

Frequency Effects in Morphologisation of Korean /n/-Epenthesis

Ogyoung Lee & Vsevolod Kapatsinski
University of Oregon

ABSTRACT

This study accounts for Korean /n/-epenthesis from a usage-based perspective, by describing the reduced productivity of epenthesis as an analogical change in progress. We found that epenthesis probability rises as whole-word frequency increases, supporting the hypothesis that analogical change begins in low-frequency words (Bybee 2002). We interpret the findings as support for the idea that frequent forms are stored and retrieved in production directly while rare words may be derived using grammar. The results further support the existence of morphological strata in Korean. We show that the constituents undergoing /n/-epenthesis are largely limited to native rather than Sino-Korean morphemes. However, not all native morphemes are able to trigger /n/ epenthesis. We argue that particular native morphemes are associated with, and able to trigger, epenthesis to the extent that they tend to occur in epenthesis-favoring contexts (Bybee 2002, Raymond & Brown 2012).

Keywords: frequency effects, productivity, lexicalisation, lexical diffusion, whole-word retrieval, computation, strata, sound change

1 Introduction

Language is not static but constantly in the process of change. According to usage-based linguistics, the grammar acquired by a language user emerges out of generalisation over experienced utterances. Phonological and morphological schemas, rules, and constraints, the elements of phonological grammar, emerge from generalisation over words. Usage-based linguistics assigns a major role in this process to frequency of use (Bybee 2001: 57). Generalisations supported by many words (i.e. ones of high type frequency) are, other things being equal, more productive (Bybee 1995, 2001). At the same time generalisations exemplified by frequent words (words of high token frequency) tend to be less productive (Bybee & Brewer 1980, Hay 2001, 2003). The basic reason is that frequent words are likely to be accessed whole, rather than through the parts (Alegre & Gordon 1999, Hay 2001, 2003, Hay & Baayen 2001, Kapatsinski 2010a, Kapatsinski & Radicke 2009, Tremblay & Baayen 2010). Therefore the listener is unlikely to parse grammatical patterns out of frequent words in perception, making patterns exemplified by frequent words less salient, and the speaker is unlikely to practice grammatical patterns when producing frequent words. When a pattern loses productivity and is replaced by a competitor (what Bybee, 2001, calls ‘analogical change’), the few words that exceptionally continue to obey the old pattern tend to be of high frequency. For example, as Old English vowel-changing past tense schemas were replaced by -ed, a few frequent verbs, now known as irregular verbs, resisted. Thus, *break-broke* but *stake-staked*. The reason again is that the past tense forms of frequent verbs could be directly retrieved from the lexicon without the need to use now unproductive vowel change schemas they seem to support.

In this paper, we document that a Korean grammatical pattern, called /n/-epenthesis, has morphologised and is now obeyed only in front of one of three ‘special’ morphemes. Following

Bybee (2002), Kapatsinski (2010b), Raymond & Brown (2012) and Torres Cacoullos (1999), we argue that these morphemes have become associated with /n/ epenthesis because they often occur as second constituents of compounds following a consonant-final initial constituent, the morphophonological environment favoring /n/ epenthesis. We show that even in front of a ‘special’ morpheme, /n/ epenthesis is not fully productive, being more likely in frequent words than in rare words. We argue that this loss of productivity is due to competition with the Korean Coda Rule, a pattern of broader scope and consequently higher type frequency.

Korean /n/-epenthesis prefixes an /n/ to a morpheme (word or suffix) that begins with /i/ or /j/ when the immediately preceding morpheme (word or prefix) ends in a consonant. That consonant itself also changes into an /n/, resulting in a geminate. This rule features prominently in descriptions of Korean phonology (Choi 2002, Eom 1999, Kang 2003, Kim-Renaud 1991, Sohn 2001) and is prescribed in Clause 29 of the second part of *Standard Pronunciation of the Regulations on Standard Korean, Regulations on Spoken and Written Korean, 88-1*, issued by the Korean Ministry of Education¹. The linguistic context that triggers the rule is described in (1):

/n/ epenthesis

C → n.n / ____ +{i,j}

If a morpheme begins with /i/ or /j/, and is preceded by another morpheme that ends in a consonant, then that consonant changes into a geminate /n/.

It has been repeatedly reported that /n/ epenthesis is not fully productive, having many exceptions and irregularities. Lee (1996) noted that the application of /n/ epenthesis depends highly on individual speakers and words. Research on the use of standard pronunciation (Kim, Kwon, & Hwang 2003) also reported that both between-speaker and within-speaker variability is evident in actual pronunciation regardless of the regulations on standard pronunciation (p. 23). It also demonstrated how the same linguistic context could result in varying pronunciation practices depending on individual words as well as speakers (Kim *et al.*, 2003: 22-32, 65-68). Traditional literature has treated this as a rule but the observed failures of application suggest that either the context for the rule’s application has not been specified in sufficient detail or the rule is lexicalised and at best only partially productive.

It was also observed in a pilot study that some of the words listed in Sohn (2001) did not always undergo the rule even when they met the rule’s structural condition. For instance, *kkoch+i.rum* ‘flower+name’ can be pronounced as /k*on.ni.rum/ with an additional /n/, which nasalises the preceding consonant, or as /k*o.ti.rum/ without an /n/, as in Table 1.

Table 1. Two instances that did not seem to always undergo /n/-epenthesis

Word	Meaning	Sohn (2001)	Alternative
<i>kkoch+i.rum</i>	‘name of a flower’	/k*on-ni.rum/	/k*o-ti.rum/
<i>os+ip.ko</i>	‘wearing clothes’	/on-nip.k*o/	/o-tip.k*o/

One might suspect that the partial productivity is the result of competition with an alternative rule (e.g., Albright & Hayes 2003, Kapatsinski 2010b). A search of 36 Korean-language textbooks used from 1984 to 2012 revealed two rules competing with /n/ epenthesis, the Korean Coda Rule and the Palatalisation Rule (see 2.1), which were introduced to children acquiring Korean during middle school and high school. As the three rules apply in the same phonological context -- a coda followed by a vowel -- /n/-epenthesis is expected to have reduced productivity. We test productivity of /n/-epenthesis in a spoken elicited production study (“wug

test”, Berko 1958), in which participants are asked to pronounce written nonce words.

2 Prior treatments of /n/-epenthesis

Both previous descriptive studies and prescriptive textbooks suggest that /n/ epenthesis is lexicalised rather than fully productive. However, the roles of lexical strata and word frequency in the process remain controversial. Thus, Lee (1996, pp.167-170) stated that /n/ epenthesis occurred more in some words/phrases than others but also suggested that words of Sino-Korean origin tend not to undergo /n/ epenthesis (pp.167-171). However, Kim et al. (2003) observed that even words with the same phonological, etymological and morphosyntactic characteristics differ vastly in probability of /n/-epenthesis. For all 62 target words in their study, the probability ranged from 0.86% (*kwe.tam+i.seol* ‘strange story’) to 96.86% (*nam.won+yeo.jung* ‘Namwon Middle School’). The 29 native Korean compounds varied between 1.14% probability of /n/ epenthesis (i.e. *twes.pak+i.ma* ‘a round forehead that looks like a gourd bowl’) and 94.86% (*ya.keum+ya.keum*, ‘(eating) little by little’). The 7 native derived nouns varied from 10.86% (*pon+im.ja* ‘original owner’) to 93.14% (*ssang+yok* ‘swearword’). The 11 Sino-Korean compounds varied from 0.86% (*kwe.tam+i.seol* ‘strange story’) to 96.86% (*nam.won+yeo.jung* ‘Namwon Middle School’). Finally, the 15 Sino-Korean derived nouns varied from 14% (*kong+yeom.pul* ‘empty prayer’) to 89.71% (*pul+yu.kwe.ha.ta* ‘to be unpleasant’). Thus word formation types did not play a determining role in Kim et al.’s data: rule application probability varied across the entire probability spectrum for both native and foreign source words, whether they were compounds or simple nouns bearing derivational suffixes.

To examine how Korean children are taught about /n/ epenthesis, and how the process is viewed in the prescriptive literature, we examined 36 Korean-language textbooks that were used from 1984 to 2012, the years when our participants were in school. We found /n/ epenthesis to be either not presented or presented alongside competing rules that are applicable in the same phonological context: the Korean Coda Rule (i.e., resyllabification of a coda of a syllable making it the onset of the next syllable that does not have an onset, e.g., /k^hi.u.ki/ for k^hi.u.k+i ‘letter k^hi.u.k + NOM, the letter k^hi.u.k is’) and the Palatalisation Rule (e.g., /k*o.tʃi/ for k*o.t+i ‘flower+NOM, the flower is’). Furthermore, when presented, examples of /n/ epenthesis were interspersed with examples of words undergoing competing rules. Thus, students were expected to memorise which words undergo which rule, suggesting that the rules are lexicalised.

