Korean Journal of English Language and Linguistics
Vol. 19, No. 1, Spring 2019, 1–26
10.15738/kjell.19.1.201903.1

A Maximum–Entropy Grammar of Phonotactics for the TRAP–BATH Vowel Distribution

Hyesun Cho
(Dankook University)

Cho, Hyesun. 2019. A maximum–entropy grammar of phonotactics for the TRAP–BATH vowel distribution. *Korean Journal of English Language and Linguistics* 19–1, 1–26. This study presents a probabilistic phonotactic grammar for the distribution of the TRAP and BATH vowels ([æ] and [ɑ:]). Patterns of variation exist since the Middle English [a] lengthening did not take place in all the eligible words satisfying the structural description (voiceless fricatives and nasal–consonant clusters). As a result, the vowel varies even in nearly–identical phonological environment (*brass* [ɑ:] vs. *crass* [æ]). This paper presents a probabilistic maximum–entropy grammar of phonotactics for the vowel distribution and variation. The learning simulation was run using the UCLA Phonotactics Learner (Hayes and Wilson 2008). The obtained grammar in the first simulation captured phonotactic patterns that are relatively strong and consistent, describing the phonotactic environments where one of the vowels never occur. To capture the patterns of variation, the second simulation was run, which resulted in a phonotactic grammar that assigns different probabilities for words that have variable vowels in the same phonological context. The results suggest that the probabilistic grammar that defines the probability distribution over the phonological forms is adequate for modeling variable phonotactic distribution which involves the gradient well–formedness of surface forms.

Keywords: phonotactics, maximum entropy, probabilistic grammar, TRAP, BATH, variation, frequency, distribution

1. Introduction

1.1 Patterns of Variation

This paper aims to offer a quantitatively–precise description of the phonological environment for the variable distribution of the TRAP vs. BATH vowels, i.e. short–a [æ] vs. broad–a [ɑ:] in the Present–Day British English. In modern British English and several more varieties of English (New Zealand, South African, and older Boston accents), the TRAP and BATH lexical sets (a group of words that have different vowels, Wells (1982)) have different vowels, [æ] and [ɑ:]. These two vowels both originated from the Middle English (ME) low front unrounded

short vowel [a]. The ME short [a] is raised to [æ], but in some words it lengthened to [a:] before voiceless fricatives (*staff, pass, path*) and French-origin nasal-consonant clusters (*dance*), which later merged with the low back [ɑ:] (as in *father*) (Wells 1982: 100). Patterns of variation exist since the lengthening did not take place in all eligible words.

The distribution of the TRAP and BATH vowels is predictable to a certain extent, but the environments are not precisely captured by usual natural classes, and furthermore, either vowel may appear in similar environment. The phonological context for the vowel [ɑ:] has been roughly described as tautosyllabic voiceless fricatives (/f/, /s/, /θ/) (1a) or nasal + consonant clusters (1b). ME [a]-lengthening did not take place if the post-vocalic fricative is heterosyllabic, so the vowel is [æ] as in *traffic, classical, passage.*

(1) a. staff b. demand
bath aunt
pass can't
class chance
ask branch

Describing the contexts for (1a) in terms of a general natural class, voiceless fricative, is not precise, since the environments do not include the post-alveolar fricative [ʃ]. Before [ʃ], vowel [æ] mainly occurs. A search in British English Example Pronunciation Dictionary (Robinson 1995) reveals that among monosyllabic words, [æʃ] occurs in 24 words, e.g., *ash, bash*, whereas [ɑ:ʃ] shows up in 3 words, e.g., *harsh, marsh, tache*.

Vowel distribution is not entirely predictable from its phonological environment. Phonological contexts where both vowels may appear are described in (2) (adapted from Wells 1982: 232-233). [æ] is found before a voiceless fricative: *math* [mæθ] and *lass* [læs]. The vowels are variable even if the phonological environment is nearly identical: e.g., *brass* [ɑ:] vs. *crass* [æ], *aunt* [ɑ:] vs. *ant* [æ].

(2)	RP /ɑ:/	RP /æ/
_f#	staff, laugh, giraffe, calf, half	gaff, gaffe
_fC	craft, shaft, after, laughter	Taft
_θ#	path, bath	math(s), hath, strath
_s	brass, pass, class	crass, lass
_st	last, past, mast	hast, bast, enthusiast, aster

_sp#	clasp, grasp, rasp, gasp	asp
_sk	ask, flask, mask	casque, gasket, ascot, mascot
_sl	castle	tassel, hassle, vassal
_sn	fasten	Masson
_ns	dance, chance, france	manse, romance, expanse
_nt	grant, slant, aunt	rant, ant, extant, banter, canter, antic
_n(t)ʃ	branch, blanch	mansion, expansion, scansion
_nd	demand, command, remand	stand, grand, hand
_mpl	example, sample	ample, trample

The distribution of [ɑ:] and [æ] is not entirely predictable, which is natural because they are phonemes, not allophones. However, the distribution is also partly regular to a varying degree, for example, before [ʃ], [æ] is preferred, and before a velar nasal, *[ɑ:] is absolutely prohibited (Gussmann 2002: 68). In brief, the distribution is complex, so it can only be adequately accounted for in a theory that reflects the gradient nature of the well-formedness of phonological forms.

1.2 Historical Development of the TRAP-BATH Split

The variable distribution of the vowels [ɑ:] and [æ] is at least partly due to their mixed historical sources. This section thus reviews the historical development of the TRAP-BATH vowels based on Wells (1982), Barber (1997), and Migglestone (1995). The vowels [æ] and [ɑ:] both originated from the Middle English vowel [a] (low front unrounded short). In the Early Modern English period (1500-1700, eModE), triggered by the raising of the Middle English /ɛ/, the ME short-/a/ began to move up and front to [æ] (Barber 1997: 109). The vowel [æ] had become a generally accepted pronunciation by the first half of the seventeenth century. The words affected by this change are called the TRAP lexical set. The [æ] is found in words with the spelling <a>, e.g., *cat, hat, man, hand* in the Present-Day English (PDE).

The BATH vowel [ɑ:] has mixed historical origins, so it occurs various phonological contexts: before a voiceless fricative, before a nasal, and before /ɹ/. Firstly, post-vocalic voiceless fricatives /s/, /f/, and /θ/ (tautosyllabic) caused the lengthening of Middle English /a/ (later 17th century). This inhibited the raising to [æ] that took place in the TRAP lexical set (e.g., *man, hat*). The lengthened low front unrounded long vowel [a:] later developed to a back vowel [ɑ:] in PDE, as in (3).

(3) ME ask [ask] > eModE [aːsk] (later 17th century) > PDE [ɑːsk]
ME staff [staf] > eModE [staːf] (later 17th century) > PDE [stɑːf]

However, the lexical set that underwent the pre-fricative lengthening was not stable. Not all eligible words have undergone the [a] lengthening: in some words, the vowel [a] did not lengthen (e.g., *asp, lass, mass*). Thus these words preserve a short vowel in PDE (Barber 1997: 122).

