer a syllable whose onset is maximized to contain all medial consonants, as shown below in (12a).

(11) Onset: Syllables must have onsets.

(12)

	zebra	Onset	$*_{\text{ons}}[\text{L}$	$*_{\text{ons}}[\text{O}$	$*\text{O}]_{\text{coda}}$	$*\text{L}]_{\text{coda}}$
a.	☞[zi.brə]			*(b)		
b.	[zib.rə]		*!(r)			*(b)
c.	[zibr.ə]	*!			*(r)	

However, there are cases where a medial consonant cluster splits into two syllables. In such cases, syllabification is further moderated by Legal(C-initial) and SonContour as the following tableaux demonstrate. Here, maximized onsets always violate one or both of these constraints (13a, c, e), but syllables with a split CC incur no violation (13b, d, f).

(13)

		Legal(C-init)	SonContour
a.	te.mper	*(mp)	*(mp)
b.	tem.per		
c.	ki.dney	*(dn)	
d.	kid.ney		
e.	cry.stal		*(st)
f.	crys.tal		

All these constraints combined together still do not properly deal with a medial cluster composed of three or more consonants. In cases like *textile*, for example, we cannot distinguish [tɛk.staɪl] from [tɛks.taɪl], since both forms satisfy Legal(C-initial) but violate SonContour, as well as the onset/coda markedness constraints.

(14)

	textile	Onset	Legal(C-init)	SonContour	$*_{ons}[O$	$*O]_{coda}$
a.	[tɛ.kstaɪl]		*(kst)	*(kst)	*(k)	
b.	[tɛk.staɪl]			*(st)	*(s)	*(k)
c.	[tɛks.taɪl]			*(ks)	*(t)	*(s)
d.	[tɛkst.aɪl]	*		*(kst)		*(t)

Noting that these two forms both contain a legal cluster in either onset or coda position, we introduce constraints penalizing complex consonants depending on the position of the syllable in which they occur (Kager 1999: 97).[12]

(15) a. *Complex(coda): Codas are simple. (*CC]$_\sigma$)
b. *Complex(onset): Onsets are simple. (*$_\sigma$[CC)

The last factor to consider is morphological boundaries, whose effect is basically to align a morpheme boundary with a syllable boundary. It is evident enough that this factor

[12] When we apply these constraints, the number of consonants matters; three consonant onsets and codas are penalized more severly (with two violation marks) than two consonant onsets and codas.

can be accounted for with alignment constraints (McCarthy and Prince 1993), but before we specify them, a few remarks are in order regarding the status of morpheme boundaries we are assuming here. First, to avoid any unnecessary controversy over whether a certain word is morphologically simple or complex, this paper assumes word-based morphology in which the input and output of a morphological operation is a word (Aronoff 1976). Second, we consider only "transparent" morpheme boundaries, that is, the boundaries between two morphemes that are clearly segmentable and compositional in their meanings (cf. Lieber 2009). Lastly, since the morphological effect can differ depending on the type of the boundary, compound boundaries are distinguished from affix boundaries, which are further differentiated into prefix and suffix boundaries. Considering all these points, this paper proposes the following three alignment constraint to capture morpheme boundary effects.

(16) a. Align-Prefix: Align(Prefix, R, Syllable, R)
The right edge of a prefix coincides with the right edge of a syllable.
b. Align-Suffix: Align(Suffix, L, Syllable, L)
The left edge of a suffix coincides with the left edge of a syllable.
c. Align-Compound: Align(Compound, L/R, Syllable, L/R)
The edges of a compound member coincide with an edge of a syllable.

So far 20 constraints have been motivated to account for the syllabification factors discussed in subsection 2.1. Although some of these constraints are ranked on a fixed scale, the rankings for most constraints are yet to be determined. We will see in the next section how these constraints are assigned ranking values, and how the resulting constraint system addresses variable syllabification.

3. English medial syllabification in a weighted constraint grammar

3.1 Data

The data for this paper were obtained from Eddington et al. (2013), in which a survey of native speakers' syllabification judgments was conducted for 4,990 two-syllable words with one to four medial consonants. In this survey, syllabification judgments were elicited by asking the subjects to choose their preferred location of the syllable boundary (ambisyllabicity was not considered as a choice), and for each test item, 20 to 26 subject responses were collected.[13]

This paper used 4,824 words out of the original 4,990 after excluding 56 words with four medial consonants and 110 in which the factors for syllabification were not clearly represented.[14] To these 4,824 words (2,503 with one medial consonant, 1,889 with two medial consonants, and 432 with three medial consonants), we added stress and morphological boundaries, two important factors of syllabification that were not indicated in the original survey data. Information about word stress was first obtained from *Carnegie Melon Pronouncing Dictionary* (version 0.7b), and if not available there, from *Cambridge English Pronunciation Dictionary* (17th edition). Morphological information was marked, first based on the CELEX lexical database (Baayen et al. 1999), and then double-checked by consulting *Shorter Oxford English Dictionary* (6th edition, a CD version) to make sure that only transparent morphological boundaries were left in the data.[15]

Before we move on, let us take a brief look at how the syllabification factors are manifested in the adjusted data, especially in examples with a single medial consonant. Let us first consider the table below, in which the examples are classified into factor combinations (column A). For each factor combination, response percentages of syllabification options are presented with the number of tokens and a representative example of the factor combination.[16]

[13] Eddington et al. (2013)'s syllabification survey data are available at https://linguistics.byu.edu/faculty/deddingt/BYU Syllabification Survey.xls.