For example, the nation-wide textbook during the 5th National Curriculum from 1991 to 1996) introduced 7 words for /n/-epenthesis (pp. 91-94), along with other competing examples that share the same phonological environment (i.e., coda followed by /i/ or /j/) but are to be pronounced following either the Korean Coda Rule or the Palatalisation rule (when the coda is one of the /t, t^h, t*, ts/). The seven words were: *mul+yak* ‘liquid medicine,’ *se.ul+yeok* ‘Seoul station,’ *hot+i.pul* ‘single blanket,’ *mak+il* ‘physical work,’ *a.res+i* ‘lower teeth,’ *neuk.mak+yeom* ‘pleurisy,’ and *ne.pok+yak* ‘a kind of medicine’. These were interspersed among examples of words that undergo the Korean Coda Rule (i.e., *ssah+i.da* /s*a.i.da/ ‘to accumulate,’ *ti.keut+i* /ti.ku.si/ ‘letter *ti.keut* NOM,’ *ki.euk+i* /ki.u.ki/ ‘letter *ki.euk* NOM,’ *hi.euh+i* /hi.u.i/ ‘letter *hi.euh* NOM,’ *yuk+i+o* /yu.ki.o/ ‘the Korean War’ (literally ‘6.25’), *song+pyeol+yeon* /soŋ.pjʌ.tʃʌn/ ‘a farewell party,’ *teung+yong+mun* /tuŋ.joŋ.mun/ ‘gateway’) and one word that undergoes Palatalisation Rule (i.e., *kkot+i* /k*o.tʃi/ ‘flower_{NOM}’). This interspersing of examples of the three rules was also the case for all high-school textbooks on Korean Grammar that we found.

Middle-school curricula during the other time periods either did not deal with pronunciation (i.e., those during the 4th and the 6th curriculum) or introduced only examples of

competing rules (i.e., palatalisation examples during the 7th curriculum, pp. 83-84, e.g., *kut+i* ‘obstinately,’ *pyeot+i* ‘sunshine_{NOM}’). In high schools throughout the investigated time frames, what we call /n/-epenthesis was presented as an example of Sais-Sori Phenomenaⁱⁱ ‘in-between sound phenomena’ in the Korean Grammar textbook.ⁱⁱⁱ Koreans thus were taught about /n/-epenthesis if their school chose to teach the optional subject ‘Grammars’.^{iv} Again, consistently with lexicalisation, the textbooks explained that the Sais-Sori Phenomenon apply to some words but not others. However, they noted that Sino-Korean words would generally not undergo the alternation (all the 4th, 5th, 6th, and 7th generation textbooks, e.g., pp. 165-166, the 4th Curriculum High School Grammar), suggesting an effect of lexical strata (e.g., Chomsky & Halle 1968, McCawley 1968).

While previous research thus agrees on the fact that the /n/ epenthesis pattern is lexicalised, there is disagreement on which kinds of words are likely to undergo it, with most researchers suggesting that the origin of the word (Sino-Korean vs. native) has a role, but cf. Kim et al. (2003). In this paper, we revisit the role of origin and additionally investigate the role of frequency of use, for both words and their parts. With respect to the role of word frequency, Kim et al. (2003) considered the possibility that frequent words are more likely to undergo /n/ epenthesis than infrequent words. However, they rejected this possibility on the basis of observing a very high rate of /n/ epenthesis with the compounds *nam.won-yeo.jung* ‘Namwon Middle School’ and *pul+yu.kwe.ha.ta* ‘to be unpleasant’, which they considered to be infrequent (Kim et al. 2003, p.32). However, Kim et al. did not evaluate the predictiveness of frequency statistically, and did not obtain any empirical measures of frequency, instead using their intuition to decide on whether a word was frequent or infrequent. Furthermore, it may well be the case that the behavior of these words might be explained by the frequencies or identities of individual morphemes within them. For instance, the high probability of /n/-epenthesis for the compound *nam.won-yeo.jung* ‘Namwon Middle School’ could be due to the second stem of the compound. If so, compounds derived using the same second stem combined with other first stems would behave similarly to the tested one. As a pilot, we asked a few native speakers of Korean how they would pronounce a set of words, where we combined different first morphemes with the same second morpheme *yeo.jung* ‘middle school’ and different second morphemes with the same first morpheme *nam.won* ‘Namwon.’ We found that speakers epenthesised an /n/ when the second morpheme was ‘middle school,’ which suggests their pronunciation depended on the identity of the second morpheme.

In addition, frequency of a word or morpheme in a specific sense might be important. Gahl (2008) showed that durations of ostensibly homophonous English words like *time* and *thyme* were significantly different, with the more frequent homophone having a shorter duration. Kim et al. (2003) tested a lexical form *kheun+il* in two different senses, ‘a big matter/happening’ and ‘an important household matter.’ Their conclusion was that the same linguistic form might behave differently depending on the semantic content. The word in the former sense had 38% of /n/-epenthesis whereas the same form in the latter sense received an /n/ 61% of the time (p. 27). However, they did not explain how the different meanings resulted in different probability of /n/-epenthesis. Plausibly the probability of /n/ epenthesis differs for the two meanings of the word because the different meanings occur with different frequencies (as argued by Gahl 2008). What we can say for certain is that probability of /n/-epenthesis is not always constant across all compounds containing the same second stem. Thus another compound with the same second stem *il* ‘work’, *mak+il* ‘physical work’, had 51% of /n/-epenthesis. To explain the differences in /n/ epenthesis probability across the various compounds containing *il* ‘work’ we need to look beyond the identity of the second morpheme. Again, one plausible influence is whole-word

frequency; perhaps, sense-specific frequency.

3 Method

3.1 Material

The materials for this study were a total of 67 words embedded in a carrier sentence. The 67 words came from 5 word-formation types. All words consisted of two morphemes, the first of which ended in one of 7 codas (/p, t, k, m, n, ŋ, l/) while the second one started with /i/. 53 words were real words of Korean while the other 14 were nonce words. We used both native and loan real morphemes. Korean native morphemes were either native bound (NB) or native free (NF). Loan morphemes were all Sino-Korean stems, which are inherently free (FF). Nonce words (Type H) are compound nonwords, whose first constituent is a nonce morpheme (NC)^v followed by a native free morpheme (NF). Table 2 summarises the organisation of the stimuli. The following paragraphs describe in more detail how the materials were designed.

Table 2. Structure of the 67 lexical stimuli

Word Formation			Real Words				Nonce NC+NF	Total
			NB+NF	NF+NF	NF+NB	FF+FF		
Coda preceding /i/								
Stop	Labial	/p/	2	2	2	2	2	10
	Alveolar	/t/	4	2	2	0	2	10
	Velar	/k/	2	1	2	2	2	9
Nasal	Labial	/m/	0	1	2	4	2	9
	Alveolar	/n/	3	2	2	2	2	11
	Velar	/ŋ/	0	1	2	2	2	7
Lateral		/l/	3	2	2	2	2	11
Total			14	11	14	14	14	67

Note. The first letter of the abbreviations for the word formation types denotes the source, whether the word is native or foreign: N refers to a native morpheme and F to a foreign (Sino-Korean) morpheme. The second letter is for morpheme type, indicating whether the morpheme is bound (B) or free (F), if not nonce (C). ‘+’ marks a morpheme boundary. Numbers are the numbers of words. Three words were excluded from the NF+NF type due to experimenter error.

The phonological context focused on in this study involved 7 codas of the first morpheme and /i/ as the initial sound of the second morpheme. While the literature says both /i/ and /j/ trigger the rule (Sohn 2001: 167), /i/ was reported to involve more variation in /n/-epenthesis than /j/ (Lee 1996: 167-169).^{vi} The coda of the preceding morpheme was one of /p, t, k, m, n, ŋ, l/. These constitute all possible codas in Korean.

Native (NB, NF) vs. loan source (FF) morphemes were distinguished to examine whether the speakers’ /n/-epenthesis probability varies depending on whether the word/morpheme is native or loan source. Free and/or bound morphemes are combined in three ways to see whether the type of the first morpheme and/or that of the second morpheme affects rule application. The formation types were bound followed by free (NB+NF), free followed by free (NF+NF and FF+FF), and free followed by bound (NF+NB).

Fourteen nonce words (NC+NF) were included to test productivity of Korean /n/-epenthesis. Our nonce words consist of a nonce first morpheme (NC) and a following real native free (NF) morpheme. We used two real morphemes, *ip* ‘leaf’ and *i.rum* ‘name’.

The words were embedded in a carrier sentence *han.peon.ssik ~ ham.ni.ta* ‘I say ~ once’ and then randomised in a list by means of Excel RAND function, where real words (Part A of the

experiment) were grouped into a separate block from the nonce words (Part B). The list was presented in Korean orthography. The meanings of the real words were defined to the right of each sentence in order to allow for different probabilities of /n/ epenthesis in homophones.

3.2 Participants

53 adult native speakers of Korean participated in the experiment. They were undergraduate and graduate students from a university in Seoul, Korea, 23-35 years old. All were born and have lived in Korea. They reported that they had not stayed abroad for more than six months. Their exposure to foreign languages was self-reported to be minimal enough not to affect their Korean. They spoke in Seoul accent, which is the standard Korean accent. No participant reported any hearing or speech problem.

In order to reduce sampling bias towards an overly specific group that might not represent the pronunciation patterns of young native speakers of Korean (in their twenties or thirties), we tried to recruit speakers of different social background; particularly, with respects to gender, major, and dialect (c.f. Labov *et al.* 1991). 31 of the participants were male and 22 were female. They were studying humanities ($n = 20$), engineering ($n = 16$), or physical science ($n = 17$), majors that belong to one of the three primary divisions in the Korean educational curricula. The participants were exposed to different accents during childhood, as they were raised in different areas: Choongcheng ($n = 1$), Jeonra ($n = 3$), Kangwon ($n = 1$), Kyongsang ($n = 7$), or Seoul ($n = 41$). However, there were no significant interactions between social and linguistic predictors, which motivates treating all participants as a single group in the rest of the paper.