The vowels in the BATH lexical set differ by region, as well. The lengthening took place mostly in accents of Southern English, so it is a regional marker that distinguishes southern from northern accents in British English. For example, according to a survey by Gupta (2005: 23-24), *grass* is pronounced more with a short vowel in Northern regions (93% of respondents), more with a long vowel in Southern regions (96% of respondents).

Secondly, in a group of French origin words, ME [a] had developed to [ɑː] (*aunt, chant, chance, dance, demand, example*) in Standard English in the eighteenth century. Postvocalic consonants were a nasal plus consonant cluster. The vowel was originally spelled as <au> (*chaunt, daunce* in the sixteenth century), which was a diphthong at that time but later monophthongized into long [ɑː] in Southern Standard British English.

Thirdly, the post-vocalic /ɹ/ was another source of long [ɑː]. The word *yarn* had the ME short /a/, [jarn] (16th century). The consonant [r] inhibited the raising of [a] to [æ][1]. In the second half of the seventeenth century, [r] causes lengthening of the preceding [a], so [jarn] changed to [jaːrn]. In the middle of the eighteenth century [r] is lost (non-rhotic dialect), and [aː] has developed into [ɑː], which results in [jɑːn] for the pronunciation of *yarn* in PDE (Barber 1997: 117-118).

Not all ME [a] vowels have become [æ] or [aː]. After [w], [a] became [ɔ] (ME water [watər] > EModE [wɔtə]) (e.g., *wander, wash*). Before a tautosyllabic /l/, the ME [a] had developed into [aʊ], which later became PDE [ɔː] (e.g., *all, altar, always, chalk*, etc.). In some words of French origin, the ME [a] became [eɪ] (e.g., *ancient, angel, chamber, change*). However, if a tautosylabic /l/ is followed by labials (/f/, /m/), [a] developed into [ɑː], e.g., ME half [half] > EModE [haulf] > [hauf] > [hɔːf] > [haːf]. Thus, post-vocalic [l]-labial is another source of PDE [ɑː].

[1] According to Barber (1997: 116-117), the quality of /r/ was uncertain, though it was probably a trill in Old and Middle English. Since then, the /r/ sound has been weakened: it was perhaps a fricative in early Modern English. In PDE, it is weakened to an approximant /ɹ/.

In summary, since PDE [ɑ:] has developed from various sources (pre-fricative lengthening, French-origin nasal-C clusters, before [r], and before [l]+labial clusters), the surface phonological environment for [ɑ:] is not homogeneous in PDE. The environment cannot be captured by a usual natural class: voiceless fricatives (excluding [ʃ]), nasals, and [l] do not belong to any single typical natural class, except that they are [+consonantal], which is not even accurate because stops should be excluded. The next section describes a probabilistic model that can account for the variable and complicated phonological distribution patterns of [ɑ:] and [æ].

1.3 Models of Phonological Variation

The patterns of variation shown in (2) cannot be described in terms of traditional generative phonological rules or standard Optimality Theory constraints, because either vowel can appear in the same context. Labov (1969) introduced variable rules, which assigns probability of rule application depending on the presence and absence of the context. While this offers a probabilistic component to phonological rules, it cannot explain the TRAP-BATH vowel variation because either vowel can occur in nearly identical contexts. The constraint-based models for phonological variation (Anttila 1997, Kiparsky 1993) would fail to account for the variation for the same reason. Partially Ordered Constraints (POC, Anttila 1997) predicts possible and impossible phonological systems, but would not distinguish the well-formedness among existing, grammatical forms. There are constraint-based models for phonological variation, such as Noisy Harmonic Grammar (Coetzee and Pater 2011), but this is not adequate for static phonotactic variation because they take input and output forms.

The patterns of phonotactic variation can be more adequately captured in a model that defines the probability distribution over possible phonological forms. The current paper aims to provide a probabilistic model for the variable distribution of TRAP and BATH vowels, in the framework of a maximum-entropy (MaxEnt) grammar of phonotactics (Hayes and Wilson 2008). The grammar is learned from a corpus, instead of a hand-crafted list of words, so the probability of a vowel [ɑ:] or [æ], given a context, can be quantitatively-precisely modeled. This takes into account not just immediate phonological contexts, but all the segments in a word can affect the probability of the form. The MaxEnt grammar captures both categorical and gradient nature of phonotactic distribution (Cho 2012, Hayes and Wilson 2008). The grammar can not only distinguish phonotactically impossible (ungrammatical) forms from possible

(grammatical) forms, but it can also distinguish the well-formedness among possible, existing forms, in a gradient manner. For example, there is a strict ban on a back vowel before a velar nasal, e.g., *[ɑːŋ] is phonotactically ungrammatical (Gussmann 2002: 68) and never attested. Thus it is expected that such sequences will be assigned near-zero probability. On the other hand, when both vowels occur in the same immediate environment (*brass* vs. *crass*), the probabilistic grammar may assign different probabilities to different words, based on the patterns of distribution in the entire corpus, even though both forms are phonotactically acceptable.

In addition, usage frequency will also be examined for the words with variable vowels. It is known that variable phonological processes tend to apply to frequent words more than infrequent words (Bybee 2001, Coetzee and Kawahara 2013). For example, t/d deletion in English targets frequent words than infrequent words (*just* vs. *jest*), and schwa deletion occurs more in frequent words (*memory* vs. *mammary*). The MaxEnt model distinguishes the well-formedness among variable words, and a possible source of variation will be identified by looking at the words' usage frequency.

2. A Probabilistic Model of Phonotactic Grammar

The phonological environment that determines between [æ] or [aː] is not adequately captured by traditional description of segmental environment. Patterns of variation can better be explained based on probabilities of occurrence of a specific vowel, depending on surrounding segments. Generalization over the statistical patterns of segments is necessary to make predictions of the occurrence of a vowel. Non-adjacent segments as well as immediately adjacent segments may affect the choice of vowel (e.g., *chant* [ɑː] vs. *brand* [æ]). Such detailed phonotactic knowledge may be a part of unconscious knowledge of speakers. Studies have shown that speakers can generalize over gradient statistical patterns in the lexicon (Albright and Hayes 2003, Hayes and Londe 2006, Jun 2010).

The UCLA Phonotactics Learner (Hayes and Wilson 2008) is used to find the probabilistic phonotactic grammar accounting for the distribution of [æ] and [ɑː]. The output grammar obtained from the learning simulation assigns probability of every word in the space of possible words. The grammar distinguishes phonotactically well-formed words from ill-formed words in terms of the probability of word (Kager and Pater 2012).

The learning data and phonological feature specification are fed to the UCLA Phonotactics Learner. Testing data can be optionally entered. The learning data consists of a set of words in a text file format - only surface forms of words are needed, without underlying forms, since phonotactic grammar is a pattern of generalization over surface forms. Feature specifications are given in a text file that specifies feature values for every segment that appears in the learning data. Using the feature specification, the Learner creates all the possible natural classes. From these natural classes, OT-style constraints are generated. The maximum number of feature matrices for a constraint can be set by the user: for example, a constraint *[-back][+continuant] would be of size 2, since it has two natural classes. If unspecified, the Learner will find constraints of any length.