[14] Specifically, we excluded 88 words showing variable stress (e.g. *bállet*, *ballét*; *ádverse*, *advérse*; *lócate*, *locáte*), 21 words in which one medial consonant simultaneously belongs to two morphemes (e.g. *ca*[*nn*]*ot*, *gent*[*l*]*y*, *eigh*[*t*]*een*), and one word, *folkway*, of which the medial consonants did not seem to be correctly represented in the original data (that is, in Eddington et al.'s representation, the medial consonants of *folkway* were given as /lkw/, not as /kw/).

[15] There was one more adjustment made to rule out seeming outliers, that is, responses that were exceptionally high or low compared to others of a similar pattern. See section 4 for details about how we adjusted outliers. The complete data set for this paper, which contains information about stress and morphological boundaries with adjusted outliers, is provided at https://blog.naver.com/shh_61/222026007296.

[16] My aim here is not to provide a statistical analysis, so here the results are presented only descriptively. See Eddington et al. (2013) for elaborated statistical analyses. As for medial consonants, we examine only those which are phonologically legal in initial and final positions. Thus, /ŋ/, which is not allowed as an onset, and /h/ and the glides, which are not allowed as codas, are not considered in Table 1.

Table 1. Syllabification factors in words with one medial consonant

(A) Factor	(B) 1st syllable is stressed				(C) 1st syllable is unstressed			
combinations	% .C	% C.	# tokens, ex		% .C	% C.	# tokens, ex	
Lax, NB, O	70.10	29.29	481	ábbe	91.04	8.29	236	negáte
Lax, NB, N	56.15	43.28	103	bánish	83.20	15.24	57	venéer
Lax, NB, L	41.07	58.13	170	bárrel	80.69	18.64	86	alárm
Lax, +C, O	83.61	16.39	5	cúp-board	94.16	5.36	39	de-cámp
Lax, +C, N	NA	NA	0		95.45	4.55	5	re-néw
Lax, +C, L	NA	NA	0		93.99	5.36	14	de-ráil
Lax, C+, O	57.69	41.81	153	cútt-ing	32.11	67.89	7	with-óut
Lax, C+, N	40.54	58.76	25	dímm-er	NA	NA	0	
Lax, C+, L	26.78	72.50	38	léer-y	12.69	86.27	4	there-ín
Tns, NB, O	86.95	12.18	479	cíder	90.59	6.90	36	tycóon
Tns, NB, N	80.21	19.00	81	clímate	86.86	12.57	8	omít
Tns, NB, L	60.15	38.88	117	dóllar	76.86	23.14	7	tiráde
Tns, +C, O	97.23	2.46	49	rów-boat	94.41	5.59	8	pre-júdge
Tns, +C, N	97.02	2.98	12	séa-man	100	0	1	re-móunt
Tns, +C, L	95.43	4.12	10	ców-lick	NA	NA	0	
Tns, C+, O	76.18	23.28	188	créep-y	73.33	25.00	3	pos-éur
Tns, C+, N	62.46	36.94	29	dréam-er	NA	NA	0	
Tns, C+, L	56.26	43.60	32	fáll-en	NA	NA	0	

Abbreviations used:
NB = no morphological boundary (i.e. morphologically simple);
+C = morphological boundary before a medial consonant;
C+ = morphological boundary after a medial consonant;
O = obstruents; N = nasals; L = liquids;
% .C = response percentage for syllable boundary before a medial consonant;
% C. = response percentage for syllable boundary after a medial consonant;
# tokens = number of tokens; ex = example; "Tns" = tense.

The table above demonstrates how the main factors work for the syllabification of a medial consonant. First, we can see the effect of the sonority slope. If we compare the sonority of the factor combinations (column A), we can see that coda preference ("% C.") increases from O to L, which suggests that more sonorous consonants tend to be syllabified as codas. In terms of onset preference ("% .C"), the percentage increases in the opposite direction, from L to O. These tendencies appear stronger when the first syllable is stressed and no morpheme boundary intervenes. Second, we can observe the effect of stress; if we examine the factor combinations where no morpheme boundary intervenes, we can see that stressed initial syllables (column B) show a higher coda preference but a lower onset preference than unstressed syllables (column C). Third, the table shows that syllable boundaries tend to align with morpheme boundaries. Compared to cases without a morpheme boundary, words with a boundary before a medial consonant ("+C") exhibit more onset preference, while those with a boundary after a medial consonant ("C+") show more coda preference. Lastly, we can see the effect of vowel quality of the first syllable, especially when it is stressed. If we compare the cases with a lax vowel to those with a tense vowel, we can observe that stressed lax vowels tend more to be followed by a coda than stressed tense vowels.

3.2 A weighted constraint grammar analysis of English medial syllabification

3.2.1 Weighted constraint grammars

Weighted constraint grammars represent a version of OT grammar in which numerical constraints are employed, rather than ordinal ones. Numerical constraints refer to constraints that are assigned a fixed ranking value (or a constraint weight) along a real-number scale where a higher value corresponds to a higher-ranked ordinal constraint. There are two earlier versions of weighted constraint grammars: Stochastic OT (Boersma 1997; Boersma and Hayes 2001; Hayes and Londe 2006) and Harmonic Grammar (Smolensky and Legendre 2006). Both theories posit numerical constraints, but they differ in how the number of violations matters. In Stochastic OT (SOT), strict domination of constraints is assumed as in standard OT, so that a candidate form violating a higher ranked constraint is excluded first. In Harmonic Grammar (HG), however, it is also important to count the number of violations since candidate forms are evaluated based on a 'Harmony (*H*) score,' which is the sum of the ranking values (*w*) multiplied by the number of violations (*s*), as shown below in (17).

(17) $$H = \sum_{k=1}^{K} w_k \cdot s_k$$

Note that *H*-scores are usually negative since the number of violations is expressed as a negative integer, and thus, the candidate with the *H*-score closest to zero is selected as optimal.