3.3 Data Collection Procedure

Participants were recruited at a university in Seoul, Korea. Prior to recording, participants were given the word list and invited to check through the list to see if they know the listed words in the sense given next to each word. The lexical meanings were clarified when the participants reported some confusion. Then, the participants were asked to read aloud the sentences three times, going through the entire list thrice from beginning to end. They were recorded individually by the same researcher. Recording was done by means of Sound Devices 722 (recording device) and AGK C420 microphone. The sampling rate was 44,100Hz (16bit).

3.4 Analyses

3.4.1 Spoken lexical frequency estimates

Prior to main analyses, the spoken lexical frequency of the stimuli was determined by means of subjective ratings of lexical frequency^{vii}. This method of collecting subjective measures of frequency has been widely used to collect information about lexical frequency (e.g., Balota *et al.* 2001, Gernsbacher 1984, Gordon 1985). Based on the literature, we considered that, in the absence of a large spoken corpus of Korean, rated frequency is the best way to approximate the frequencies of the words in our participants' spoken experience. For instance, Gernsbacher (1984), in six experiments attempting to resolve inconsistent results from two decades of word recognition studies, demonstrated that subjectively rated experiential familiarity but not written lexical frequency was a reliable predictor of word recognition. Balota *et al.* (2001) have argued that subjective frequency estimates more accurately reflect the actual relative frequencies of exposure to a set of words and are less affected by sampling biases, compared to objective frequency counts based on written corpora and traditional familiarity ratings.

We followed Balota *et al.*'s (2001) recommendations for collecting subjective frequency estimates. Discussing some advantages of and concerns about familiarity ratings, Balota *et al.* (2001) suggested that (1) participants should be explicitly asked to rate how frequently they have encountered the words as a way to minimise the influence of semantic, orthographic, and/or phonological factors on the familiarity ratings; and (2) participants should differentiate the source of exposure: seen, heard, written, or spoken.

The subjective frequency estimates of the words were collected from a separate group of 21 native speakers of Korean, who did not participate in the production experiment but are of equivalent demographic background (age, education, gender, dialect) to those in the production experiment. We specifically asked the respondents to rate how frequently they have actually heard each of the presented words being spoken. The subjects rated the words with respect to the frequency of encounters on a 20-point scale, ranging from 'Never' (coded as 1) to 'Every day'

(coded as 20). We averaged the 21 subjects' answers to get a spectrum of whole-word spoken frequencies. In order to test the reliability of the responses, we presented 12 words twice during the ratings task.^{viii} Three words were randomly selected from each of the four word-formation types: NB+NF, NF+NF, NF+NB, and FF+FF. The participants' responses were fairly consistent across repetitions ($t(251) = -1.772, p > .05$).

The ratings were supplemented with corpus frequencies to estimate how often the words and stems occur in epenthesis-favoring vs. disfavoring contexts. We searched the morphologically tagged modern spoken and written Korean corpus of the 21st Century Sejong Project, accessible at the website of the National Institute of the Korean Language. We retrieved the frequencies with which each target morpheme was used (i) as the second/third component of whole words and (ii) as an independent word. Within the two search result sets, we also calculated the number of tokens eligible for /n/ epenthesis and the number of tokens that were not eligible. The eligible cases were those in which the target morpheme immediately followed a consonant. The non-eligible cases were those in which the target morpheme comes after a non-consonant, including vowels, punctuation, or orthographic indications of a juncture.

3.4.2 Determination of /n/-epenthesis application

A total of 10,653 collected tokens (53 speakers * 67 words * 3 repetitions) were analyzed in Praat (Boersma & Weenink 2013) by one of the researchers (i.e., Lee) to determine whether /n/-epenthesis applied. For words whose initial morpheme ended in an obstruent, an /n/ was coded as having been epenthesised when there was a nasal voice bar (strong low frequency energy) in the targeted location.

For words whose first morpheme ended in /n/ prior to epenthesis, existence of a geminate /n/ was considered evidence of epenthesis. For instance, the word /mæn+ip/ 'bare + mouth' 'for free' should have a geminate /n/, produced as /mæn.nip/, if /n/ epenthesis applied. While the distinction between singletons and geminates was primarily determined by ear, acoustic correlates were also consulted in order to resolve uncertainty. We compared the durations of each speaker's realisation of single onset or coda /n/'s to the duration of the target /n/ to determine whether the /n/ was a singleton or a geminate. Geminates have longer consonant duration than matched singletons (Han 2008, Idemaru & Guion 2008, 2010, Oh & Redford 2012), although the specific duration ratio of singletons to geminates varies across languages (e.g., 1:3 for Japanese and 1:1.8 for Italian, in Idemaru & Guion 2008). While there have been arguments for other acoustic correlates, we mainly relied on duration to distinguish singletons from geminates in our data because Oh and Redford (2012) found that the other cues are not very reliable for heteromorphemic word-internal geminates, which is our case: they rarely show pitch changes or pauses that could be present in geminates across word boundaries.

For words whose first morpheme ended in /l/, /n/ epenthesis was seen to have applied if either /n/ or /l/ was inserted as the onset of the initial syllable of the second morpheme. For example, after /n/ epenthesis, a word /sol+ip/ 'pine + leaf, leaf of a pine tree' can be pronounced either as /sol.nip/ or /sol.lip/. Following Sohn (2001), we assume the latter pronunciation to be the product of liquidisation of the epenthesised /n/.

Coding reliability was assessed by means of interrater agreement: a male Seoulian (age 27) listened to 1,500 randomly selected samples without using Praat and marked whether epenthesis applied. The two native speakers agreed on all but 10 instances, where the speakers did not articulate the stimuli clearly.

3.4.3 Statistical modeling

Following Kuperman & Bresnan (2012) and Barth & Kapatsinski (2014), we used multimodel inference (operationalised in the MuMIn R package, Bartoń 2013) in combination with mixed-effects regression (operationalised in the lme4 R package, Bates 2013). The dependent variable was rule application. Since it is binary (1=application, 0=non-application), we used binary logistic regression (with logit link function). We included random intercepts for participants and either items or individual first and/or second morphemes. In addition, random slopes for the predictors of interest were also included as outlined in more detail within the results section, following the recommendation to maintain maximal random effects structure to minimise Type I errors (Barr *et al.* 2013). Fixed factors included word-formation type (Type A ~ Type F) and judged frequency (of the whole-words).

Multimodel inference was used to deal with model selection uncertainty (Burnham & Anderson 2002). With multimodel inference, instead of selecting the single best regression model, we derive regression coefficients by averaging them over all regression models that can be derived by selecting a subset of our predictors (including the full set), with each model weighted by its predictiveness (measured using the Akaike Information Criterion, as modified for small samples by Sugiura 1978). Thus, for example, the regression coefficient for judged frequency is the average of the coefficient from a model that included no predictors (zero), a model that included only compound type (where the coefficient for frequency is still zero), a model that included only word frequency (where the coefficient is maximally different from zero), a model that included all predictors, etc. The degree to which each model influences the estimate of the coefficient is based on how predictive the model is: the estimated coefficient is a *weighted* sum of coefficients estimated by the individual models. Thus, if the null model with no predictors is a very bad model, it is not allowed to influence the estimated coefficients much.

4 Results

First, plotting the probability of /n/-epenthesis by the 5 word-formation types, Figure 1 shows that /n/ was only inserted in words whose second morpheme was NF.

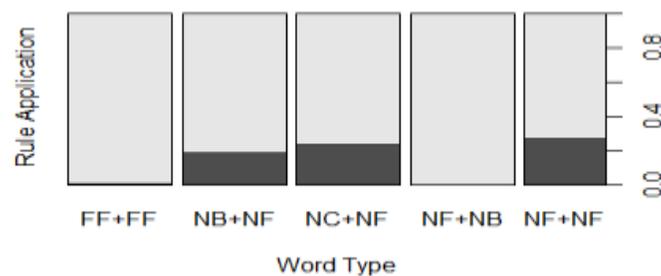


Figure 1. /n/-epenthesis by word-formation type: applied to words whose second morpheme is NF.

Second, if we limit ourselves to the subset of words in which the rule *can* apply, i.e., ones in which the second morpheme is native free, high-frequency words had higher probability of rule application. As shown in Figure 2, rule application was well correlated with judged frequencies of the whole words.

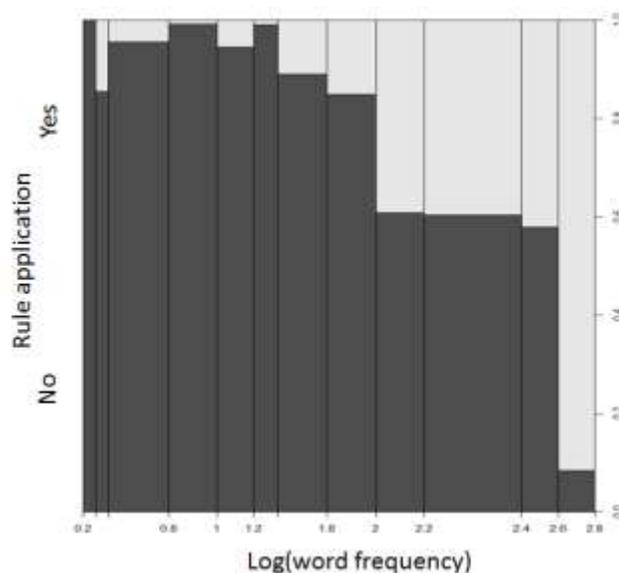


Figure 2. The effect of judged word frequency (log transformed) on rule application, where the second morpheme is NF (native free).

