Behav Res (2017) 49:1939–1950
DOI 10.3758/s13428-016-0836-8

K-SPAN: A lexical database of Korean surface phonetic forms and phonological neighborhood density statistics

Jeffrey J. Holliday[1] · **Rory Turnbull**[2] · **Julien Eychenne**[3]

Published online: 2 February 2017

Abstract This article presents K-SPAN (Korean Surface Phonetics and Neighborhoods), a database of surface phonetic forms and several measures of phonological neighborhood density for 63,836 Korean words. Currently publicly available Korean corpora are limited by the fact that they only provide orthographic representations in Hangeul, which is problematic since phonetic forms in Korean cannot be reliably predicted from orthographic forms. We describe the method used to derive the surface phonetic forms from a publicly available orthographic corpus of Korean, and report on several statistics calculated using this database; namely, segment unigram frequencies, which are compared to previously reported results, along with segment-based and syllable-based neighborhood density statistics for three types of representation: an "orthographic" form, which is a quasi-phonological representation, a "conservative" form, which maintains all known contrasts, and a "modern" form, which represents the pronunciation of contemporary Seoul Korean. These representations are rendered in an ASCII-encoded scheme, which allows users to query the corpus without having to read Korean orthography, and permits the calculation of a wide range of phonological measures.

Keywords Korean · Phonological neighborhood density · Lexicon · Lexical database

Electronic supplementary material The online version of this article (doi:10.3758/s13428-016-0836-8) contains supplementary material, which is available to authorized users.

✉ Julien Eychenne
jeychenne@hufs.ac.kr

Jeffrey J. Holliday
holliday@korea.ac.kr

Rory Turnbull
rory.turnbull@ens.fr

[1] Department of Korean Language and Literature, Korea University, 145 Anam-ro Seongbuk-gu, Seoul 02841, South Korea

[2] Laboratoire de Sciences Cognitives et Psycholinguistique (ENS, EHESS, CNRS), Département d'Etudes Cognitives, Ecole Normale Supérieure, PSL Research University, 29, rue d'Ulm, 75005 Paris, France

[3] Department of Linguistics and Cognitive Science, Hankuk University of Foreign Studies, Mohyeon, Yongin, Gyeonggi 17035, South Korea

Introduction

This article presents K-SPAN (Korean Surface Phonetics and Neighborhoods), the first lexical database of Korean to include transcriptions of surface phonetic forms and neighborhood density statistics. The database includes 63,836 entries, drawn from the Modern Korean Usage Frequency Survey 2 corpus (Kim 2005; Korean title: "현대 국어 사용 빈도 조사 2").

Developing experiments with carefully controlled stimuli is a common activity of those who are interested in spoken language processing, such as speech scientists, experimental psychologists, and linguists. Especially for tasks involving speech production or perception it can be important to be able to control the phonetic or phonological content of stimulus items. For that reason, phonetized databases—which list phonetic transcriptions of words—are an invaluable resource. However, Korean, like other understudied languages, does not have such a database. The development of K-SPAN was motivated by a desire to remedy this gap.

A related motivation was to calculate the phonological neighborhood density (ND) of Korean words. Phonological ND is commonly used as a measure of word similarity in studies of the phonological structure of the lexicon. It is typically operationalized as the number of other words in the lexicon that differ from a target word by a phonological edit distance of one: that is, by the addition, deletion, or substitution of a single phoneme. ND has been shown to be relevant in explaining performance on a host of linguistic tasks.

In the realm of speech perception, it has been found that high ND is correlated with slower lexical access. For example, Luce and Pisoni (1998) showed that listeners exhibited longer reaction times to high ND words in lexical decision, identification in noise, and naming tasks. In speech production, many studies have shown evidence for hyperarticulation in high ND words (Munson & Solomon, 2004; Scarborough, 2004; Wright, 2004), which has been interpreted in support of a listener-oriented view of phonetic reduction. Other studies have shown how ND may be useful in characterizing word learning: children's lexicons tend to contain more high ND words at first, and over time expand to low ND words (Coady & Aslin, 2003; Stokes, 2010). In short, the concept of ND is applicable to a wide range of questions in language processing and acquisition.