The output phonotactic grammar consists of a set of weighted constraints. From all possible constraints, constraints are added to the grammar based on the O/E criteria (Frisch, Pierrehumbert and Broe 2004, Pierrehumbert 1994). In general, the O/E value is the ratio of the observed number of given forms to the expected number of the forms if segments are combined at random. In the Hayes and Wilson model, the O/E is computed for constraints, which is the ratio of the observed number of constraint violations to the expected number of constraint violations. The O/E indicates the accuracy of a constraint, so the O/E value of 0.001 means that the constraint was violated once among 1000 possible forms. Such a constraint is a strong, accurate constraint, compared with those with the O/E of, say, 0.1. The possible constraints are added to the grammar in the ascending order of accuracy, i.e., starting with more accurate ones.

After that, a set of weights for the constraints is searched for following the principle of maximum entropy, i.e., the set of weight values that maximizes the probability of the given learning data (P(D)). The weights are searched for by 'iterative ascent', which is guaranteed to converge because the search space for a maximum entropy grammar is always convex, where there is one maximum (Hayes and Wilson 2008: 387). This is where P(D) has the maximum value under a given set of constraints and weights. The weight of a constraint implies the importance of the constraint: higher-weighted constraints are more important to satisfy. Forms violating higher-weighted constraints will be less likely to surface.

In a maximum entropy grammar, probability of a word form is the well-formedness of the form. The weighted sum of constraint violations by a form is the 'score' of the form (Hayes and Wilson 2008: 383). Thus, for a given form, violations of any

constraint will increase the score, so the higher the score, the worse the form. The maximum entropy (MaxEnt) value for a form ($P^*(x)$) is e^{-score}, and probability of a phonological representation x ($P(x)$) is $P^*(x)/Z$, where Z is the sum of MaxEnt values of all the possible forms (Hayes and Wilson 2008: 384). In brief, the more well-formed a form (with less constraint violations), the lower the score, and the higher the probability. Using the score, the probabilistic model gives gradient probabilities for any phonological forms, instead of making binary judgments on whether a form is grammatical or ungrammatical.

The maximum entropy Phonotactics Learner is chosen because it can generalize over detailed patterns of segments neighboring the vowels, giving gradient well-formedness judgments for words. The output phonotactic grammar will help understand the detailed gradient patterns involving the choice of a vowel. Not only can we better understand the already-known environments in terms of more precise quantitative weights, but also new patterns that have eluded human linguists' observation may be found.

3. Learning Simulation

The learning data for the simulation was collected from British English Example Pronunciation Dictionary ('BEEP', Robinson 1995), consisting of 253093 word forms. This dictionary contains words with inflectional suffixes (*-ed* (past tense), *-'s* (possessive), and *-s* (plural)). Words ending with these suffixes were all removed using scripts in R (R Core Team 2018) with a text-matching function *grepl,* a base function in R. After that, monosyllabic words with either /æ/ or /ɑː/ were collected, which resulted in 624 words. Vowels /æ/ and /ɑː/ also appear in polysyllabic words (*advantage* [ədvɑːntɪʤ], *adversary* [ædvəsəri]), but the learning data was limited to monosyllabic words only, in order to control for the confounding factors such as morphological complexity and word-position. Thus, words with syllabic consonants in the second syllable (e.g., *ample* [æmp̩l]) were removed in the learning data as well.

In addition to the learning data, a feature chart is provided, by which the Learner generates a list of all possible natural classes. The feature chart for English phonemes was created based on Clements and Hume (1995), Halle (1995), and mainly Hayes and Wilson's (2008: 396) simulation for English onset phonotactics (See Appendix). Following Hayes and Wilson (2008), contrastive and privative underspecifications were used in order to control the number of natural classes. Features [anterior] and

[strident] were specified only for feature [coronal], and features [voice] and [continuant] were specified only for obstruents ([-sonorant]). Features [high] and [back] were specified only for vocoids (vowels and glides). Features [short], [long] were specified only for vowels. Privative underspecification (using only + value) was used for [nasal], [spread], and place features. Feature [approximant] was not specified for vowels. Vowels were given [+approximant] specification in Clements and Hume (1995), but, for the purpose of phonotactics learning, grouping approximants (/l/, /r/, /j/, /w/) together can be more useful than grouping /l/, /r/, /j/, /w/ and vowels together. Approximants act together in English onset clusters (e.g., play) (Kenstowicz 1994: 34, 251), while approximants and vowels usually do not form a natural class that acts together.

The UCLA Phonotactics Learner was run to obtain the phonotactic constraints that regulate the distribution of /æ/ and /ɑː/. The maximum O/E for constraints was set to 0.3. This means that the Learner will stop adding constraints once it reaches O/E greater than 0.3. The same value was used in Hayes and Wilson's (2008) learning simulation for English onset constraints. The maximum number of constraints to discover was not limited so the Learner will find all relevant constraints under the given accuracy level (O/E). The maximum constraint size was the default value, size of 3, meaning that the constraints will include maximally three feature matrices. All other parameters were left unspecified, using default values.

Hayes and Wilson (2008) conducted simulations to find a phonotactic grammar for possible English onset clusters. They ran the same simulations 10 separate times. Constraints in the learned grammars are slightly different in each run, since constraint selection is stochastic. They reported the grammar that showed the lowest correlation with gradient judgments of English speakers (from Scholes 1966), saying that all the ten grammars were very similar.

In the present phonotactics learning, I conducted 10-fold cross validation (Kohavi 1995) in order to ensure that grammars from 10 simulations did not have significant differences. For 10-fold cross validation, the learning data were divided into 10 parts, and each one of them was set aside as a testing data in turn, while the remaining data was the learning data for each run. For example, the first 1/10 of the learning data was used as the testing data for the first simulation, while the rest of the data (9/10) were the learning data. In the second simulation, the second 1/10 of the learning data was the testing data, while the rest of them was the learning data. This repeats until the last 1/10 of the data was the testing data and the rest of the data was the

learning data. This way, learning simulation was run 10 times. Each output grammar gave the testing words scores. Since the words were all actual observed forms, it was assumed that the grammar that had the lowest total scores for the testing words performed the best. The scores from each testing set were analyzed using *R* Statistical Software. A linear regression model was fitted with scores as dependent variable and each simulation (1~10) as predictor. It turns out that scores are not significantly different across simulations ($F(9, 614) = 1.035$, $p = 0.41$). A Tukey's HSD Post-Hoc test for multiple comparison revealed that none of the simulations were significantly different from each other. This implies that the output grammar of the Phonotactics Learner is robust, and we can be ensured that any one of the multiple simulation results can be representative and reliable.

After this, learning simulations were run 10 separate times with the entire training data (624 words)[2]. The Learner settings were the same as before. The simulations found 67 constraints on average ($SD = 3.2$). The output grammars from the 10-fold cross-validation procedure contained 61 constraints on average ($SD = 2.3$). In the next section, the output grammar of one of the simulations is reported.