To confirm the reliability of the frequency effect and further investigate the differences among word types, while controlling for peculiarities of individual morphemes, three planned comparison analyses were conducted using mixed-effects models in which a subset of the word formation types was included. The first analysis investigated the influence of the type of the first morpheme by comparing the word types in which the second morpheme was native free (NB+NF, NF+NF, FF+NF). There were random intercepts for individual first and second morphemes, random intercepts for participants and random slopes for the fixed effects within participants. Table 3 shows that the effect of judged word frequency was significant within this subset of stimuli (confirming the trends shown in Figures 1 and 2): frequent words were more likely to have /n/ epenthesis than rare ones. Beyond these effects, there was a significant effect of whether the first morpheme was bound, free or nonce with bound morphemes favoring rule application. In addition, there was an interaction between Type and Repetition. Rule application probability increased from the first exposure to a word to the third for the nonce words but there was no effect of Repetition for existing words.

Table 3. Generalised linear mixed effects model results for the subset of words with native free second morphemes. Frequency was rank-transformed for this analysis to remove skew.

	<i>b</i>	<i>se(b)</i>	<i>z</i>	<i>p</i>	
(Intercept)	-4.75	0.83	-5.72	.00000	***
Repetition2	-0.16	0.24	-0.65	.51	
Repetition3	-0.27	0.26	-1.05	.29	
TypeNC+NF	-2.29	0.91	-2.52	.01	*
TypeNF+NF	-2.19	0.78	-2.82	.00	**
Frequency	0.0006	0.0001	4.18	.00003	***
Repetition2:TypeNC+NF	0.57	0.34	1.68	.09	.
Repetition3:TypeNC+NF	1.09	0.35	3.15	.002	**
Repetition2:TypeNF+NF	0.05	0.34	0.14	.89	
Repetition3:TypeNF+NF	-0.05	0.36	-0.13	.89	

Note. Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1

To examine possible differences between Sino-Korean and native Korean words, an additional mixed-effects regression analysis was conducted on the NF+NF and FF+FF stimuli. Again, we included random intercepts for the individual morphemes and participants as well as within-participant random slopes for the fixed effects, which were judged Frequency (rank transformed to reduce skew), Type (native vs. foreign), and Repetition (1 vs. 2 vs. 3). There was a significant effect of Frequency: frequent words were more likely to feature /n/ epenthesis than rare ones. There was also a significant effect of Type: Native Korean words favored /n/ epenthesis compared to Sino-Korean ones. There was no significant effect of Repetition on these real words. While Figure 1 suggests an interaction between Type and Frequency, the interaction could not be fit because of lack of variation within the Sino-Korean words, which almost never undergo /n/ epenthesis in our data.

Table 4. Generalised linear mixed effects model results for words consisting of native Korean (NF+NF) or Sino-Korean (FF+FF) free morphemes.

	<i>b</i>	<i>se(b)</i>	<i>z</i>	<i>p</i>	
(Intercept)	-16.22	2.97	5.47	.00000	***
Frequency	0.0009	0.0002	3.13	.002	**
TypeNF+NF	6.99	1.92	3.65	.0003	***
Repetition2	-0.11	0.26	0.40	.69	
Repetition3	-0.26	0.26	0.99	.32	

Note. Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1

Finally, within the set of words containing native second morphemes, we examined the influence of whether the second morpheme was bound or free. There were again random intercepts for individual morphemes and participants as well as random slopes for the fixed-effects predictors within participants. Table 5 shows that there was an influence of Type, such that native free second morphemes significantly favored epenthesis compared to native bound ones, which did not trigger epenthesis in our data. There was still an effect of judged Frequency (rank-transformed for the analysis), such that probability of epenthesis increased with increases in word frequency.

Table 5. Generalised linear mixed effects model results for words consisting of native Korean morphemes where the second morpheme is free (NF+NF) or bound (NF+NB).

	<i>b</i>	<i>se(b)</i>	<i>z</i>	<i>p</i>	
(Intercept)	-15.24	2.11	-7.23	.00000	***
Repetition2	-0.04	0.21	-0.21	.83	
Repetition3	-0.18	0.21	-0.85	.39	
TypeNF+NF	6.36	1.64	3.88	.0001	***
Frequency	0.001	0.0002	5.29	.00000	***

Note. Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '' 1

While Type was a significant predictor in the presence of variability in rule application associated with individual morphemes, it was found that both first and second morphemes played a role in predicting rule application. Models that contained random intercepts for individual second morphemes performed better than ones that did not include the random intercepts: ($\chi^2(1) = 1101.8, p < .0001$, for the full dataset). Models that had random intercepts for individual first morphemes also performed better than ones that did not ($\chi^2(1) = 245.4, p < .0001$, for the full dataset). This result motivated searching for particular morphemes that are strongly associated with presence or absence of /n/ epenthesis using conditional inference trees (Hothorn, Hornik, & Zeileis 2006).

Identity of the first morpheme, identity of the second morpheme and word frequency were entered as predictors of rule application into a conditional inference tree. Conditional inference trees are binary-branching. Thus, to identify the strongest predictor, each of the predictors entered into the model is transformed into a binary variable in every possible way. For categorical predictors like identity of the second morpheme, this has the effect of identifying clusters of predictor levels (here, second morphemes) that behave alike. For continuous predictors like word frequency, different splits into low and high values are attempted. The best binarised variable is then placed at the top of the tree. In our case, this was identity of the second morpheme of the compound, which was split into two groups that differed greatly in the probability of rule application. One of these groups consisted of the three morphemes that strongly favored /n/ epenthesis, *il*¹ 'work', *ip*² 'mouth', and *ip*³ 'leaf', which we refer to as 'special' morphemes.

Conditional inference trees are conditional because they consider entering additional predictors into the tree based on whether they improve predictiveness of the model within subsets of data defined by the values of the stronger predictors already in the tree. For example, in the present case, the predictive values of word frequency and the identity of the first morpheme would be assessed and compared separately for 1) words containing the 'special' second morphemes, and 2) words not containing the 'special' second morphemes.

Word frequency was selected as an important predictor while first morpheme identity was not, indicating that 1) the second morpheme and word frequency are better predictors of rule application, and 2) once the identity of the second morpheme and word frequency are known, knowing the identity of the individual first morphemes does not increase the predictive power of the model. Word frequency was predictive only for words whose second morpheme is one of the three special morphemes favoring /n/ epenthesis. More frequent words (whose log-frequency was higher than 2.573) mostly had an epenthesised /n/ while the relatively infrequent words (whose log-frequency was equal or lower than 2.573) received an /n/ half of the time. /n/

epenthesis largely failed to apply without a special morpheme to trigger it regardless of word frequency.

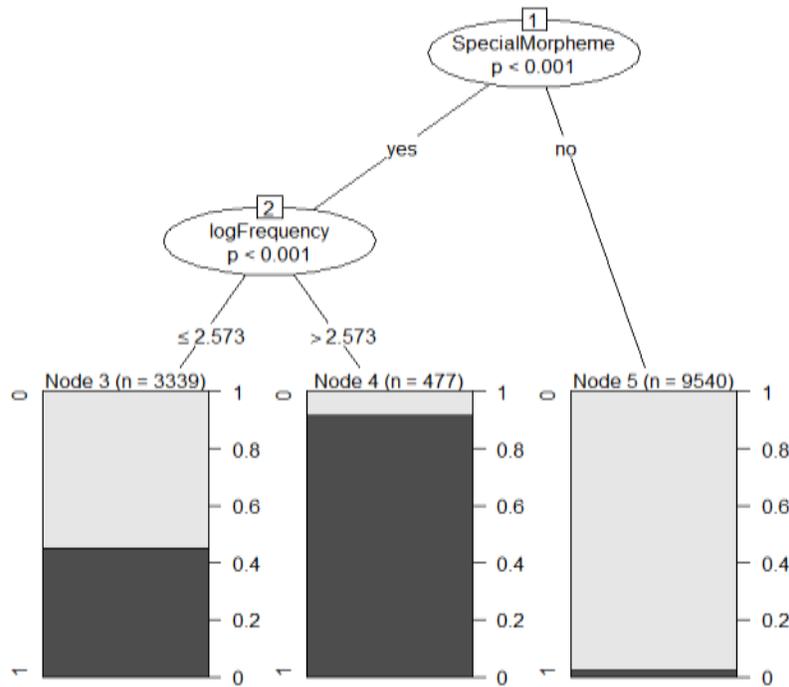


Figure 3. Conditional inference tree depicting the effects of Frequency and Special Morpheme (*il*¹ ‘work’, *ip*² ‘mouth’, and *ip*³ ‘leaf’)

Taking into account the possible existence of special morphemes, we used multimodel inference with generalised logistic mixed-effects models to determine whether the effects of Type and Frequency remain significant when Special Morpheme (yes or no) is entered into the model as a predictor. The Type and Frequency effects in Tables 3-5 remained, supporting their robustness across individual morphemes.

5 Discussion

We have documented that the Korean /n/ epenthesis rule has undergone morphologisation and has limited productivity. Morphologisation is indicated by the finding that /n/ epenthesis occurs almost exclusively preceding three ‘special’ morphemes, all other contexts instead favoring the application of the more general, and more type-frequent Korean Coda Rule.