However, the great majority of this literature is based on English. Although ND is not inherently a language-specific concept (i.e., all languages have words, which are composed of phonemes), the validity of ND as a meaningful psycholinguistic measure is understudied in non-English languages. (Holliday & Turnbull 2015[1] and Vitevitch & Stamer 2006 are notable exceptions.) Accordingly, one of the aims of the current paper is to extend research on ND effects to Korean.

Previous research on neighborhood density in Korean

Although there exists some previous work on the effects of ND in Korean, it has been limited in several ways. First, these studies have focused mostly on visual word recognition (Kwon, Lee, Lee, & Nam, 2011; Kwon & Nam, 2011; Kwon, 2014). While such work is valuable in its own right, it addresses a different set of questions and does not lend itself to cross-linguistic comparison with the studies on spoken language processing cited above. Second, all of these studies used a measure of ND that is based on the *eumjeol*, a unit of Korean orthography that usually, but not always, corresponds to a phonological syllable. This measure is not analogous to the phoneme-based measure used in most of the spoken language processing literature.

[1]Holliday and Turnbull (2015) calculated ND using a grapheme-to-phoneme algorithm that could not handle the exceptions discussed in Section 3, and was a precursor to the current study.

Although syllable-based ND measures have proven meaningful in studies of visual word recognition (e.g., Carreiras, Alvarez, & de Vega, 1993; Perea, & Carreiras, 1998), they may be unsuited to the study of spoken Korean, owing to the many-to-many mapping between Korean orthography and pronunciation. The same spelling can be used for different pronunciations (e.g., 발병 [palpjʌŋ] 'occurrence of a disease' vs. 발병 [palp*jʌŋ] 'foot disease'), and different spellings can be used for the same pronunciation (e.g., 무난 [munan] 'easy' vs. 문안 [munan] 'regards'). This fact is discussed in more detail in Section 3. To our knowledge, the only published study of ND effects involving spoken Korean, Song, Nam, & Koo, (2012), investigated the effects of word frequency and ND on spoken word segmentation. This study used a syllable-based ND measure, as was done in other studies of spoken word segmentation (e.g., Cutler, Mehler, Norris, & Segui, 1986 and Mehler, Dommergues, Frauenfelder, & Segui, 1981), but it renders it nevertheless incompatible with the majority of ND studies on English, which used a phoneme-based measure.

Perhaps the most substantial gap in this body of research on Korean lies in the representations upon which ND was calculated, irrespective of being syllable- or phoneme-based. The aforementioned studies calculated ND from the orthographic forms of words, taking advantage of the fact that the Korean orthography, Hangeul, is relatively shallow: in many cases, the phonological form of a Korean word can be reliably derived from its orthographic form. Nevertheless, there are many exceptions, including the examples cited above, whose phonological form *cannot* be derived from its orthographic form. Sometimes this is the result of phonological processes that apply selectively based on morphological factors (e.g., whether or not the word is a compound noun), or only to words in certain lexical strata (e.g., Sino-Korean words). Lastly, there are some words whose phonological form is irregular, and is simply unpredictable from the orthographic form. Thus, given only the orthographic forms in a lexicon, the calculation of any kind of ND (other than orthographic syllable-based) becomes intractable.

Currently available Korean corpora

There exist several large corpora of Korean, but, to our knowledge, none of them provide both orthographic and phonological representations of words. The Sejong Corpus (Kim, 2006) is the largest publicly available Korean corpus. This corpus contains a large amount of annotated textual data written in Hangeul (part of the corpus also contains Chinese characters). It contains two core subparts: a "spoken" corpus, which includes orthographic transcriptions of conversations and interviews, and a "written" corpus, which includes material such as press articles, textbooks, novels

and poems from the 20th century. The written part of the corpus contains nearly 34 million tokens, corresponding to about 2.7 million types. The spoken part is much smaller and contains about 800,000 tokens, corresponding to nearly 115,000 types. Part of the Sejong Corpus is annotated for part of speech and includes lemmatic information, but no disambiguation is provided for lemmas that are homographs. This is an important issue since, as mentioned earlier, homographs in Korean are not necessarily homophones. As a result, it is not possible to reliably derive surface phonetic forms from the Sejong Corpus.