One empirical consequence of HG, which makes it different from SOT, is that a candidate form that violates a higher ranked constraint can still be selected as the optimal output if the *H*-score of a competing form that has racked up multiple violations of lower ranked constraints is lower than that of the optimal form. This is called the “ganging-up” effect (Pater 2009), which is illustrated in the following tableau. Here the first candidate, which violates two constraints with lower weights, is ruled out because its *H*-score is lower than that of the second candidate, which violates only the most highly weighted constraint.

(18)

	Constraint1 w = 70	Constraint2 w = 50	Constraint3 w = 40	H
Candidate1		-1	-1	(-1*50)+(-1*40) = -90
☞ Candidate2	-1			-1*70 = -70

Noisy Harmonic Grammar (Pater et al. 2007; Pater 2009; Boersma and Pater 2016) is HG with ‘noise (*N*),’ a concept inherited from SOT to deal with variation. When evaluating output candidates in Noisy Harmonic Grammar (NHG), constraint weights are perturbed by noise randomly selected from a standard normal distribution with a mean of zero and standard deviation of 1.[17] The *H*-score incorporating noise is thus calculated as in (19), and if two constraint ranking values are sufficiently close, ranking reversal may take place. Such reversals give rise to variable outputs as illustrated in (20) due to the different noise values assigned at the time of evaluation.

(19) $H = \sum_{k=1}^{K} (w_k + N_k) \cdot s_k$

(20)

	Con1 w = 60 N = 1.1	Con2 w = 58 N = -0.6	H
☞ Cand1		-1	-57.4
Cand2	-1		-61.1

	Con1 w = 60 N = -1.4	Con2 w = 58 N = 1.7	H
Cand1		-1	-59.7
☞ Cand2	-1		-58.6

As we have already seen, the products of the constraint weight and the number of violations would normally add up negatively since constraint violations are given in negative numbers. If the constraint weight happens to be negative, however, the multiplication of the weight by the violation score results in a positive increase of the

[17] The actual noise utilized at the time of evaluation is obtained by multiplying the Gaussian random variable (i.e. the noise value randomly selected from the normal distribution) with a ‘ranking spread,’ a manually specifiable scale variable whose default value is 2 (Boersma 1997).

H-score, contrary to the expectation that constraint violations in OT would act against the overall harmony. Thus, if we want an OT-like evaluation from HG constraints, their weights must be limited to positive real numbers (Prince 2002; Pater et al. 2007). Thus, some approaches impose a non-negativity condition on weights so that all negative post-noise ranking values (w_k+N_k) are replaced with zero (Keller 2000, 2006).

However, simply resetting the weight of a negative constraint to zero still does not provide a satisfactory solution since it treats all negative constraints equally as contributing nothing to the harmony. Following Pater (2009) and Boersma and Pater (2016), this paper adopts Exponential NHG, a version of NHG in which a constraint's violation score is multiplied by the exponent of the sum of a constraint's weight and the noise added to this weight, as detailed in (21) below.

(21) $$H=\sum_{k=1}^{K} s_k \cdot e^{w_k+N_k}$$

Exponentiation ensures that the violation score is always multiplied by a positive number, even when w_k+N_k is negative. In this way, the weights of negative constraints are accommodated in the standard harmony calculation, in which constraints with lower weights contribute less to the harmony than those with higher weights. The tableau below illustrates how a constraint with a negative weight, Con3, contributes to the evaluation of the overall harmony.

(22)

	Con1 w = 7 N = -0.9	Con2 w = 4 N = 1.2	Con3 w = -4 N = 0.7	H
Cand1		-2	-1	(-2*$e^{5.2}$)+(-1*$e^{-3.3}$) = -362.54 -0.04 = -362.58
☞ Cand2	-1		-1	(-1*$e^{6.1}$)+(-1*$e^{-3.3}$) = -445.86 -0.04 = -445.90

3.2.2 An Exponential NHG analysis of medial syllabification

For any constraint-based analysis, a set of constraints and a ranking among them are necessary. Since we have already introduced the constraints, what remains to do is to determine the ranking values or weights of these constraints. Constraint weights are obtained here by conducting learning simulations in Praat's constraint grammar module (Boersma and Weenink 2020, version 6.1.09), which requires an initial grammar and distribution information for output candidates. The initial grammar of this analysis consists of the 20 constraints introduced in 2.2, each with an initial default weight of 100, and constraint tableaux for 4,824 bisyllabic words, in which syllabification options are posited as output candidates and their constraint violations are indicated.[18] Among the constraints, only those in the sonority markedness hierarchies, (9) and (10), are marked with their rankings fixed in the initial grammar. In addition to the initial grammar, we also need distribution information, which is basically the frequency with which each variant form occurs, drawn from the syllabification responses in Eddington et al. (2013)'s survey data.

The initial grammar and the distribution information are then submitted to Praat to initiate a learning process. In running the learning simulations, we set the 'decision strategy' to Exponential NHG for reasons discussed in the previous section. All other settings are kept at Praat's defaults.[19] Since the weights learned at each simulation are slightly different due to the noise, the simulation is run ten times and the average values are used. The average weights learned after the ten learning trials and the weights with added noise are as follows:[20]

[18] The initial grammar was constructed using the Perl scripts written by the author. The input file for this paper, which contains the initial grammar and the distribution information, is provided at https://blog.naver.com/shh_61/222026089600.

[19] To be specific, the default settings are as follows: the initial ranking values are 100, an evaluation noise is 2.0, and the initial learning rate or plasticity is 1.0 with 4 decrements of 0.1 at every 100,000 replications.

[20] The learned grammar with the entire constraint tableaux is available at https://blog.naver.com/shh_61/222026091669.