As noted in Bybee (2001), morphologisation is usually accompanied by loss of productivity. Thus, even in the context of the special morphemes, /n/ epenthesis is not fully productive. In particular, /n/ epenthesis is more likely in high-frequency words than in low-frequency words. The direction of this frequency effect is consistent with usage-based views of analogical change (Bybee 2001, Phillips 1984, 2001) and dual-route models of the relationship between lexical retrieval and grammatical computation (Baayen *et al.* 1997, Berko 1958, Hay 2003, Kapatsinski 2010a). If /n/-epenthesis was triggered solely by the special morphemes, all words with a certain special morpheme would be equally likely to exhibit /n/ epenthesis. However, the data show decreasing probability of /n/-epenthesis in the whole words that share ‘leaf’ as a function of the whole-word frequency: the probability decreases from HF whole words

(e.g., *tanphungip*³ ‘maple leaf’, judged frequency = 14.29, application count = 155), to less frequent whole words (e.g., *solip*³ ‘pine-tree leaf’, judged frequency = 6.57, application count = 91), to nonce whole words (e.g., *matam-ip*³ ‘NONCE-leaf’, judged frequency = 2.1, application count = 54). This can be accounted for only if we assume whole-word storage and retrieval.

As we noted above, /n/ epenthesis has lost productivity and is being analogically replaced by the Korean Coda Rule. Analogical changes start with low-frequency words: once a rule loses productivity, it cannot derive new forms, yet seemingly rule-obeying forms can still be produced if they are retrieved from the lexicon. Low-frequency words are harder to retrieve from the lexicon than high-frequency words (Oldfield & Wingfield 1965 et seq) and are more likely to be produced piece-by-piece using the grammar (for production, see Berko 1958, Bybee 1995, 2001, Hay 2003, Kapatsinski 2010a). As a result, loss of productivity of a rule is immediately reflected in low-frequency words and non-words (the basic idea behind the wug test, Berko 1958). In contrast, high-frequency words containing an /n/ that was historically the product of /n/ epenthesis can simply be retrieved from the lexicon, as long as their morphologically complex surface forms are stored in memory. Thus a rule can lose productivity but still seem to apply in high-frequency words. For example, while the *i*→*ε* stem change is no longer productive in English past tense verbs, the high-frequency *keep/kept* is much less likely to regularise (**keeped*) than a low-frequency verb like *weep/wept* or *weeped* (Bybee 2001; see also Tagalog nasal substitution in Zuraw 2000, Russian velar palatalisation in Kapatsinski 2010b).

5.1 What makes special morphemes special: occurrence in epenthesis-favoring context

What is it about ‘leaf’ and the other special morphemes that makes them favor /n/ epenthesis? Importantly, the special morphemes appear to be defined morphosyntactically, rather than phonologically, indicating that the reason is likely not to be found in the phonology of the special morphemes. In earlier work, Kim *et al.* (2003) reported different probability of /n/-epenthesis in the same form *kheun+il* in two different senses, one in ‘a big matter/happening’ (38% of /n/-epenthesis) and the other in ‘an important household matter’ (61%). In the present study, we observed that *il*² ‘day’ and *ip*¹ ‘enter’ are not special morphemes whereas *il*¹ ‘work’, *ip*² ‘mouth’, and *ip*³ ‘leaf’ are. Analogous cases are discussed by Bybee & Scheibman (1999), where *going to* reduces when used to mean ‘future’ but not (as much) when used to denote locomotion in a direction; and by Gahl (2008), who shows that even content words that differ in frequency, like *thyme* and *time*, have slightly different pronunciations. The finding of different pronunciation patterns for apparent homophones suggests the existence of separate stored production representations for each word in a homophone pair (Gahl 2008).

We suggest that the special morphemes have become associated with /n/ epenthesis, or indeed fused with the preceding /n/ into a single schema, by virtue of frequently occurring in contexts favoring /n/ epenthesis (as suggested by Bybee 2002, Kapatsinski 2010b, Raymond & Brown 2012, Torres Cacoullos 1999). For example, one of the special morphemes is *ip*³ ‘leaf’. The top four words that underwent /n/-epenthesis most often all contain ‘leaf’ as their second component: ‘maple leaf’ (155/159), ‘cotyledon’ (153/159), ‘flower leaf’ (149/159), and ‘camellia leaf’ (141/159). Even nonce words containing *ip*³ are likely to undergo /n/ epenthesis, indicating that /n/ epenthesis is reliably triggered by *ip*³. For example, *ma.tal+ip*³ ‘NONCE+leaf’, judged unfamiliar (1.33/20), received a high rate of application of /n/-epenthesis (70/159). As we would expect, *ip*³ is usually used in combination with other morphemes as in *tanphung-ip*³ ‘maple leaf’ (raw frequency = 14.29), *kkot-ip*³ ‘flower leaf’ (13.67), *tongpaek-ip*³ ‘camellia leaf’ (7.48), and so on. ‘Leaf’ is special among the native free morphemes tested in that it often occurs as the second morpheme of a compound (in fact, it is more likely to occur in a compound than as a

separate word; 10 vs. 5 in the spoken subcorpus and 3317 vs. 2082 in the written subcorpus of the 21st Century Sejong Project, by the National Institute of the Korean Language^{ix}). Furthermore, within this context, *ip*³ ‘leaf’ is unique in almost always occurring in a phonological context for /n/ epenthesis (spoken: 10 vs. 0; written: 3064 vs. 253). By virtue of frequently occurring in contexts favoring /n/ epenthesis, *ip*³ could have become associated with /n/ epenthesis more than any other native free morpheme. By contrast, it can be predicted that bound second morphemes may not be associated with /n/ epenthesis. Such morphemes (e.g., *-i* ‘ADVERBIALISER’ and *-ira* ‘is’ in our experiment) are highly frequent grammatical units in Korean, and are therefore used across varying contexts, only some of which favor /n/ epenthesis (cf. Bybee 2002 for the English *-ed*).

Differences among the types of first morphemes are also consistent with the importance of frequently occurring in an epenthesis-favoring context. Bound first morphemes, which by definition always occur in the epenthesis-favoring word-internal context, favor /n/ epenthesis when context is controlled as in the present study. Free morphemes, which often occur word-finally, disfavor it. Most strikingly, when first introduced, nonce first morphemes are relatively unlikely to undergo /n/ epenthesis. With three exposures, all in an epenthesis-favoring context, the nonce morphemes come to undergo /n/ epenthesis significantly more often. We do not observe this effect of repetition for existing morphemes, since the proportion of occurrences in epenthesis-favoring contexts is much more stable for an existing morpheme and therefore its pronunciation is not as affected by three additional tokens (Goldinger 1998).

Because our stimuli containing special morphemes were all compounds with a consonant-final first morpheme, there is some ambiguity regarding *which* unit becomes associated with /n/ epenthesis as a result of frequently occurring in epenthesis-favoring contexts. There are at least three alternative possibilities, which differ in how context-specific the effect of the special morphemes on the probability of /n/ epenthesis is predicted to be. One possibility, which might be suggested by our wording so far, is that the output /n/ becomes associated with a morpheme or a word, forming the onset of that word or morpheme. This kind of process can be argued to occur, for example, in the formation of the word *nother* from the frequent co-occurrence of *an* and *other* in *another*. If this is the applicable process for the present data, special morphemes are expected to favor /n/ epenthesis even when they are word-initial.

Another possibility is that /n/ still becomes part of a unit but the relevant unit is a partially lexically-specific compounding construction, in which the second stem of the compound is fully specified while the first stem is left underspecified except for the final /n/. Adopting Booij (2010)’s representation, one of these constructions could be denoted as /...n+ip/ ~ [kind of[leaf]_N]_N. Under this account, the *ip* ‘leaf’ in a compound is not the same lexical item as the *ip* ‘leaf’ when it occurs on its own as an independent noun. The compound-internal *ip* is expected to favor /n/ epenthesis while the independent word *ip* may not. Such partially lexically-specific compounding constructions have been proposed by Booij (2010) to account for the fact that a morpheme may have a consistent meaning across a number of different compounds but yet this meaning might be different from the one that the ‘same’ morpheme has outside of the compounding context. For example, the Dutch *boer* means ‘seller’ when it occurs as head of a compound but ‘farmer’ when it is a noun on its own (Booij 2010). Kapatsinski & Vakareliyska (2013) and Trasca (2012) further document that these partially lexically-specific constructions can arise in a language, acquiring productivity, by generalisation over several borrowed compounds with the same head, without the independent noun necessarily being borrowed. The partially-lexically-specific construction for +*ip* ‘leaf’ could acquire the preceding /n/ because it frequently follows /n/. Kapatsinski (2013) exposed participants to a language in which the suffix

–i usually occurred after /tʃ/, either because the /tʃ/ resulted from velar palatalisation, as in /bek/ ‘wug.SG’ ~ /betʃi/ ‘wug.PL’, or because it occurred underlyingly, as in /lutʃ/ ‘looch.SG’ ~ /lutʃi/ ‘looch.PL’. The participants came to often attach –tʃi to novel inputs, e.g. generating /bluptʃi/ ‘blup.PL’ when presented with /blup/ ‘blup.SG’. If the same process occurs with the ‘special morphemes’ we found, they should favor /n/ epenthesis even when the preceding stem ends in a vowel, thus lacking a consonant that could be transformed into /n/ by the /n/ epenthesis rule, just as /tʃ/ is inserted after /blup/, which lacks a velar that could be transformed into /tʃ/ in Kapatsinski (2013).