The KAIST Corpus[2] contains 70 million words and is available to the public, but the downloadable .zip file consists of raw, unparsed Korean text contained in 11,629 separate .txt files. An annotated subset of 1 million words is also available, but still contains only orthographic forms, like the larger version, and is not lemmatized. While this is certainly a valuable resource, it carries the same limitations as the Sejong Corpus with respect to the calculation of ND.

Shin, Kiaer, & Cha, (2013) reported phoneme frequency statistics from two corpora: the *Yonsei Korean Language Dictionary*,[3] and the Spoken Language Information Lab Corpus (Shin, 2008), a corpus of spoken dialogue recorded from 57 native speakers of Seoul Korean. Either of these corpora could potentially be used for the calculation of ND statistics, but neither the phonetic forms of the *Yonsei Korean Language Dictionary* nor the Spoken Language Information Lab Corpus (in any form) are available to the public.

The current study

This paper presents a lexical database of surface phonetic forms[4] and ND measures for Korean words, derived from a publicly available orthographic corpus. The corpus used as the basis for the current work was the Modern Korean Usage Frequency Survey 2 (Kim 2005; Korean title: "현대 국어 사용 빈도 조사 2"; henceforth MKUFS2). The MKUFS2 is a balanced lemmatized corpus containing 3,086,031 word tokens and 82,501 word types. Although it only provides orthographic forms, what sets the MKUFS2 apart from the corpora described above is its accessibility: it can be downloaded from the website of the National Institute of Korean Language (NIKL) as a single table file containing the orthographic form and lexical frequency of each word. Our work thus consisted of two main parts: phonetization of the orthographic forms, and calculation of ND based on the phonetized forms.

This work proceeded according to the following steps, to be described in detail in the next section. First, we retrieved the pronunciation of each word in the MKUFS2 that had an entry in the online Naver Korean dictionary. Second, because the pronunciation entries in the Naver dictionary are taken from the prescriptive forms in the Korean dictionary published by NIKL, we implemented several phonetic neutralizations that more accurately reflect the modern pronunciation of younger Seoul speakers. Lastly, we calculated several ND measures based on the modern pronunciations, the conservative pronunciations offered by NIKL, and the orthographic forms. These statistics are discussed in Section 3, along with a comparison between the segment-based frequencies measured in our database and those reported in previous studies.

This endeavor resulted in the creation of a database with the following information for each word: phonetic transcriptions of the modern and conservative pronunciations rendered in WorldBet (Hieronymus, 1994) and in another easy-to-process encoding scheme, and both segment- and syllable-based ND measures (to be described below). Each word is further identified by the row number of its entry in the MKUFS2, which the user can then refer to alongside the current database.

Method

Background on Korean orthography–phonology mapping

Contemporary Korean is written using an alpha-syllabic system (*Hangeul*), invented in the 15th century, which was specifically designed to transcribe this language. Modern Hangeul comprises 40 letters (*jamo*), which are organized into graphical syllable-sized blocks (*eumjeol*). For example, the word 사진 'photo' /satɕin/ is composed of two eumjeol. The first eumjeol, 사, is composed of the jamo ㅅ /s/ and ㅏ /a/. The second eumjeol, 진, is composed of the jamo ㅈ /tɕ/, ㅣ /i/, and ㄴ /n/.

The syllabic nature of this alphabet allows Hangeul to encode differences in syllabification between words. For example, the sequence /tali/ can be written as 다리 /ta.li/ 'leg' or as 달이 /tal.i/ 'moon+NOM'. Note that the second eumjeol in the latter example features the jamo ㅇ, which represents an empty (null) syllable onset. Due to

[2] http://semanticweb.kaist.ac.kr/home/index.php/Corpus1.

[3] https://ilis.yonsei.ac.kr/dic/.

[4] Throughout this paper, the "surface phonetic form", transcribed in square brackets, is intended to represent the output of phonological processes, such as those that result in neutralization. It does not, however, represent allophonic variations that are completely predictable. For example, lax stops become voiced between two other voiced segments (e.g., 바다 /pata/ "sea" → [pada]), and the sibilant fricative /s/ is palatalized when followed by /i/ or /j/ (e.g., 신 /sin/ "god" → [ɕin]). These allophonic variations are not phonologically relevant and do not enter into considerations of ND.

the phonological process of resyllabification, both of these words are pronounced identically as [ta.li]. Crucially, however, the morphological distinction between them is preserved in the spelling.