(23) Constraint weights after learning is finished

	Weight (*w*)	Weight+Noise (*w*+*N*)
Align-Prefix	298.037	296.899
Legal(C-final)	153.301	153.427
Legal(C-initial)	15.129	15.123
Align-Compound	14.102	14.265
Weight-to-Stress	12.846	13.295
Onset	11.854	11.530
*Complex(coda)	11.797	11.278
$*_{\text{onset}}$[G	11.754	11.462
$*_{\text{onset}}$[L	11.674	10.932
*O]$_{\text{coda}}$	11.368	11.707
*Complex(onset)	11.296	11.297
Legal(Vowel)	11.043	10.345
Align-Suffix	10.582	10.314
$*_{\text{onset}}$[N	10.120	11.271
$*_{\text{onset}}$[O	9.342	8.435
*N]$_{\text{coda}}$	9.144	10.702
*L]$_{\text{coda}}$	8.911	10.049
*Superheavy	-5.533	-6.246
Sonority Contour	-41.224	-40.994
Stress-to-Weight	-575.543	-575.456

Note that at the high end of the ranking are phonological legality constraints for consonants and two alignment constraints matching the boundaries of a prefix or a compound with those of a syllable.[21] Highly ranked consonant legality constraints corroborate the previous observations that phonological legality is a stronger indicator for medial syllabification than other factors such as sonority and vowel quality (Smith and Pitt 1999; Redford and Randall 2005). On the other end of the ranking are found SonContour and Stress-to-Weight. These constraints are ranked at the lowest layer because their effects substantially overlap with those of higher ranked constraints. Stress-to-Weight, for example, rules out a light syllable ending in a stressed lax vowel, which is also excluded by the higher ranked Legal(Vowel). Likewise, consonant clusters

[21] Smith and Pitt (1999), who examine suffixes only, report that the effect of morpheme boundary is minimal. Their findings are partially confirmed by the constraint ranking in (23), in which the suffix aligning constraint is assigned a much smaller weight than the prefix/compound boundary alignment constraints (cf. Laeufer 1995). As for the prefix boundary, whose effect on syllabification is by far the strongest, an anonymous reviewer pointed out that the highest ranking of Align-Prefix might be because this study has not considered a sufficient number of cases with a consonant-final prefix followed by a vowel-initial root (e.g. *dis-arm*, *dis-own*), in which a mismatch between prefix and syllable boundaries is highly likely to occur. Further study on the issue of asymmetry between prefix and suffix boundaries on syllabification is needed.

violating Sonority Contour, except for a small number of sonority-violating but legal clusters such as *sp*-, *st*-, *sk*- (see Appendix for more examples), also violate the higher ranked consonant legality constraints.

Provided below are sample tableaux extracted from the learned grammar. The tableaux illustrate how the proposed constraints evaluate output candidates with the *H*-scores provided in the rightmost column.[22] Note, among others, that the forms violating Align-Prefix and Legal(C-final), by far the two highest ranked constraints, are assigned extremely low *H*-scores (see (24c, ii), (24d, iii), (24e, iii), (24f, iv)). The forms violating the next two highest constraints, Legal(C-initial) and Align-Compound, also receive severe penalties and end up with fairly low *H*-scores (see (24b, i), (24e, i), (24f, i)).

(24) Sample tableaux from the learned grammar (with *w*+*N* as given in (23))

	Inputs: a. palate b. temper c. remount d. approve e. pipeline f. huntress	Align-Prefix	Legality(C-final)	Legality(C-initial)	Align-Compound	Weight-to-Stress	*O]coda	Onset	*onset[G	*Complex(ons)	*Complex(coda)	*onset[N	*onset[L	*N]coda	Legality(V)	Align-Suffix	*L]coda	*onset[O	*Superheavy	SonContour	Stress-to-Weight	*H*
a.	i. pǽ.lət												-1		-1						-1	$-8.702*10^{4}$
	ii. pǽl.ət							-1									-1					$-1.248*10^{5}$
b.	i. tɛ́.mpər			-1						-1		-1			-1					-1	-1	$-3.886*10^{6}$
	ii. tɛ́m.pər													-1				-1				$-4.904*10^{4}$
	iii. tɛ́mp.ər						-1	-1			-1								-1			$-3.021*10^{5}$
c.	i. rɪ.+máʊnt					-1						-1										$-6.729*10^{5}$
	ii. rɪ+m.áʊnt	-1				-1		-1						-1					-1			$-8.742*10^{128}$
d.	i. ə.prúv									-1					-1			-1				$-1.163*10^{5}$
	ii. əp.rúv					-1	-1						-1									$-7.717*10^{5}$
	iii. əpr.úv		-1			-1		-1			-1						-1		-1	-1		$-4.289*10^{66}$
e.	i. páɪ.p+làɪn				-1					-1								-1				$-1.653*10^{6}$
	ii. páɪp.+làɪn						-1						-1						-1			$-1.773*10^{5}$
	iii. páɪpl.+àɪn		-1		-1			-1			-1						-1		-1	-1		$-4.289*10^{66}$
f.	i. hʌ́.ntr+ɪs			-1						-2		-1			-1	-1				-1	-1	$-3.997*10^{6}$
	ii. hʌ́n.tr+ɪs									-1				-1		-1		-1				$-1.598*10^{5}$
	iii. hʌ́nt.r+ɪs						-1				-1		-1			-1			-1			$-2.865*10^{5}$
	iv. hʌ́ntr.+ɪs		-1					-1			-2						-1		-1	-1		$-4.289*10^{66}$

[22] The *H*-scores are calculated using the formula given in (21). The *H*-score for [pǽ.lət], for example, is $-(e^{10.932}+e^{10.345}+e^{-575.456}) = -8.702*10^{4}$.