4. Analysis of the Resulting Grammar

The resulting grammar learned from all 624 words is reported in this section. Only the constraints concerning the vowels of interest ([ɑː], [æ]) will be reported. These are the constraints containing any of the vocoid features ([−high], [±back], [±short], [±long]) (hereafter 'vocoid constraints'). Feature [+high] is excluded because the [+high] natural class contains only glides ([w], [j]). In all 10 grammars, there were a total of 25 different vocoid constraints, with the average 13.4 (SD = 1.6) in each grammar. Among the 25 vocoid constraints, 13 constraints repeatedly occurred at least five times across all ten simulations. One of the simulations (8th) contained all of these 13 most-frequent constraints (total 15), so the results from the 8th simulation will thus be analyzed, presented in Table 1. The first column is constraints. The second column gives the description of the constraints, the weights are in the third

[2] I also ran simulations with the CELEX word frequency, but the output grammar contained 935 constraints, of which 456 constraints are concerned with [æ] and [ɑː] vowels. Thus, I only report the result of the simulations without using word frequency. It does not seem plausible for speakers to consider 456 constraints when choosing between the two vowels.

column, and the number of occurrences of a constraint across simulations is shown in the fourth column.

Table 1. Constraints in the Learned Grammar

	Constraints	Description	Weight	# of occurrences
1	*[−high][−consonantal]	Vowel−vowel sequences and diphthongs are not allowed.	3.735	10
2	*[−high][−word_boundary] [+sonorant]	Words may not end with sonorants.	2.783	1
3	*[−word_boundary] [+approximant,−back]	Words may not end with [j].	2.667	10
4	*[+back][+sonorant,+dorsal]	*[ɑːŋ] (No [ɑː]−velar nasal)	2.32	10
5	*[−consonantal][−back] [+coronal]	No VV or VG allowed. No glide in onset. After glide−[æ], coronal codas are not allowed.	2.314	10
6	*[−back][+word_boundary]	Words may not end with [j] and [æ]	2.15	10
7	*[+voice,+strident][+back] [−approximant]	[ɑː] may not occur between voiced strident and non−liquid. (e.g., *[ʤɑːt])	2.027	10
8	*[+back][+sonorant][+labial]	After [ɑː], sonorant−labial complex codas are not allowed.	1.883	6
9	*[−back] [+approximant,−anterior]	*[æɹ]	1.867	10
10	*[+back][−syllabic][+voice]	After [ɑː], codas may not be [lb], [lv], [nd], [nʤ], [ŋd].	1.854	7
11	*[+back][+approximant] [−continuant]	After [ɑː], codas may not be [lb], [lk]. [lp], [lt].	1.851	5
12	*[+back][−continuant] [−continuant]	After [ɑː], codas may not be [kt], [pt].	1.775	2
13	*[+labial][+back] [−voice,+labial]	[ɑː] may not occur between labial onset and voiceless labial coda.	1.751	5
14	*[−strident][+back] [−continuant,+labial]	[ɑː] may not occur between non−strident onset and labial stop.	1.745	9
15	*[+continuant,−strident] [−short]	[ɑː] may not occur with [ð], [θ] onsets.	1.643	9

The learned grammar includes constraints many of which represent well−known phonotactic constraints in English. The constraints are classified in terms of subunits of syllable structure: nucleus (VV and V#), rhyme and complex coda (VC(C)), and onset−nucleus (CV), and onset−coda co−occurrence (CVC) constraints. The constraints are presented in a descending order of weights in each table.

4.1 VV and V# Constraints

Table 2 shows the constraints related to a single vowel or vowel sequences. The second column shows the segments corresponding to the natural classes in the constraints. Constraints 1 and 6 are strong (high weights) and robust, discovered in all of the 10 simulations. Constraint 1 is the highest-weighted among all constraints. This constraint bans any sequences of vowels (e.g., *[ɑːɑː], *[ɑːæ]) or vowel and glide (e.g., *[ɑːw], *[æj]). The ban against VV sequences is obvious because the learning data included monosyllabic words only. The ban on *[ɑː,æ]-[w,j] is also natural because these vowels do not form a diphthong with [w] or [j]. These VV or VG sequences never occur in the learning data, so the constraint has a high weight. Constraint 6 captures a well-known phonotactic restriction in English, namely, a short vowel cannot occur in word-final position in monosyllabic words (Carr 2013: 61).

Table 2. VV and V# Constraints

	Constraints	Segments	Description	Weight
1	*[-high][-consonantal]	*[ɑː,æ] [ɑː,æ,w,j]	Vowel-vowel sequences and diphthongs are not allowed.	3.931
6	*[-back][+word_boundary]	*[æ,j]#	Words may not end with [j] or [æ].	2.15

4.2 VC(C) Constraints

Table 3 shows constraints that penalize the rhyme (VC) of a syllable. Constraint 4 indicates the restriction against the illegal sequences [ɑːŋ] and [wŋ]. This constraint captures a phonotactic restriction in English which does not allow a long vowel and velar nasal sequence. In English, a long vowel or a diphthong does not occur before velar nasal e.g. *[luːŋ], *[blaʊŋ] (Gussmann 2002: 68). In our learning data, there were 53 words that contained [æŋ] (e.g., *bang, twang, thank*), but no words appear with [ɑː,w]-[ŋ] sequences.

Table 3. VC Constraints

	Constraints	Segments	Descriptions	Weight
4	*[+back][+sonorant,+dorsal]	*[ɑː,w],[ŋ]	[ɑː] may not have coda [ŋ].	2.32
9	*[-back][+approximant,-anterior]	*[æ,j], [ɹ]	[æ] may not have coda [ɹ].	1.867

Constraint 9 penalizes the [æɹ] sequence. This is a well-established restriction in English: [æ] does not occur before a tautosyllabic [ɹ] (e.g., *carry* [kæ.ɹi] vs. *car* [kɑɹ], not *[kæɹ]), except in truncated forms (*Larry* [læ.ɹi], *Lar* [læɹ]) (Benua 1995, Kahn 1976, Kager 1999: 260)). In our learning data, there was no [æɹ] sequence, so the weight is relatively high. On the other hand, there were 14 words with [ɑːɹ] sequences, most of them exist as alternative pronunciations for [ɑːɹ], along with [ɑː] pronunciations (e.g., car [kɑːɹ], [kɑː]) in the learning data.

Table 4. VCC Constraints

	Constraints	Segments	Description	Weight
2	*[−high][−word_boundary] [+sonorant]	*[ɑː, æ][ ][l, m, n, ŋ, ɹ, w, j]	Words may not end with sonorants.	2.783
8	*[+back][+sonorant] [+labial]	*[ɑː, w][l, m, n, ŋ, ɹ, w, j][b, f, m, p, v, w]	After [ɑː], sonorant-labial complex codas are not allowed.	1.883
10	*[+back][−syllabic] [+voice]	*[ɑː, w][b, ʧ, d, ð, f, g, h, ʤ, k, l, m, n, ŋ, p, ɹ, s, ʃ, t, θ, v, w, j, z, ʒ][b, d, ð, g, ʤ, v, z, ʒ]	After [ɑː], codas may not be [lb],[lv],[nd],[nʤ],[ŋd].	1.854
11	*[+back][+approximant] [−continuant]	*[ɑː, w][l, r, w, j][b, ch, d, g, jh, k, p, t]	After [ɑː], codas may not be [lb], [lk]. [lp], [lt].	1.851
12	*[+back][−continuant] [−continuant]	*[ɑː, w][b, ch, d, g, jh, k, p, t] [b, ch, d, g, jh, k, p, t]	After [ɑː], codas may not be [kt], [pt].	1.775

Note: [ ] means any segments.