Finally, it is possible that the special morphemes come to favor a rule that transforms consonants into /n/ (becoming part of the structural description of a variable rule, as in Labov 1969; see also Gouskova & Becker 2013, Pierrehumbert 2006 for recent examples of productive arbitrary alternations). In this case, the special morphemes should have an effect on /n/ epenthesis only when the preceding morpheme ends in a consonant, as otherwise there is no source for an /n/ in the input to the rule. We leave it to future research to decide amongst these alternatives.

5.2 Occurrence in epenthesis-favoring context as an alternative to strata

The application of /n/-epenthesis was found to be blocked in Sino-Korean words. This result would traditionally be accounted for with reference to strata (e.g. Chomsky & Halle 1968, McCawley 1968), where Sino-Korean words would be assumed to not feature /n/ epenthesis *because* they are Sino-Korean. However, whereas the Sino-Korean words in our experiment rarely applied /n/-epenthesis, other studies found cases in which Sino-Korean words did trigger /n/-epenthesis, often almost to the same degree as the native-source words. For instance, Kim *et al.* (2003) found *yeodjung* ‘middle school’ in *namwonyeodjung* ‘Namwon Middle School’ attracted /n/-epenthesis from 96.86% of the speakers, which is an extremely high probability of /n/-epenthesis. This inconsistency in the behavior of Sino-Korean words makes a stratal analysis problematic.

Instead, we could note that Sino-Korean morphemes are inherently free. Thus they are relatively unlikely to occur following another morpheme within the same word, which is the context where /n/ epenthesis applies. As a result, Sino-Korean morphemes would be expected to not become associated with /n/ epenthesis and therefore should fail to trigger it even when they do follow another morpheme within a word. Sino-Korean words that do trigger /n/-epenthesis are then expected to be the Sino-Korean words that do tend to occur inside words following other morphemes. Kim *et al.*’s (2003) example of *yeodjung* ‘middle school’ is of this kind, as it often occurs in names of middle schools following other morphemes. On this account, such Sino-Korean words should behave the same way as native Korean words.

However, we should also note that the Sino-Korean morpheme *il*² ‘day’ does not often trigger /n/ epenthesis, despite being homophonous with a special morpheme *il*¹ ‘work’ that does trigger /n/ epenthesis and often occurring in compounds that are eligible to undergo /n/ epenthesis. The Sino-Korean *il*² ‘day’ has 252 spoken occurrences in compounds vs. 95 outside of compounds; 17629 written occurrences in compounds vs. 1129 outside; of the compounds containing ‘day’, 68% of spoken compound tokens, and 48% of written ones eligible to undergo /n/ epenthesis. The native *il*¹ ‘work’ almost never occurs in compounds 0 occurrences in compounds vs. 932 outside in speech, 813 in compounds vs. 31788 outside in writing. Despite this difference, the native morpheme is more likely to trigger /n/ epenthesis. Thus, the effect of strata does not appear to be reducible to frequency of occurrence in epenthesis-favoring contexts. Overall, the results are consistent with Zuraw’s (2009) proposal that the grammar restricts where

alternations can and cannot occur but word frequency influences the likelihood of alternation within the contexts that the grammar has defined as being variable.

5.3 Explaining productivity loss

Why did /n/ epenthesis become morphologised and is losing the competition to the Korean Coda Rule, even in front of the special morphemes? We suggest that the Korean Coda Rule is favored by a number of factors. Perhaps, foremost among them is its higher scope and, as a result, higher type frequency (Albright & Hayes 2003, Bybee 1985, 1995, 2001, Richtsmeier 2011, Wedel 2003). The phonological contexts that are described as triggering /n/-epenthesis (a coda-final morpheme followed by /i/ or /j/ within a word), are also described as triggering competing rules: the Korean Coda Rule and the Palatalisation Rule. While the Palatalisation Rule can apply only to those that end in /t/ followed by an /i/-initial syllable, and /n/-epenthesis can apply only to a coda followed by an /i/- or /j/-initial morpheme, the Coda Rule can apply to any coda-vowel sequence.

High scope and high type frequency have long been hypothesised to be associated with productivity (e.g. Albright & Hayes 2003, Bybee 1985, 1995, 2001). More recently, a causal link between the two has been established in experimental studies with miniature artificial languages in which rule scope or type frequency were directly manipulated and productivity measured (Kapatsinski 2010b, 2013, Richtsmeier 2011).

In particular, Kapatsinski (2010b, 2013) introduced participants to languages in which a plural suffix *-i* triggered palatalisation of a preceding velar stop ($k \rightarrow tʃi$). All participants were exposed to the same examples of rule application, $k \rightarrow tʃi$. However, half of the participants learned languages in which *-i* also often attached to [p] and [t] while the other half learned languages in which [p] and [t]-final stems took the suffix *-a*. Kapatsinski (2010b, 2013) showed that when a rule-triggering suffix is often used outside of the context in which the rule can apply, participants do not learn to productively apply the rule in front of the suffix. In contrast, when the suffix is usually used in the context in which the rule can apply, the participants learn to apply the rule in front of the suffix.

Using Albright & Hayes' (2003) Minimal Generalisation Learner, Kapatsinski (2010b) modeled this result as resulting from competition between $k \rightarrow tʃi$ and a more general $0 \rightarrow i/C_$ rule that competes with $k \rightarrow tʃi$ for [k]-final inputs. The higher the type frequency of $0 \rightarrow i/C_$, the lower the productivity of $k \rightarrow tʃi$. In the current context, /n/ epenthesis is applicable to a subset of contexts for which Coda Rule can apply, just as $k \rightarrow tʃi$ was applicable to a subset of contexts in which $0 \rightarrow i/C_$ could apply in Kapatsinski (2010b, 2013). High type frequency of the Coda Rule should thus lower productivity of /n/ epenthesis just as high type frequency of $0 \rightarrow i/C_$ lowered productivity of $k \rightarrow tʃi$.

The Korean Coda Rule, a simple resyllabification process, may also be favored because its application does not require a stem change. Inductive biases against stem-changing rules, and particularly rules enacting major changes have recently been documented by Kerkhoff (2007), Krajewski et al. (2011), Stave *et al.* (2013), White (2014) and Zuraw (2000), though cf. Bybee & Newman (1995).

Morphologisation of /n/ epenthesis may have also been favored by the high token frequencies of the special morphemes. As Korean speakers encountered /n/-epenthesis largely in front of the three special morphemes, they could learn that /n/-epenthesis applies only to the 'gang' of words with these morphemes (Bybee 2001), thus leading to loss of productivity of the /n/-epenthesis pattern itself. Finally, once restricted to high-frequency words, /n/ epenthesis could also further lose productivity because high-frequency words are likely to be retrieved from

the lexicon directly in production and perception, thus failing to provide the speaker with practice applying /n/ epenthesis and providing no evidence for /n/ epenthesis to the listener (Hay 2001, 2003, Hay & Baayen 2001, Kapatsinski 2010a).

5.4 Comparison to Kim *et al.* (2003)

Kim *et al.* (2003) suspected that whole-word frequency may influence the probability of /n/ epenthesis but abandoned the idea concluding that the data were contradictory. They noted that we sometimes see unusually high probability of /n/-epenthesis for some low-frequency words. In particular, they mentioned two low-frequency whole words (based on the authors' intuition) that often undergo /n/-epenthesis: *nam.won-yeo.jung* 'Namwon Middle School' (/n/ epenthesised in 96.86% of the speakers) and *pul+yu.kwe.ha.ta* 'to be unpleasant' (89.71%). We would argue that, although the proper name 'Namwon Middle School' itself is of low frequency, the stem 'middle school' is a frequent word, which is furthermore usually used as part of a compound following the name of a middle school, which at least half the time would end in a consonant. It would therefore be expected to become associated with /n/ epenthesis on our account.

Then, when the speakers are to produce a whole (or say compound) word that they have never heard before, such as 'Namwon Middle School', they pronounce the proper noun 'Namwon' following the general pronunciation rules in Korean, as they do not have stored representation of such a word, and retrieve either the word 'middle school' -- or the construction [N['middle school']]_N (Booij 2010) -- from the lexicon. Because of its association with /n/ epenthesis, gained through often occurring in epenthesis-favoring contexts, 'middle school' is a 'special morpheme', which has the power to trigger /n/ epenthesis despite /n/ epenthesis being generally unproductive.

This process could also be used to explain exceptional *non*-epenthesis. For instance, Kim *et al.* (2003: 27) discussed the low probability (1.14%) of /n/-epenthesis in words like *twes.pak+i.ma* 'a forehead that looks like a upside-down gourd bowl' was due to the fact that the participants read the words syllable by syllable. However, our participants did not show syllable-by-syllable reading, yet the same result was obtained. We would therefore interpret the lack of epenthesis in this word as being due to 1) the fact that the rule is morphologised coupled with 2) the fact that the second morpheme *i.ma* 'forehead' does not tend to occur in epenthesis-favoring contexts and thus is not able to trigger epenthesis.

6 Conclusion

We have argued that the Korean /n/-epenthesis rule has become morphologised and is no longer fully productive, largely because of competing with a rule of broader scope and consequently higher type frequency, the Korean Coda Rule (higher type frequency being a major determinant of productivity, as documented by Kapatsinski 2010b, 2013, and Richtsmeier 2011). Three special second morphemes, *il*¹ 'work', *ip*² 'mouth', and *ip*³ 'leaf', were found to play a crucial role in /n/-epenthesis. When one of the three morphemes is the second morpheme of a compound, /n/ epenthesis may apply. Following Bybee (2002), Kapatsinski (2010b), Raymond & Brown (2012) and Torres Cacoullós (1999), we argued that the special morpheme 'leaf' is special because it tends to occur in compounds following consonants, the context favoring /n/ epenthesis.