The learned grammar, if successful, is expected to replicate the distributions of the observed responses. To verify whether this is the case, we feed the learned grammar into Praat's "To output Distributions" function to produce the predicted distributions of the outputs. Since slightly different output patterns are predicted each time the function is run due to the evaluation noise, this procedure is repeated ten times, and the average output distributions over the ten repetitions are used.[23] The predicted average output distributions obtained for the selected examples are as follows.

(25) Predicted output distributions for selected examples

	Inputs	Outputs	Predicted responses[24]		Observed responses	
a.	palate	pǽ.lət	8.62	(41.05%)	10	(47.62%)
		pǽl.ət	12.38	(58.95%)	11	(52.38%)
b.	temper	tɛ́.mpər	0.18	(0.82%)	1	(4.55%)
		tɛ́m.pər	20.38	(92.64%)	17	(77.27%)
		tɛ́mp.ər	1.44	(6.55%)	4	(18.18%)
c.	remount	rɪ.+máʊnt	21.00	(100%)	21	(100%)
		rɪ+m.áʊnt	0	(0%)	0	(0%)
d.	approve	ə.prúv	15.70	(74.76%)	16	(76.19%)
		əp.rúv	5.30	(25.24%)	5	(23.81%)
		əpr.úv	0	(0%)	0	(0%)
e.	pipeline	páɪ.p+làɪn	5.08	(23.09%)	4	(18.18%)
		páɪp.+làɪn	16.92	(76.91%)	18	(81.82%)
		páɪpl.+àɪn	0	(0%)	0	(0%)
f.	huntress	hʌ́.ntr+ɪs	0.17	(0.74%)	1	(4.35%)
		hʌ́n.tr+ɪs	17.41	(75.73%)	19	(82.61%)
		hʌ́nt.r+ɪs	5.41	(23.53%)	3	(13.04%)
		hʌ́ntr.+ɪs	0	(0%)	0	(0%)

3.2.3 Evaluating the proposed analysis

The last step of the analysis is to evaluate the quality of the predictions by comparing predicted responses to observed ones. For the evaluation, we use two error measures, Root Mean Square Error (RMSE) and R^2. RMSE calculates the prediction error rate by taking the square root of the mean square error, which is the average of the squared difference between the original (O) and predicted (P) values over the data set.[25]

[23] Praat's default settings are kept when running the "To output Distribution" function. The average output distributions predicted by the learned grammar for the entire 4,824 test words are provided at https://blog.naver.com/shh_61/222026093301.

[24] Praat's "To output distributions“ function gives the output values in proportions, but to compare with the observed responses, the predicted proportion for each syllabification option is converted to a ”quasi-frequency“ value by multiplying it with the total number of responses.

$$(26)\ \mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(O_i - P_i)^2}$$

The closer the values of RMSE are to zero, the more accurate the predicted performance will be.

R^2, known as the coefficient of determination and the square of the correlation coefficient, represents the proportion of the variance for the observed response that is predicted by the proposed model. The following formula is used here to obtain R^2. ($\overline{O}$ is the mean of the observed responses.)

$$(27)\ R^2 = 1 - \frac{\sum_{i=1}^{N}(O_i - P_i)^2}{\sum_{i=1}^{N}(O_i - \overline{O_i})^2}$$

The second term of the formula represents the ratio of the residual sum of the proposed model to that of the "mean" model, where all predictions are the same (i.e. the mean of the observed responses ($\overline{O}$)). R^2 ranges from 0 to 1, which is interpreted as a percentage of improvement over the mean model, with zero indicating that the proposed model does not improve prediction over the mean model, with 1 representing a perfect prediction.

The error measures are applied after a preprocessing step is taken to exclude the highest 5% of the observed-predicted differences from the whole set of 12,401 data points, which is the total number of syllabification options obtained from 4,824 words (that is, 2,503*2=5,006 from words with one medial consonant, 1,889*3=5,667 from words with medial CC, and 432*4=1,728 from words with medial CCC). This procedure is necessary to avoid the effect of rarely occurring high prediction errors that may substantially reduce the quality of prediction for the whole data set. The values of RMSE and R^2 measured for the remaining 95% are 2.21 and 0.93, respectively. Both values seem to suggest good predictive performance according to the criteria recommended by

[25] The other commonly used measure is Mean Absolute Error (MAE), which represents the average difference between the observed and predicted values across the whole data set. MAE gives the same weight to all errors, while RMSE penalizes large errors. RMSE is used here because of a concern about outliers.

Alexander et al. (2015). The R^2 value of 0.93 means that 93% of the data points are accounted for by the proposed analysis. This is well above 0.6, the advised threshold for being considered a good model. As for RMSE, it is suggested that the value should be less than 10% of the range of target property value in order for a model to be considered useful. Since there are 0 to 26 responses elicited for each syllabification option, the suggested RMSE threshold for being considered a good model is 2.6, which the current model meets within a fair margin.

4. Conclusions and discussion

In this paper we have presented a weighted constraint grammar analysis that best models the syllabifications reported in Eddington et al. (2013). Focusing on the syllabification of medial consonants, we first reviewed important findings of previous studies, based on which OT-based constraints were formulated. We then constructed an initial grammar, to which we submitted distribution information about medial syllabifications gathered from Eddington et al.'s (2013) survey data. After that, we performed learning simulations using Exponential NHG, a weighted constraint grammar especially designed to deal with negative constraints effectively. As a result of these learning processes, we obtained an output grammar where each constraint was assigned a numerical weight, and from this output grammar we generated a predicted response for each syllabification option. We then evaluated the predictive accuracy by comparing the predicted to the observed responses in terms of RMSE and R^2, and the results confirmed a good predictive performance for the proposed analysis.