Table 4 shows another set of rhyme constraints, with complex codas. Constraint 2 is just an artifact from removing words with syllabic consonants in the learning data. The segments [+sonorant] are the ones that may serve as a syllabic nucleus in that position. In order to have only monosyllabic words in the learning data, words with the second syllable with a syllabic nucleus were all removed from the learning data (e.g., *ample* [æmp̩]).

On the other hand, Constraints 8, 10, 11, and 12 illustrate the restrictions on the consonant cluster after [ɑː]. Complicated as it may seem, not all the combinations of the segments in the constraints are permissible sequences in English. For example, among all the possible combinations from Constraint 8, [lb], [mp] are phonotactically permissible and attested, but *[mm], *[nv] are not.

As described in Section 2, constraints are added to the grammar based on their

accuracy, or the O/E, the observed number of constraint violations over the expected number of violations. A lower O/E value suggests that the number of constraint violations is small, so the constraint is strong and more likely to be added to the grammar. In other words, sequences that are penalized by strong constraints are less likely to occur in the observed data. So we consider sequence frequencies in Table 5. Table 5 shows the frequencies for complex codas after [æ] for each constraint. Complex codas with zero frequencies are omitted. None of these allowed complex codas occur after [ɑ:] (frequency of zero), while quite a few of them occur after [æ].

Table 5. Frequency of Attested Complex Codas after [æ]

Cons8	Freq	Cons10	Freq	Cons11	Freq	Cons12	Freq
[lb]	1	[lb]	1	[lb]	1	[kt]	6
[lf]	2	[lv]	2	[lk]	3	[pt]	6
[lp]	3	[nd]	13	[lp]	3		
[mp]	17	[ndʒ]	2	[lt]	2		
		[ŋd]	1				

From Table 5, it can be seen that vowel [æ] occurs more often in the environment described in these constraints. Recall that absence of [ɑ:] before [lf] is due to the historical change where ME [a] developed into [ɑ:] before a tautosylabic /l/ followed by labials (/f/, /m/), e.g., ME half [half] > PDE [ha:f] (Section 1.2).

4.3 CV(C) Constraints

Table 6 shows the constraints concerning the CV (onset-nucleus) part of a syllable. Constraint 3 indicates that any segment plus [j] is ill-formed. This constraint is not much concerned with a vowel, but with [j]. It can penalize [ɑ:,æ]-[j] sequences, none of which occur in the learning set. Constraint 15 penalizes [ðɑ:] and [θɑ:]. Neither exists in our learning data, though [ðæ] and [θæ] were found with a low frequency (2 and 3 each) (*that, than; thatch, tanh* ('hyperbolic tangent' [θæn]), *thank*).

Table 6. CV Constraints

	Constraints	Segments	Description	Weight
3	*[-word_boundary] [+approximant,-back]	*[ ][j]	Words may not end with [j].	2.667
15	*[+continuant,-strident][-short]	*[ð, θ][ɑː]	[ɑː] may not occur with [ð,θ] onsets.	1.643

Table 7 illustrates CVC constraints, concerning co-occurrence restrictions on onset-nucleus-coda sequences. Constraint 5 can best be understood to mean that, after a glide and [æ], coda must be non-coronal (e.g., *yak* [jæk], but not *[jæt]). The first natural class in Constraint 5 consists of vowels [ɑː, æ] and glides [w, j]. Considering first the vowels [ɑː] and [æ], it simply means the ban on VV and VG sequences (possible combinations of [ɑː, æ] and [æ, j]), which is similar to Constraint 1 above. The VV/VG sequences have frequencies of zero, so the weight is high. This is correct but not very informative. Glide onsets are more interesting. The phonological representations banned by the constraint are *[w][æ][+coronal] and *[j][æ][+coronal]. In the learning data, eleven instances of [wæ] sequences are observed but none of them have a coronal coda, whereas labial and velar codas appear ([m]:1, [k]: 3, [g]: 3, [ŋ]: 4) (e.g., *swag* [swæg], *twang* [twæŋ]). There are four [jæ] sequences, none of them had a coronal coda (*yak, yam, yank, yaep*), either.

Table 7. CVC constraints

	Constraints	Segments	Description	Weight
5	*[-consonantal][-back] [+coronal]	*[ɑː, æ, w, j][æ, j][ʧ, d, ð, ʤ, l, n, ɹ, s, ʃ, t, θ, z, ʒ]	No VV or VG allowed. No glide in onset After glide-[æ], coronal codas are prohibited.	2.314
7	*[+voice,+strident] [+back][-approximant]	*[ʤ, z, ʒ][ɑː, w][b, ʧ, d, ð, f, g, h, ʤ, k, m, n, ŋ, p, s, ʃ, t, θ, v, z, ʒ]	[ɑː] may not occur between voiced strident and non-liquid.	2.027
13	*[+labial][+back] [-voice,+labial]	*[b, f, m, p, v, w][ɑː, w][f, p]	[ɑː] may not occur between labial onset and voiceless labial coda.	1.751
14	*[-strident][+back] [-continuant,+labial]	*[d, ð, l, n, ɹ, t,θ][ɑː, w][b, p]	[ɑː] may not occur between non-strident onset and labial stop.	1.745

Constraint 7 suggests underrepresentedness of sequences [ʤ, z, ʒ]−[ɑ:], where the frequencies are [ʤɑ:] (2), [zɑ:] (2), and [ʒɑ:] (0). As the third natural class suggests ([−approximant]), the only observed coda is [ɹ], as alternative pronunciation (*jar* [ʤɑ:]/[ʤɑ:ɹ], *tsar* [zɑ:]/[zɑ:ɹ], *[ʒɑ:]). In summary, voiced strident fricatives rarely can be an onset to [ɑ:], and in that rare case, the only possible coda is [ɹ].

Constraint 13 is an OCP (Obligatory Contour Principle) constraint. In general the OCP prohibits adjacent identical elements (McCarthy 1986). In many languages, the same place of articulation is avoided in nearby consonants such as onset and coda (C̲VC̲), often analyzed with OCP−Place constraints (Arabic (Frisch et al. 2004), Muna (Coetzee and Pater 2008), Japanese (Kawahara et al. 2006), Latin (Berkley 2000)). English also exhibits OCP−Place effects in (s)CVC words and syllables (Berkley 2000, Taylor 2011, Oh and Hong 2013)) for labial and dorsal places. Constraint 13 captures labial OCP−Place effects when the vowel is [ɑ:]. None of such forms occur in the learning data (e.g., *[pɑ:f]). On the other hand, if the vowel is [æ], labial co−occurrence is found in two words: *map* [mæp] and *pap* [pæp].

Finally, Constraint 14 penalizes CVC sequences of non−strident coronal consonant, vowel [ɑ:], and labial stop. None of such sequences appear in the learning data. On the other hand, the same onset−coda combinations are frequently observed with [æ], as shown in Table 8.