Other examples of many-to-one mappings between Hangeul and pronunciation relate to phonological mergers and neutralizations. First, a number of phonemes that were formerly distinct, such as the vowels /e/ and /ɛ/, have merged in Modern Korean and are now pronounced identically by most speakers (see Eychenne & Jang, 2015; Hong, 1988, 36–89; Shin et al., 2013, 99–101 among others). Second, Korean possesses a rich set of phonological processes that neutralize some phonological contrasts in certain environments (Ahn, 1998; Shin et al., 2013, ch. 8). For instance, the contrast between the three bilabial plosives /p/ (lenis), /p^h/ (aspirated) and /p*/ (fortis) is lost in coda position, where these phonemes are all realized as an unreleased bilabial stop [$p^{\urcorner}$]. Such phenomena are generally not problematic for a phonetization system since they are fully predictable.

There are, however, a number of processes that are sensitive to morphological information, involve a large amount of lexical idiosyncrasy, and are not reflected in standard Hangeul spelling. To take but one example, in some compound words in which the second morpheme starts with /i/ or /j/, an /n/ is inserted between the two morphemes. For example, 담요 /tam#jo/ 'blanket' is a compound of the Sino-Korean morpheme 담 /tam/ 'blanket' and the native Korean word 요 /jo/ 'Korean-style mattress'. This word, which has the morphophonological structure /tam#jo/, undergoes [n]-insertion and is pronounced [tamnjo]. Note that this inserted [n] is not reflected in the spelling. However, not all words containing an /i/- or /j/-initial morpheme trigger this process. Thus, the word 금요 /kɨm#jo/ 'Friday', which contains the morphemes 금 /kɨm/ 'gold' and 요 /jo/ 'shining', is transparently realized as [kɨmjo]. We will not delve into the complicated issues surrounding the range of morphology-sensitive processes in Korean in this paper (but see Shin et al., 2013, ch. 9, for an overview of the most important ones); for our purposes, it suffices to say that these processes make it extremely difficult to derive completely reliable phonetic transcriptions from orthographic forms alone.

Phonetization of MKUFS2

To deal with these unpredictable grapheme-to-phoneme correspondences, we opted for a more direct phonetization strategy by relying on existing publicly available resources (see Appendix for details, including web links). The MKUFS2 corpus is freely available for research purposes and can be downloaded from NIKL's website. This corpus provides, among other things, a dictionary of grammatical morphemes and lexical items. For the purpose of this work, we only considered the dictionary of lexical items, which contains 82,501 lemmas, along with each word's token frequency, part of speech, and an optional disambiguation column. In the case of homonyms, the disambiguation column clarified which lemma the entry referred to, and in the case of Sino-Korean or other loanwords, it contained the Chinese characters (*hanja*) or source language form.

In order to obtain the surface phonetic forms for each word, we used the free online Naver dictionary.[5] For most words whose pronunciation differs from the spelling (predictably or not), Naver provides a pseudo-phonetic representation in Hangeul. For instance, the verb form 농익다 /noŋ#ikta/ 'to be ripe' is phonetized as [농닉따], which transparently corresponds to the actual phonetic realization [noŋnikt*a]. This pseudo-phonetic representation shows the application of the non-predictable /n/ insertion rule discussed above, and also the predictable rule of post-obstruent tensing, which turns the underlying /t/ into tense [t*] because of the preceding obstruent. Homonyms were generally (but not systematically) identified with a numeric code that matched forms across the two corpora. For example, no phonetization is provided for 가마$_4$ 'volume unit', indicating that it is transparently phonetized as [가마] (= [kama]), whereas 가마$_5$ 'sedan chair' is phonetized as [가:마] (= [ka:ma]), with a long vowel in the first syllable.