Although the current analysis replicated the observed responses quite well, the match was not perfect. Closing this paper, we will speculate on the possible loci of mismatche s.[26] First of all, mismatches seem to occur most likely in cases where the observed responses were markedly deviant from those of other similar examples. To verify this, 4,824 test words were classified into 433 groups based on the syllabification factors discussed in 2.1. For each word, the deviance of the observed response was obtained from the mean observed response of the group to which it belongs. A Spearman rank-order correlation test was conducted to compare this deviance with the difference

[26] Prediction performance for the entire 12,401 data points, including the list of 433 test word groups, is provided at https://blog.naver.com/shh_61/222165849635.

between the observed and predicted responses. The results showed that the two values are positively correlated (ρ(12,399) = .626, p < .001), verifying that mismatches are more likely to occur when the observed responses become more deviant from those of the other examples within the same group.

The second possible source of mismatches is "opaque" morphological boundaries, which we did not consider in this paper. We may call a morpheme boundary opaque if the morphemes before and after the boundary are not easily segmented, or their meanings are not compositional. Opaque boundaries, as well as transparent boundaries, influence medial syllabification according to Eddington et al. (2013: 60), who specifically report that when there is a preconsonantal opaque boundary in words with one medial consonant, the log odds of onset preference is 0.84. This corresponds to an odds ratio of 2.32 (= $e^{0.84}$), meaning that the probability of the medial consonant being the onset of the second syllable is 2.32 times higher than the probability of it being the coda of the first syllable. They further note that an opaque boundary inhibits onset preference when it is placed after a medial consonant with the log odds of -0.30 or the odds ratio of 0.74, which means that the medial consonant is less likely to be syllabified as an onset by 26% when there is a post-consonantal opaque boundary. Opaque boundaries may not be as influential as compound or transparent affix boundaries, but their effects are still observable; Eddington et al. (2013) note that their respective log odds for onset preference are 1.88 and 1.32 when there is a boundary before a medial consonant, and -0.77 and -3.03 when the boundaries are placed after a consonant.

Some of the mismatches may have arisen from orthography, since the syllabification survey on which this paper is based was conducted on written forms. The effect of spelling is particularly evident if a test word includes a consonant sound that is represented by two letters. This is where ambisyllabicity was often reported in previous research; ambisyllabic responses are more frequent for a consonant sound represented by an orthographic geminate (e.g. [b] in *rabbit*) than for one represented by a single letter (e.g. [b] in *habit*) (Treiman and Danis 1988a, Derwing 1992). Eddington et al. (2013: 56) also note that digraphs such as *ck* exhibit a similar behavior; the segment [k] represented by the orthographic sequence *ck* is less likely to be syllabified as an onset than when it is represented by a single letter. To handle the influence of spelling, they propose the concept of "orthographic legality," which bans a letter or a sequence of letters that does not appear at the beginning of a word from occurring in syllable-initial position.

Lastly, there may be mismatches related to gradient phonotactics. Among the phonologically illegal clusters that have a similar composition, there are some clusters that receive different degrees of penalty than others. For example, we treated obstruent-liquid onset clusters such as *vr* and *sr* equally as phonologically illegal, but unlike *sr*, which is allowed as onset in just 3.17%, *vr* is much more likely to be syllabified as an onset at 33.66%. Gradient phonotactics may be also exhibited in phonologically legal clusters with similar combinations. Obstruent-glide clusters such as *kw* and *tw*, for example, show different degrees of onset preference: the first is 82.43% while the latter is substantially lower at 20.48%.

The sources of mismatches that we have considered in this section are by no means exhaustive. Nevertheless, I believe that it will be vital to address these issues in future research in order to develop more accurately predictive models.

References

Alexander, David. L. J, Alexander Tropsha, and David A. Winkler. 2015. Beware of R2: Simple, unambiguous assessment of the prediction accuracy of QSAR and QSPR models. *Journal of Chemical Information and Modeling* 55(7): 1316-1322. Available as author manuscript at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4530125.

Aronoff, Mark. 1976. *Word formation in generative grammar*. Cambridge, MA: The MIT Press.

Baayen, Herald, Richard Piepenbrock, and Leon Gulikers. 1999. *The CELEX Lexical Database* (CDROM). Philadelphia, PA: University of Pennsylvania, Linguistic Data Consortium.

Bailey, Charles. 1978. *Gradience in English syllabization and a revised concept of unmarked syllabization*. Bloomington, IN: Indiana University Linguistics Club.

Boersma, Paul and Bruce Hayes. 2001. Empirical tests of the gradual learning algorithm. *Linguistic Inquiry* 32(1): 45-86.

Boersma, Paul and Joe Pater. 2016. Convergence properties of a Gradual Learning Algorithm for Harmonic Grammar. In John McCarthy and Joe Pater (eds.), *Harmonic Grammar and Harmonic Serialism*, 389-434. London: Equinox Press.

Boersma, Paul and David Weenink. 2020. *Praat: Doing phonetics by computer* (Version 6.1.09) [Computer program]. http://www.praat.org.

Boersma, Paul. 1997. How we learn variation, optionality, and probability. *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam* 21, 43-58.

Borowsky, Toni. 1986. *Topics in the lexical phonology of English*. PhD Dissertation. University of Massachusetts, Amherst.

Clements, George N. 1990. The role of the sonority cycle in core syllabification. In John Kingston and Mary Beckman (eds.), *Papers in Laboratory Phonology I*, 283-333. Cambridge: Cambridge University Press.

Clements, George N. and Samuel Jay Keyser. 1983. *CV phonology: A generative theory of syllable*. Cambridge, MA: The MIT Press.

Coetzee, Andries. 2016. A comprehensive model of phonological variation: Grammatical and non-grammatical factors in variable nasal place assimilation. *Phonology* 33(2): 211-246.