Table 8. Frequency of [−strident][æ][b,p]

_b	Frequency	Example	_p	Frequency	Example
[dæb]	1	*dab*	[dæp]	0	−
[læb]	4	*lab, slab*	[læp]	6	*lap, flap*
[næb]	1	*nab*	[næp]	3	*nap, snap*
[ɹæb]	3	*grab, drab*	[ɹæp]	8	*rap, crap*
[tæb]	2	*tab, stab*	[tæp]	1	*tap*
[θæb]	0	−	[θæp]	0	−
[ðæb]	0	−	[ðæp]	0	−
	11			18	

Labial places are avoided after [ɑ:] even when there is an intervening segment. Recall Constraint 8 in Table 4 above. The constraint *[+back][+sonorant][+labial] penalizes sequences of [ɑ:]−sonorant−labial (e.g., *[ɑ:lb]), whereas [æ] is allowed in that environment.

5. Phonotactic Grammar for the Variable Environments

In Section 4, phonotactic constraints for the distribution of [ɑː] and [æ] were examined. Some constraints represent the well-known phonotactic restrictions in English, such as the ban on the long vowel-velar nasal sequence (*[ɑːŋ]), the word-final short vowel ([æ]#), *[æɹ], and labial OCP-Place effects. Some less well-known restrictions were found, such as the preference for [æ] before some sonorant-obstruent clusters (e.g., [lb, lf, nd, mp]), the ban on [ɑː] after non-strident fricatives [θ, ð] (Cons15) and voiced stridents [ʤ, z, ʒ] (Cons7). In all of these constraints, the banned sequences have zero frequency in the learning data (e.g., [ɑːlb] (0) [ælb] (2)), so these constraints are effectively non-violable. They describe phonological contexts where frequency of one of the vowels is zero.

However, the previous simulations did not seem to find relatively 'weak' constraints, failing to capture the patterns of variation where both vowels variably occur in the same phonological environment. For example, in *brass* [bɹɑːs] and *crass* [kɹæs], the segments immediately surrounding the vowel are identical, but the vowels are different. As mentioned, not all words that satisfy the structural description of ME [a]-lengthening (voiceless fricatives and nasal-C clusters) have undergone lengthening, so this is where the variation is mostly found.

The frequency of each vowel before voiceless fricatives in the learning data is shown in Table 9. Both vowels occur at least once. However, which vowel is more frequent is different in each context, so the generalization with the voiceless fricative natural class as a whole is not proper.

Table 9. Frequency of Voiceless Fricative Codas

	[f]#	[fC]	[θ]#	[s]#	[st]#	[sp]#	[sk]#	[ʃ]#	
[ɑː]	8	9	5	11	9	5	6	3	56
[æ]	4	1	3	32	4	1	5	24	74

Table 10 shows the frequency of nasal-C codas in our learning data (sequences of zero frequency [nz], [nʒ] were omitted). [ɑː] is more frequent when post-nasal C is strident, but [æ] is more frequent when it is a stop. However, this is not absolute, since both vowels may occur in either context.

Table 10. Frequency of Nasal-C Codas

	[nt]	[nd]	[ns]	[nʃ]	[ntʃ]	[ndʒ]
[ɑː]	5	0	7	1	5	0
[æ]	12	13	7	0	0	2

The effect of voiceless fricatives and nasal-C environments is variable and inconsistent, compared to the environments where either of the vowels never occurs. This is why the previous learning simulation did not capture the variable environment. Thus, I created a smaller data set to focus on the variation. The words end with one of the voiceless fricative or nasal-C clusters ([f, s, ʃ, θ, st, sp, sk, ns, nt, nd, nʃ, ntʃ], i.e., all the contexts in (2)), collected from the previous learning data. The new data set had 153 words. With this the simulation was run again. The maximum O/E threshold was 0.3 as before, but the grammar still did not learn the constraints for variable environment. Thus, the maximum O/E threshold is removed, which means that the Learner will find all constraints without limiting the O/E threshold so that constraints with the O/E greater than 0.3 (weaker constraints) can be added[3].

The output grammar consists of 66 constraints, 9 of which are vocoid constraints concerning [ɑː] or [æ]. The constraints concerning [ɑː] is shown in Table 11, those concerning [æ] shown in Table 12. For more accurate and simpler presentation, only the coda consonants that exist in the learning data are presented in the segment column.

Table 11. Constraints for Variable Environments for [ɑː]

	Constraints	Segments	Weight
1	*[+back][−word_boundary][+voice]	*[ɑː][n][d]	1.694
2	*[+back][−anterior]	*[ɑː][ʃ]	1.485
3	*[−approximant][+back][+sonorant]	*^[l, ɹ, w, j][ɑː][n]	0.879
4	*[+continuant][+back][+sonorant]	*[ð, f, h, s, ʃ, θ, v, z, ʒ][ɑː][n]	0.871
5	*[+dorsal][+back]	*[g, k, ŋ][ɑː]	0.565

Note: ^[ ] means the complement natural class, any segment that is not a member of the natural class [ ].

[3] An anonymous reviewer suggested running a simulation with token frequency for the variation data set. However, with the O/E threshold unspecified, the Learner did not stop even after it found more than 1800 constraints, running for six hours. As mentioned in Footnote 2, such a large constraint set cannot be a plausible grammar for the variation given the number of relevant words, so I do not report the results here. Instead, I only compare probability (well-formedness) and token frequency later in this section.

In Table 11, Constraint 1 is effectively the same as the previous constraint 10 (banning occurrence of [ɑ:] before [nd], [nʤ], and [ŋd] (See Table 5)). Vowel [ɑ:] is penalized before [nd]. Constraint 2 penalizes [ɑ:] before [ʃ], which is consistent with our knowledge that sequence [æʃ] is more frequent than [ɑ:ʃ] (e.g., *cash, ash* vs. *harsh*). Constraints 3 and 4 penalize [ɑ:] before [n], so in general [æ] is preferred before [n]. Constraint 4 also contains the same CV restriction observed in the previous simulation (constraints banning [ð, θ, z, ʒ] before [ɑ:] in Tables 5 and 6) Constraint 5 is a CV restriction, penalizing dorsal onsets before [ɑ:]. Word−initial velar nasals are prohibited, with zero frequency, but #[g, k]−[ɑ:] sequences do appear in English, e.g., *guard* [gɑ:d], *cask* [kɑ:sk]. Because these sequences are allowed though the frequency is low ([gɑ:] 5, [kɑ:] 18), the constraint has a relative low weight.

Table 12. Constraints for Variable Environments for [æ]

	Constraints	Segments	Weight
1	*[−back][−word_boundary][−anterior]	*[æ][n][ʧ, ʃ]	1.584
2	*[−back][+labial]	*[æ][f]	1.192
3	*[−voice,+labial][−back]	*[f, p][æ]	0.737
4	*[−back][−sonorant,+anterior]	*[æ][s, θ]	0.488

In Table 12, Constraint 1 penalizes [æ]−[nʧ,nʃ], so the vowel [ɑ:] is preferred instead (unlike [nd]). Constraint 2 penalizes [æ] before [f] (e.g., *staff* [stɑ:f], *half* [hɑ:f]). In Constraint 3, [f, p] onsets are not preferred before [æ]. Constraint 4 penalizes [æ] before [s,θ], but with a tiny weight, its effect is very small, allowing for variation between [æ] and [ɑ:]. To summarize, all the constraints that are concerned with vowel variation are found, mostly consistent with previous findings: [ɑ:] is preferred before [f, s, θ, nʧ, nʃ], whereas [æ] is preferred before [nd,ʃ].