We have argued that /n/ epenthesis might not occur before most Sino-Korean morphemes because these do not tend to occur in compounds. This is a tempting hypothesis because it would allow for a unified treatment of all Korean morphemes without recourse to positing mental

representations of lexical strata and thus special mechanisms for categorising words into strata (as in Becker & Gouskova 2013). The empirical advantage of this proposal is that it can capture the behavior of exceptional Sino-Korean morphemes that do trigger /n/ epenthesis: these are Sino-Korean morphemes that tend to often occur in compounds. However, we should note that, while all three ‘special’ morphemes are native, two of the three do not often occur in compounds, and one of the Sino-Korean morphemes in our study does often occur in compounds. Thus we have to conclude that the effects of strata cannot be reduced to frequency of occurrence in an epenthesis-favoring context. We suggest that both morpheme source and frequency of occurrence in contexts favoring epenthesis play a role in predicting /n/ epenthesis. When a word belongs to the Sino-Korean stratum, it cannot trigger /n/ epenthesis even if it often occurs in epenthesis-favoring contexts. In contrast, when the morpheme is native, frequency of occurrence in epenthesis-favoring contexts comes into play (see also Zuraw 2009).

Whether /n/ epenthesis applies before a special morpheme depends on the frequency of the whole compound ending in the special morpheme: frequent compounds favor /n/ epenthesis. As /n/ epenthesis lost productivity, low token frequency words have begun to undergo the more general Korean Coda Rule, where the coda of the preceding syllable is re-syllabified into the onset of the next syllable. We interpret the findings as support for the idea that frequent forms are stored and retrieved in production directly while rare words may be derived using the grammar (Bybee 1995, 2001, Hay 2001, 2003, Hay & Baayen 2001, Kapatsinski 2010a). Thus, analogical changes start in low-frequency words like the one examined here start in low-frequency words, for which lexical retrieval is difficult or unavailable (Bybee 2001, Phillips 1984, 2001). The proposed model is summarised in Figure 4.

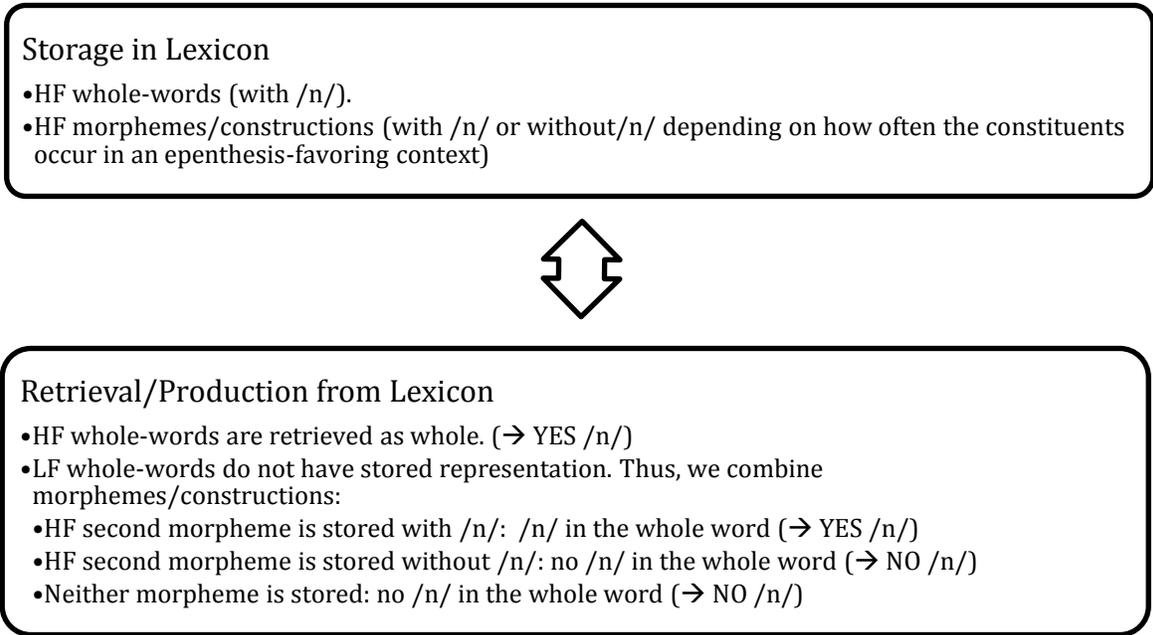


Figure 4. Proposed model

Note. HF: high frequency; LF: low frequency; YES /n/: /n/-epenthesis applies; NO /n/: /n/-epenthesis does not apply.

References

- Albright, Adams & Bruce Hayes (2003). Rules vs. analogy in English past tenses: a computational/experimental study. *Cognition* **90:2**. 119–161.
- Baayen, R. Harald, Ton Dijkstra & Robert Schreuder (1997). Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language* **37:1**. 94–117.
- Balota, David A., Maura Pilotti & Michael J. Cortese (2001). Subjective frequency estimates for 2,938 monosyllabic words. *Memory and Cognition* **29:4**. 639–647.
- Barr, Dale, Roger Levy, Christoph Scheepers, & Harry Tily (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* **68:2**. 255–278.
- Barth, Danielle & Vsevolod Kapatsinski (2014). A multimodel inference approach to categorical variant choice: construction, priming and frequency effects on the choice between full and contracted forms of *am*, *are* and *is*. *Corpus Linguistics and Linguistic Theory*. Published online ahead of print. DOI: 10.1515/cllt-2014-0022.
- Bartoń, Kamil (2013). MuMIn: Multi-model inference. R package (version 1.9.13). <http://CRAN.R-project.org/package=MuMIn>.
- Bates, Douglas, Martin Maechler & Ben Bolker (2013). lme4: Linear mixed-effects models using S4 classes. R package (version 0.999999-2). <http://CRAN.R-project.org/package=lme4>.
- Becker, Michael & Maria Gouskova (2012). Source-oriented generalizations as grammar inference in Russian vowel deletion. Ms, Indiana University and New York University.
- Berko, Jean (1958). The child's learning of English morphology. *Word* **14**. 150–177.
- Boersma, Paul, & David Weenink (2013). Praat: doing phonetics by computer [Computer program]. <http://praat.org>
- Booij, Geert (2010). *Construction Morphology*. Oxford: Oxford University Press.
- Burnham, Kenneth P. & David R. Anderson (2002). *Model selection and multimodel inference: a practical information-theoretic approach, 2nd Ed*. New York: Springer.
- Bybee, Joan (1985). *Morphology: A study of the relation between meaning and form*. Amsterdam: John Benjamins.
- Bybee, Joan (2001). *Phonology and language use*. Cambridge, New York, Port Melbourne, Madrid, & Cape Town: Cambridge University Press.
- Bybee, Joan (2002). Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change* **14:3**. 261–290.
- Bybee, Joan L. & Mary A. Brewer (1980). Explanation in morphophonemics: changes in Provençal and Spanish preterite forms. *Lingua* **52:3**. 201–242.
- Bybee, Joan L. & Jean E. Newman (1995). Are stem changes as natural as affixes?. *Linguistics* **33:4**. 633–654.
- Bybee, Joan & Joanne Scheibman (1999). The effect of usage on degrees of constituency: the reduction of *don't* in English. *Linguistics* **37:4**. 575–596.
- Choi, Taeyoung (2002). *Taeyoung Choi's research on Korean language*. Seoul: Worin.
- Chomsky, Noam & Morris Halle (1968). *The sound pattern of English*. New York: Harper & Row.
- Eom, Taesu (1999). *Research on Korean phonological rules*. Seoul: Kukhak Jaryowon.
- Eu, Jinseung (2008). Testing search engine frequencies: Patterns of inconsistency. *Corpus Linguistics and Linguistic Theory* **4:2**. 177–207.