Coetzee, Andries and Joe Pater. 2008. Weighted constraints and gradient restrictions on place co-occurrence in Muna and Arabic. *Natural Language and Linguistic Theory* 26(2): 289-337.

Coetzee, Andries and Joe Pater. 2011. The place of variation in phonological theory. In John Goldsmith, Jason Riggle, and Alan Yu (eds.), *The handbook of phonological theory*, 401-434. Oxford: Blackwell.

Coetzee, Andries and Shigeto Kawahara. 2013. Frequency biases in phonological variation. *Natural Language and Linguistic Theory* 31(1): 47-89.

Derwing, Bruce. 1992. A ‘pause-break’ task for eliciting syllable boundary judgments from literate and illiterate speakers: Preliminary results for five diverse languages. *Language and Speech* 35(1-2): 219-235.

Eddington, David, Rebecca Treiman, and Dirk Elzinga. 2013. Syllabification of American English: Evidence from a large-scale experiment, Part I. *Journal of Quantitative Linguistics* 20(1), 45-67.

Fallow, Deborah. 1981. Experimental evidence for English syllabification and syllable structure. *Journal of Linguistics* 17(2): 309–317.

Féry, Caroline. 2003. Markedness, faithfulness, vowel quality and syllable structure in French. *Journal of French Language Studies* 13(2): 247-280.

Giegerich, Heinz. 1992. *English phonology: An introduction*. Cambridge: Cambridge University Press.

Gouskova, Maria. 2004. Relational hierarchies in Optimality Theory: The case of syllable contact. *Phonology* 21(2): 201-250.

Halle, Morris and Karuvannur P. Mohanan. 1985. Segmental phonology of modern English. *Linguistic Inquiry* 16(1): 57-116.

Hammond, Michael. 1997. Vowel quality and syllabification in English. *Language* 73(1): 1-17.

Hammond, Michael. 1999. *The phonology of English: A prosodic Optimality-Theoretic approach*. Oxford: Oxford University Press

Hayes, Bruce. 2017. Varieties of Noisy Harmonic Grammar. Ms. UCLA.

Hayes, Bruce and Zsuzsa Londe. 2006. Stochastic phonological knowledge: The case of Hungarian vowel harmony. *Phonology* 23(1); 59-104.

Hoard, James E. 1971. Aspiration, tenseness, and syllabification in English. *Language* 47(1): 133–140.

Hooper, Joan. 1972. The syllable in phonological theory. *Language* 48(3): 525-540.

Hooper, Joan. 1976. *An introduction to Natural Generative Phonology*. New York, NY: Academic Press.

Jesney, Karen. 2007. The locus of variation in weighted constraint grammars. Paper presented at Workshop on Variation, Gradience and Frequency in Phonology, Stanford, CA.

Kager, René. 1999. *Optimality Theory*. Cambridge: Cambridge University Press.

Kahn, Daniel. 1976. *Syllable-based generalizations in English phonology*. PhD Dissertation. MIT.

Keller, Frank. 2000. *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality*. PhD Dissertation. University of Edinburgh.

Keller, Frank. 2006. Linear optimality theory as a model of gradience in grammar. In Gisbert Fanselow, Caroline Féry, Ralf Vogel, and Matthias Schlesewsky (eds.), *Gradience in grammar: Generative perspective*, 270-287. Oxford: Oxford University Press.

Kiparsky, Paul. 1979. Metrical structure assignment is cyclic. *Linguistic Inquiry* 10(3): 421-441.

Ladefoged, Peter. 2006. *A course in phonetics*. Boston, MA: Thomson Wadsworth.

Laeufer, Christiane. 1995. Morphology and syllabification domains. *Lingua* 97(2-3): 101-121.

Lieber, Rochelle. 2009. *Introducing morphology*. Cambridge: Cambridge University Press.

McCarthy, John and Alan Prince. 1993. Generalized alignment. In Geert Booij and Japp Van Marle (eds.), *Yearbook of morphology* 1993, 79-153. Dordrecht: Kluwer.

Murray, Robert and Theo Vennemann. 1983. Sound change and syllable structure in Germanic phonology. *Language* 59(3): 514–528.

Myers, Scott. 1987. Vowel shortening in English. *Natural Language and Linguistic Theory* 5(4): 485–518.

Pater, Joe. 2009, Weighted constraints in generative linguistics. *Cognitive Science* 33(6): 999-1035.

Pater, Joe, Chris Potts, and Rajesh Bhatt. 2007. Harmonic Grammar with linear programming. Rutgers Optimality Archive #872.

Prince, Alan. 2002. Anything goes. In Takeru Honma, Masao Okazaki, Toshiyuki Tabata, and Shin-ichi Tanaka (eds.), *New century of phonology and phonological theory*, 66-90. Tokyo: Kaitakusha.

Prince, Alan and Paul Smolensky. 1993/2004. *Optimality Theory: Constraint interaction in generative grammar*. Oxford: Blackwell.

Pulgram, Ernst. 1970. *Syllable, word, nexus, cursus*. The Hague: Mouton.

Redford, Melissa and Patrick Randall. 2005. The role of juncture cues and phonological knowledge in English syllabification judgments. *Journal of Phonetics* 33(1): 27-46.

Selkirk, Elizabeth. 1982. The syllable. In Harry van der Hulst and Norval Smith (eds.), *The Structure of phonological representations II*, 337–383. Dordrecht: Foris.

Smith, Katherine and Mark Pitt. 1999. Phonological and morphological influences in the syllabification of spoken words. *Journal of Memory and Language* 41(2): 199-222.

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

Treiman, Rebecca. 1984. On the status of final consonant clusters in English syllables. *Journal of Verbal Learning and Verbal Behavior* 23(3): 243-356.

Treiman, Rebecca and Andrea Zukowski. 1990. Toward an understanding of English syllabification.