As described in Section 2, the MaxEnt grammar assigns probabilities for words: the more well−formed, the more probable. Table 13 shows word probabilities assigned by the MaxEnt grammar for some words. In the score column, the score is the weighted sum of the constraint violations by each word (the smaller, the more well−formed). In the MaxEnt column, the MaxEnt value is computed by e^{-score}. The probability of each word (P(word)) is the portion of its MaxEnt value over the sum of all the MaxEnt values of the words in the entire data (the MaxEnt sum = 69.93, so P(*aunt*) = 1/69.93). Since each P(word) is very small, they are transformed to Log(P). So, the

higher Log(P) (closer to zero), the better the form.

Table 13 shows the clear difference between *brass* and *crass* in terms of probability. Recall that these words have different vowels in a nearly identical phonological environment. The probability values suggest that *brass* is more well-formed, with the vowel [ɑ:] as in other phonologically-similar words such as *class* and *grass*, than *crass* (Log(P): −2.22 vs. −2.43).

Table 13. Word Probability Assigned by the MaxEnt Grammar

		Score	MaxEnt ($=e^{-score}$)	P(word)	Log(P)
aunt	[ɑ:nt]	0	1	0.0142	−1.84
class	[klɑ:s]	0.442	0.642	0.0091	−2.03
grass	[gɹɑ:s]	0.872	0.418	0.0059	−2.22
brass	[bɹɑ:s]	0.872	0.418	0.0059	−2.22
crass	[kɹæs]	1.36	0.256	0.0036	−2.43
task	[tɑ:sk]	2.986	0.050	0.0007	−3.14

Table 14 shows how the scores in Table 13 were obtained for *brass* and *crass*. The relevant constraints and their weights are presented, and the number of violations by each word is shown under the constraints. The score is the weighted sum of the constraints violation, as shown in the score column. It is 0.872 for *brass*, and 1.36 for *crass*, since *crass* violates both constraints. The first constraint in Table 14 penalizes the occurrence of [æ] before an anterior obstruent ([s]), which is violated by *crass* but not by *brass*. Both words violate the second constraint equally, because the immediate phonological environments are the same ([ɹ]_[s]).

Table 14. Constraint Violations and Scores for *brass* and *crass*

	*[−back][−sonorant,+anterior]	*[−anterior][−word_boundary][−sonorant,+anterior]	Score
	0.488	0.872	
[bɹɑ:s]		1	0.872
[kɹæs]	1	1	0.488+0.872=1.36

In order to examine whether token frequency of each word can distinguish *crass* from other general patterns even more clearly, CELEX counts (COBUILD log frequency) were collected from WebCelex[4]. Table 15 shows the well-formedness, or the

[4] http://celex.mpi.nl

probability of words (Log(P(word)) and their token frequency (Log Celex Frequency). Probability decreases as frequency decreases. The results for the entire variation data are shown in Figure 1.

Table 15. Probability of Words vs. Token Frequency

		Log(P)	Log CELEX Frequency
class	[klɑːs]	−2.03	2.32
grass	[gɹɑːs]	−2.22	1.91
brass	[bɹɑːs]	−2.22	1.27
crass	[kɹæs]	−2.43	0

Figure 1 shows the plot of probability against frequency of words. The fitted line was obtained by second-order polynomial regression. Overall, there is a weak positive correlation between frequency and probability (Adjusted R^2 = .009). The more frequent, the more probable (well-formed), though the relationship is not strict ($F(2, 123) = 1.60$, $p = 0.205$). Probability is highly variable in lower frequency words. The words with coda [s] mostly follow the regression line. For these words, [ɑː] occurs with higher frequency words (*class, glass, pass, grass, brass*), whereas [æ] occurs with lower frequency words (*ass, bass, lass, crass*). In particular, *crass* has a lower probability than the others. Although the ME [a]-lengthening is not a rule of the PDE, the frequency difference in the vowel distribution suggests that the lengthening might have taken place in more frequent words than in less frequent words, in line with Bybee (2001) and Coetzee and Kawaraha (2013).

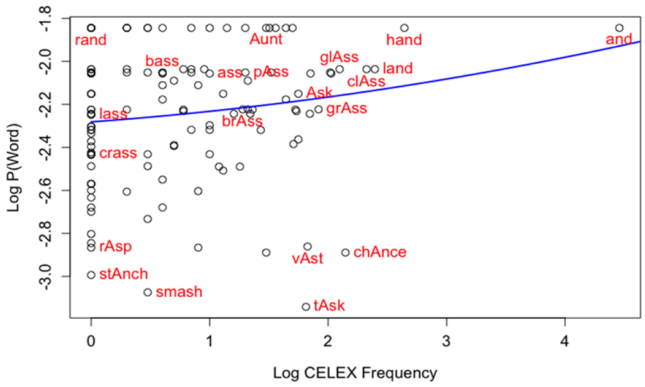

Figure 1. Log Probability of Words against Log CELEX Frequency.
(Capital 'A' indicates [ɑː], and small 'a' [æ].)

6. Conclusion

This paper examined the phonotactic grammar for the distribution of TRAP-BATH vowels ([æ] and [ɑ:]) in British English, based on the constraints obtained from learning simulations with 624 monosyllabic words. The constraints captured nearly absolute bans as well as mere preference in the phonotactic grammar. The phonological environment where the vowels variably occur (voiceless fricatives and nasal-C clusters) was not captured by the constraint found with the entire learning set, because the effects of these environments were variable and inconsistent. As a result, the relevant constraints were not strong enough to be included in the grammar. Thus the learning simulation was conducted with the learning data with the variable environments only. The patterns of variation were captured by the constraints, assigning different probabilities reflecting the well-formedness of the forms.

The findings in this research suggest that a probabilistic grammar that defines the probability distribution over the phonological forms is adequate for modeling variable phonotactic distribution such as this. The resulting grammar reflects the gradient nature of the variation. Furthermore, the variable distribution can be partly accounted for by usage frequency, in that more frequent words tend to appear with [ɑ:] vowels, which is the trace of historical ME [a]-lengthening. In conclusion, different vowels in the same environment (e.g. *brass-crass*) can be adequately modeled in terms of differences in probability, revealing that *brass* is more 'regular' than *crass,* and the differences may be due to differences in their usage frequency.

References

Albright, A., A. Andrade and B. Hayes. 2001. Segmental environments of Spanish diphthongization. In A. Albright and T. Cho, eds., *UCLA Working Papers in Linguistics 7: Papers in Phonology* 5, 117-151. LA.: UCLA.

Albright, A. and B. Hayes. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition* 90, 119-161.

Anttila, A. 1997. Deriving variation from grammar. In F. Hinskens, R. van Hout and L. Wetzels, eds., *Variation, Change, and Phonological Theory*, 35-68. Amsterdam: John Benjamins.