- Gahl, Susanne (2008). *Time and thyme* are not homophones: The effect of lemma frequency on word durations in spontaneous speech. *Language* **84:3**. 474–496.
- Gernsbacher, Morton A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General* **113:2**. 256–281.
- Goldinger, Stephen D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review* **105:2**. 251–279.
- Gordon, Barry (1985). Subjective frequency and the lexical decision latency function: implications for mechanisms of lexical access. *Journal of Memory and Language* **24:6**. 631–645.
- Gouskova, Maria & Michael Becker. (2013). Nonce words show that Russian yer alternations are governed by the grammar. *Natural Language & Linguistic Theory* **31**. 735–765.
- Han, Kyung-im (2008). Geminate consonants in Korean. *British and American Language and Literature Association* **86:3**. 251–266.
- Hay, Jennifer (2001). Lexical frequency in morphology: is everything relative? *Linguistics* **39:6**. 1041–1070.
- Hay, Jennifer (2003). *Causes and consequences of word structure*. New York-London: Routledge.
- Hay, Jennifer & R. Harald Baayen (2001). Parsing and productivity. In Geert Booij and Jaap van Marle (eds.) *Yearbook of Morphology*. Dordrecht, Boston, London: Kluwer Academic Publishers. 203–235.
- Hothorn, Torsten, Kurt Hornik & Achim Zeileis (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics* **15:3**. 651–674.
- Idemaru, Kaori & Susan G. Guion (2008). Acoustic covariants of length contrast in Japanese stops. *Journal of the International Phonetic Association* **38:2**. 167–186.
- Idemaru, Kaori & Susan Guion-Anderson (2010). Relational timing in the production and perception of Japanese singleton and geminate stops. *Phonetica* **67:1**. 25–46.
- Kang, Okmi (2003). *Korean phonology*. Seoul: Thaeaksa.
- Kapatsinski, Vsevolod (2010a). What is it I am writing? Lexical frequency effects in spelling Russian prefixes: Uncertainty and competition in an apparently regular system. *Corpus Linguistics and Linguistic Theory* **6:2**. 157–215.
- Kapatsinski, Vsevolod (2010b). Velar palatalisation in Russian and artificial grammar: constraints on models of morphology. *Journal of Laboratory Phonology* **1:2**. 361–393.
- Kapatsinski, Vsevolod (2013). Conspiring to mean: Experimental and computational evidence for a usage-based harmonic approach to morphophonology. *Language* **89:1**. 110–148.
- Kapatsinski, Vsevolod & Joshua Radicke. (2009). Frequency and the emergence of prefabs: Evidence from monitoring. In Roberta Corrigan, Edith A. Moravcsik, Hamid Ouali and Kathleen Wheatley (eds.) *Formulaic language. Volume 2: Acquisition, loss, psychological reality, and functional explanations*, 499–520. Amsterdam: John Benjamins.
- Kapatsinski, Vsevolod & Cynthia Vakareliyska (2013). [N[N]] compounds in Russian: A growing family of constructions. *Constructions and Frames* **5:1**. 69–87.
- Kerkhoff, Annemarie O. (2007). *Acquisition of morpho-phonology: The Dutch voicing alternation*. Doctoral Dissertation, Utrecht: Utrecht University.
- Kim, Suncheol, Miyoung Kwon & Yunshin Hwang (2003). *Research on the practice of standard pronunciation II, 2003-1-8*. National Institute of Korean Language.

- Kim-Renaud, Young-key (1991). *Korean consonantal phonology*, 2nd ed. Seoul: Hanshin Publishing Company.
- Krajewski, Grzegorz, Anna L. Theakston, Elena V. M. Lieven & Michael Tomasello (2011). How Polish children switch from one case to another when using novel nouns: Challenges for models of inflectional morphology. *Language and Cognitive Processes* **26:4-6**. 830-861.
- Kuperman, Victor & Joan Bresnan (2012). The effects of construction probability on word durations during spontaneous incremental sentence production. *Journal of Memory and Language* **66:4**. 588–611.
- Labov, William (1969) Contraction, deletion, and inherent variability of the English copula. *Language* **45:4**. 715-762.
- Lee, Ho-young (1996). *Korean phonetics*. Paju: Thaeaksa.
- Lee, Jong-woon (2006). *Korean orthographic representation*, 2nd ed. Seoul: Sechang Media.
- McCawley, James D. (1968). *The phonological component of a grammar of Japanese*. The Hague: Mouton.
- Oh, Grace E. & Melissa A. Redford (2012). The production and phonetic representation of fake geminates in English. *Journal of Phonetics* **40:1**. 82–91.
- Oldfield, Richard C. & Arthur Wingfield (1965). Response latencies in naming objects. *Quarterly Journal of Experimental Psychology* **17:4**. 273–281.
- Pae, Joochae (1996). *Introduction to Korean phonology*. Seoul: Shingu Publishing.
- Phillips, Betty S. (1984). Word frequency and the actuation of sound change. *Language* **60:2**. 320-342.
- Phillips, Betty S. (2001). Lexical diffusion, lexical frequency, and lexical analysis. In Joan L. Bybee and Paul J. Hopper (eds.) *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins. 123–136.
- Pickett, Emily R., Sheila E. Blumstein & Martha W. Burton (1999). Effects of speaking rate on the singleton/geminate consonant contrast in Italian. *Phonetica* **56:2**. 135–157.
- Pierrehumbert, Janet (2006). The statistical basis of an unnatural alternation. In Louis Goldstein, Doug Whalen, and Catherine Best (eds.), *Laboratory phonology VIII*. Berlin: Mouton de Gruyter. 83-108.
- Raymond, William D. & Esther L. Brown (2012). Are effects of word frequency effects of context of use? An analysis of initial fricative reduction in Spanish. In Stefan. Th. Gries and Dagmar Divjak (eds.) *Frequency effects in language learning and processing*. Berlin: Mouton de Gruyter. 35–52.
- Richtsmeier, Peter T. (2011). Word-types, not word-tokens, facilitate extraction of phonotactic sequences by adults. *Journal of Laboratory Phonology* **2:1**. 157–183.
- Shin, Ji-young & Jae-eun Cha (2003). *The system of Korean Sounds: for the basis of research on Korean phonology*. Seoul: Hankuk Munhwasa.
- Sohn, Ho-Min (2001). *The Korean Language*. Cambridge University Press.
- Stave, Matthew, Amy Smolek & Vsevolod Kapatsinski (2013). Inductive bias against stem changes as perseveration: Experimental evidence for an articulatory approach to output-output faithfulness. In *Proceedings of the Annual Meeting of the Cognitive Science Society* **35**. 3454–3459.
- Sugiura, Nariaki (1978). Further analysts of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics-Theory and Methods* **7:1**. 13–26.
- Taft, Marcus (2004). Morphological decomposition and the reverse base frequency effect. *The Quarterly Journal of Experimental Psychology* **57A:4**. 745–765.

- Torres Cacoullos, Rena (1999). Variation and grammaticization in progressives Spanish-ndo constructions. *Studies in Language* **23:1**. 25-59.
- Trasca, Luminita (2012) Noun+noun constructions in written Romanian: What do *escrocii internet, tribunalul Timis* and *presedintele SIF* have in common? *Linguaculture* **3:2**. Iasi: A.I.Cuza University Press.
- Tremblay, Antoine & R. Harald Baayen (2010). Holistic processing of regular four-word sequences: A behavioral and ERP study of the effects of structure, frequency, and probability on immediate free recall. In David Wood (ed.) *Perspectives on formulaic language: Acquisition and communication*. London: Continuum. 151–173.
- Wedel, Andrew B. (2003). Self-organization and categorical behavior in phonology. *Berkeley Linguistics Society Meeting* **29**. 611–622.
- White, James (2014). Evidence for a learning bias against saltatory phonological alternations. *Cognition* **130:1**. 96–115.
- Zuraw, Kie R. (2000). *Patterned exceptions in phonology*. Doctoral dissertation, Los Angeles, CA: University of California, Los Angeles.
- Zuraw, Kie R. (2009). Frequency influences on rule application within and across words. *Chicago Linguistic Society* **43**. 283–309.

ⁱ Available also on the website of *The National Institute of the Korean Language* (pp. 112-113), http://www.korean.go.kr/09_new/data/etc_view.jsp?idx=82, and also appended in Lee (2006:511-512).

ⁱⁱ Sais-Sori Phenomenon requires that a <^> orthographic symbol (pronounced as /s/) should be orthographically inserted to the end of the first morpheme when the first morpheme ends in a vowel. For example, *is+mom* is the orthographic representation for the word-formation *i+mom* ‘tooth+body, gum,’ from p. 74 of the 7th Curriculum High School Grammar.

ⁱⁱⁱ Following generations’ textbooks started to introduce the name /n/-epenthesis with the aforementioned linguistic context (i.e., C _ /i, j/) (e.g., the 7th Revised Curriculum, from 2011: *Mirae-en, Pisang Kyoyuk, Jihak-sa*).

^{iv} The available sources did not give information as to the proportion of the population was taught this subject.

^v It was assumed that the nonce morpheme would be interpreted as a native nonce. Even if the assumption is not true, we believe that it would not affect the results of the study in the identity of the first morpheme did not exert a major influence on /n/-epenthesis (Figure 3).

^{vi} Lee (1996:167-169) noted that the /i/ context would usually be associated with lexical, individual, and boundary variations whereas the /j/ environment would trigger /n/-epenthesis (optionally for onomatopoeia and mimetic words) unless the words were separated by a pause.

^{vii} Initially, these lexical items were constructed based on a national report on word frequency in spoken Korean (*hyeondae kukeo sayong pindo josa kyeolkwa* 2003-1.20 ‘the results of a research on the frequency of modern Korean language use 2003-1.20’, released by the National Institute of Korean Language: http://www.sejonghakdang.org/user.do?jsessionid=hzm0WpraMqmVnrlCmnlBsut1FWRCZrXRObHIBHtzITPnJQGyH3aPzna8cnciqZJF.was_servlet_engine1?mode=view&page=P40020100&dc=&boardId=1002&cp=1&boardSeq=897). However, the (small) size of the corpus did not allow us to create a complete word satisfying the phonological, morphological, and frequency criteria of the current study. Moreover, although we checked the whole-word frequencies by the hits in Google search, it did not seem to reflect the spoken frequency appropriately as some advertisements and written expressions appeared to distort the targeted spoken frequency measures (see also Eu 2008). Thus, in order to better reflect the spoken whole-word frequencies, we finally decided to conduct a frequency survey with another homogeneous subset of the population.

^{viii} For the frequencies of these repeated words we used the average of the two ratings.

^{ix} Available from <http://ithub.korean.go.kr/user/corpus/corpusSearchManager.do>.