Journal of Memory and Language 29(1): 66-85.

Treiman, Rebecca and Catalina Danis. 1988a. Syllabification of intervocalic consonants. *Journal of Memory and Language* 27(1): 87-104.

Treiman, Rebecca and Catalina Danis. 1988b. Short-term memory errors for spoken syllables are affected by the linguistic structure of the syllables. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14(1): 145-152.

Treiman, Rebecca, Kathleen Straub, and Patrick Lavery. 1994. Syllabification of bisyllabic nonwords: Evidence from short-term memory errors. *Language and Speech* 37(1): 45-60.

Vennemann, Theo. 1988. *Preference Laws for Syllable Structure and the Explanation of Sound Change*. Berlin: Mouton de Gruyter.

Wells, John C. 1990. Syllabification and Allophony. In Susan Ramsaran (ed.), *Studies in the pronunciation of English: A commemorative honor of A. C. Gimson*, 76–86. London: Routledge.

Yavaş, Mehmet. 2011. *Applied English phonology*. Oxford: Wiley-Blackwell.

Appendix

Onset and coda CC clusters in English

		C2	Obstruents																	Nasals			L		G	
	C1		p	t	k	b	d	g	ʧ	ʤ	f	θ	s	ʃ	h	v	ð	z	ʒ	m	n	ŋ	l	r	j	w
Onset CC clusters	O	p	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	/	/	/	○	○	○	/
		t	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	/	/	/	/	○	○	○
		k	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	/	/	/	○	○	○	○
		b	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	/	/	/	○	○	○	/
		d	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	/	/	/	/	○	○	○
		g	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	/	/	/	○	○	○	○
		ʧ	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	/	/	/	/	/	/	/
		ʤ	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	/	/	/	/	/	/	/
		f	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	/	/	/	○	○	○	/
		θ	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	/	/	/	/	○	○	○
		s	⊗	⊗	⊗	×	×	×	×	×	×	×	×	×	×	×	×	×	×	○	○	/	○	/	/	○
		ʃ	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	/	/	/	/	○	/	/
		h	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	/	/	/	/	/	○	○
		v	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	/	/	/	/	/	○	/
		ð	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	/	/	/	/	/	/	/
		z	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	/	/	/	/	/	/	○
		ʒ	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	/	/	/	/	/	/	/
	N	m	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	/	/	○	/
		n	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	/	/	○	/
		ŋ	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	/	/	/	/
	L	l	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	/	/
		r	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	/	/
	G	j	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
		w	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×

		C2	Obstruents																	Nasals			L		G	
	C1		p	t	k	b	d	g	ʧ	ʤ	f	θ	s	ʃ	h	v	ð	z	ʒ	m	n	ŋ	l	r	j	w
Coda CC clusters	O	p	×	⊗	×	×	×	×	×	×	×	⊗	⊗	×	×	×	×	×	×	×	×	×	×	×	×	×
		t	×	×	×	×	×	×	×	×	×	⊗	⊗	×	×	×	×	×	×	×	×	×	×	×	×	×
		k	×	⊗	×	×	×	×	×	×	×	×	⊗	×	×	×	×	×	×	×	×	×	×	×	×	×
		b	×	×	×	×	⊗	×	×	×	×	×	×	×	×	×	×	⊗	×	×	×	×	×	×	×	×
		d	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	⊗	×	×	×	×	×	×	×	×
		g	×	×	×	×	⊗	×	×	×	×	×	×	×	×	×	×	⊗	×	×	×	×	×	×	×	×
		ʧ	×	⊗	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
		ʤ	×	×	×	×	⊗	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
		f	×	⊗	×	×	×	×	×	×	×	⊗	⊗	×	×	×	×	×	×	×	×	×	×	×	×	×
		θ	×	⊗	×	×	×	×	×	×	×	×	⊗	×	×	×	×	×	×	×	×	×	×	×	×	×
		s	⊗	⊗	⊗	×	×	×	×	×	×	⊗	×	×	×	×	×	×	×	×	×	×	×	×	×	×
		ʃ	×	⊗	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
		h	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
		v	×	×	×	×	⊗	×	×	×	×	×	×	×	×	×	×	⊗	×	×	×	×	×	×	×	×
		ð	×	×	×	×	⊗	×	×	×	×	×	×	×	×	×	×	⊗	×	×	×	×	×	×	×	×
		z	×	×	×	×	⊗	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
		ʒ	×	×	×	×	⊗	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
	N	m	○	/	/	/	○	/	/	/	○	○	/	/	/	/	/	○	/	×	×	×	×	×	×	×
		n	/	○	/	/	○	/	○	○	/	○	○	/	/	/	/	○	/	×	×	×	×	×	×	×
		ŋ	/	/	○	/	○	/	/	/	/	○	/	/	/	/	/	○	/	×	×	×	×	×	×	×
	L	l	○	○	○	○	○	/	○	○	○	○	○	○	/	○	/	○	/	○	○	/	×	×	×	×
		r	○	○	○	○	○	○	○	○	○	○	○	○	/	○	/	○	/	○	○	/	⊗	×	×	×
	G	j	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	×	×
		w	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	/	×	×

Symbols used:
× Both the Sonority Contour Principle (SCP) and Phonological Legality (PL) are violated;
○ both the SCP and PL are met;
⊗ the SCP is violated but PL is met;
/ the SCP is met but PL is violated.

Sung-Hoon Hong
Professor
Department of English Language and Language Technology
Hankuk University of Foreign Studies
107 Imun-ro, Dongdaemun-gu
Seoul 02450, Korea
E-mail: hongshoon@hufs.ac.kr

Received: 2021. 02. 10.
Revised: 2021 03. 14.
Accepted: 2021. 03. 15.