Barber, C. 1997. *Early Modern English*. Edinburgh: Edinburgh University Press.

Benua, L. 1995. Identity effects in morphological truncation. In J. Beckman, L. Walsh Dickey and S. Urbanczyk, eds., *Papers in Optimality Theory: University of Massachusetts Occasional Papers in Linguistics* 18, 77-136. Amherst, Mass.: Graduate Linguistic Association.

Berkley, D. 2000. *Gradient Obligatory Contour Principle Effects.* Doctoral dissertation, Northwestern University.

Bybee, J. L. 2001. *Phonology and Language Use*. Cambridge: Cambridge University Press.

Carr, P. 2013. *English Phonetics and Phonology.* Oxford: Wiley-Blackwell.

Cho, H. 2012. Statistical learning of Korean phonotactics. *Studies in Phonetics, Phonology, and Morphology* 18(2), 339-370.

Clements, G. N. and E. Hume. 1995. The internal organization of speech sounds. In J. Goldsmith, ed., *The Handbook of Phonological Theory,* 245-306. Oxford: Blackwell.

Coetzee, A. W. and S. Kawahara. 2013. Frequency biases in phonological variation. *Natural Language and Linguistic Theory* 31(1), 47-89.

Coetzee, A. W. and J. Pater. 2008. Weighted constraints and gradient restrictions on place co-occurrence in Muna and Arabic. *Natural Language and Linguistic Theory* 26, 289-337.

Coetzee, A. W. and J. Pater. 2011. The place of variation in phonological theory. In J. Goldsmith, J. Riggle, and A. Yu, eds., *Handbook of phonological theory: 2nd Edition*, 401-434. Cambridge: Blackwell

Frisch, S., J. Pierrehumbert and M. Broe. 2004. Similarity avoidance and the OCP. *Natural Language and Linguistic Theory* 22, 179-228.

Gussmann, E. 2002. Domains and phonological regularities. In E. gussmann, ed., *Phonology: Analysis and Theory*, 45-65. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139164108.004

Gupta, A. F. 2005. Baths and becks: A report on two prominent dialectal variables in England, *English Today* 81(21), 21-27.

Halle, M. 1995. Feature geometry and feature spreading. *Linguistic Inquiry* 26, 1-46.

Hayes, B. and Z. C. Londe. 2006. Stochastic phonological knowledge: The case of Hungarian vowel harmony. *Phonology* 23, 59-104.

Hayes, B. and C. Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39(3), 379-440.

Jun, J. 2010. Stem-final obstruent variation in Korean. *Journal of East Asian Linguistics* 19(2), 137-179.

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

Kager, R. and J. Pater. 2012. Phonotactics as phonology: Knowledge of a complex restriction in Dutch. *Phonology* 29(1), 81-111.

Kahn, D. 1976. *Syllable-Based Generalizations in English Phonology.* Doctoral dissertation, Massachusetts Institute of Technology, MA, USA.

Kawahara, O., H. Ono and K. Sudo. 2006. Consonant co-occurrence restrictions in Yamato Japanese. *Japanese/Korean Linguistics* 14, 27-38.

Kenstowicz, M. 1994. *Phonology in Generative Grammar.* Oxford: Blackwell.

Kiparsky, P. 1993. *An OT perspective on phonological variation*. Ms. Stanford University. Paper presented at the Rutgers Optimality Workshop. October, 1993. Retrieved from http://www.stanford.edu/~kiparsky/Papers/nwave 94.pdf

Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence.* San Mateo, CA: Morgan Kaufmann, 2(12), 1138-1143.

Labov, W. 1969. Contraction, deletion, and inherent variability of the English copula. *Language* 45, 715-762.

McCarthy, J. 1986. OCP Effects: Gemination and antigemination. *Linguistic Inquiry* 17, 207-63.

Migglestone, L. 1995. *Talking Proper: The Rise of Accent as Social Symbol.* Oxford: Claredon Press.

Oh, Y-.L. and S.-H. Hong. 2013. A noisy harmonic grammar analysis of gradient OCP effects in English syllables. *Studies in Phonetics, Phonology and Morphology* 19(3), 433-455.

Pierrehumbert, J. 1994. Syllable structure and word structure: A study of triconsonantal clusters in Englis. In P. Keating, ed., *Phonological Structure and Phonetic Form: Papers in Laboratory Phonology III*, 168-188. Cambridge: Cambridge University Press.

R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org

Robinson, T., J. Fransen, D. Pye, J. Foote and S. Renals. 1995. WSJCAMO: A British English Speech Corpus for large vocabulary continuous speech recognition. In

Proceedings of ICASSP (International Conference on Acuoustics, Speech, and Signal Processing) (Detroit, MI), 81-85.

Robinson, T. 1995. *British English Example Pronunciation Dictionary (BEEP)*. Retrieved from http://svr-www.eng.cam.ac.uk/comp.speech/Section1/Lexical/beep.html

Sholes, R. 1966. *Phonotactic Grammaticality*. The Hague: Mouton.

Taylor, B. 2011. Why not "Spop"? OCP and prominent position effects on the English lexicon. In *Proceedings of the GMU Working Papers in Linguistics 8.* Retrieved from http://www.gmu.edu/org/lingclub/WP/texts/8_Taylor.pdf

Wells, J. C. 1982. *Accents of English*, Vols. I-III. Cambridge: Cambridge University Press.

Examples in: English
Applicable Languages: English
Applicable Level: Tertiary

Cho, Hyesun
Professor, Dankook University
Department of Education
Graduate School of Education
152, Jukjeon-ro, Suji-gu, Yongin-si
Gyeonggi-do, Korea
Tel: 031) 8005-3968
E-mail: hscho@dankook.ac.kr

Received: February 11, 2019
Revised: March 10, 2019
Accepted: March 19, 2019

Appendix

Feature Chart

	syllabic	consonantal	approximant	sonorant	continuant	nasal	voice	spread	labial	coronal	anterior	strident	lateral	dorsal	high	back	short	long
b	−	+	−	−	−		+		+									
ʧ	−	+	−	−	−		−			+	−	+						
d	−	+	−	−	−		+			+	+	−						
ð	−	+	−	−	+		+			+	+	−						
f	−	+	−	−	+		−		+									
g	−	+	−	−	−		+							+				
ʤ	−	+	−	−	−		+			+	−	+						
k	−	+	−	−	−		−							+				
p	−	+	−	−	−		−		+									
s	−	+	−	−	+		−			+	+	+						
ʃ	−	+	−	−	+		−			+	−	+						
t	−	+	−	−	−		−			+	+	−						
θ	−	+	−	−	+		−			+	+	−						
v	−	+	−	−	+		+		+									
z	−	+	−	−	+		+			+	+	+						
ʒ	−	+	−	−	+		+			+	−	+						
h	−	+	−	−	+		−	+										
m	−	+	−	+		+			+									
n	−	+	−	+		+				+	+	−						
ŋ	−	+	−	+		+								+				
l	−	+	+	+						+	+	−	+					
ɹ	−	+	+	+						+	−	−						
w	−	−	+	+					+						+	+		
j	−	−	+	+											+	−		
ɑː	+	−													−	+	−	+
æ	+	−													−	−	+	−