Thus, the first step of the phonetization procedure was to obtain the pseudo-phonetic transcription in Hangeul, as provided by Naver, for each word in the MKUFS2 corpus. The text file containing the lexical items, which is provided in a legacy encoding (Windows code page 949), was first converted to Unicode (UTF-8).[6] For each word form, we retrieved the first result page(s), up to five. For unambiguous words, we extracted the only entry that was returned; for homonyms, we relied on a combination of the word's numeric code and hanja disambiguation (where available) to attempt to identify the target entry, giving precedence to the hanja disambiguation in case of conflict. For each entry, we extracted the pseudo-phonetic form when one was provided; otherwise, we used the orthographic form as a pseudo-phonetization since, in that case, the pronunciation was totally transparent. Words that could not be identified were discarded. Failure to identify a word in Naver could have

[5] Naver's dictionary is itself based on an online dictionary made available by NIKL, but it provides a more convenient interface since the entry for a given word can be accessed via a URL that contains the target word (along with other options). This enabled us to fetch and retrieve the relevant page(s) for an entry using the Common Gateway Interface protocol. In addition, unlike NIKL's dictionary, Naver renders search results as plain HTML that is easy to parse.

[6] We identified 343 forms whose disambiguation contained invalid hanja characters; 314 of these words appear in the final database.

two causes. First, some words from the MKUFS2 corpus were simply not listed at all in Naver. Many of the unknown words were complex verbs composed of a base verb + helping verb (such as 하다 'to do', 주다 'to give' or 되다 'to become'), such as 간략화되다 'to become simplified'. Second, some words were redirected to a similar, but different entry. For example, 가가소소하다 /kakaso:sohata/, a rare word with only one occurrence in the MKUFS2 corpus, was redirected to 소소하다 /so:sohata/; although Naver does provide several entries for the latter form, the first of which is phonetized as [소:소하다], this cannot be used to automatically and reliably derive the phonetization of 가가소소하다. Therefore, search results such as this one were excluded. In total, out of the 82,501 lemmas found in the MKUFS2 corpus, 18,665 forms were discarded; we obtained 63,836 phonetic forms, representing 77.4 % of the original corpus.[7]

For a large number of words (5018 items, 7.9 % of the database), Naver provided two different pseudo-phonetic forms, representing two pronunciation variants, with or without the application of a number of optional (though widespread) processes, such as the reduction of /je/ to /e/ after a velar stop (/sikje/ 'watch' → [sike]), or the neutralization of ㅚ /ø/ to ㅞ /we/ (see Table 1). Although each process, taken in isolation, was systematically applied to either the first or second pronunciation variant, the phonetization was not entirely consistent regarding what type of pronunciation each variant was supposed to represent. For example, many processes that characterize a typical modern pronunciation in Seoul Korean were applied to the second variant, but some (such as the insertion of the glide /j/ between /i/ and /ʌ/) were applied to the first variant. In addition, a number of features found in Modern Korean (e.g., loss of the length contrast, merger between ㅔ /e/ and ㅐ /ɛ/) were not indicated at all.

In order to alleviate these problems and to make the database maximally useful, we created two pronunciation variants, labeled as "conservative" and "modern". The conservative variant represents a somewhat archaic, if not artificial, pronunciation where all potential contrasts have been preserved. For example, the vowels ㅟ and ㅚ are transcribed as the monophthongs /y/ and /ø/, respectively, which corresponds to the normative pronunciation known as the "Standard Korean Pronunciation" (Shin et al., 2013, 97-99). The modern variant, on the other hand, represents a pronunciation typical of contemporary Seoul Korean. In order to obtain surface phonetic forms for the modern and conservative pronunciations, we first linearized each Hangeul pseudo-phonetic form using a standard code point decomposition algorithm (The Unicode Consortium, 2015, §3.12) which decomposes each eumjeol into its constituent jamo. As an example, the string "가공" was linearized into "ㄱㅏㄱㅗㅇ" (=/kakoŋ/). For all the forms in the database that were phonetized with two variants, we aligned the two strings using the Minimum Edit Distance algorithm, as implemented in Cock et al. (2009). We then built a conservative and modern pronunciation by assigning each mismatched character in Naver's pseudo-phonetic forms to the appropriate variant. The conservative forms, as mentioned above, retain all of the contrasts.

After the conservative and modern pronunciations were generated, we checked for potential errors (that is, cases when the phonetization provided by Naver was obviously incorrect) by searching for illegal phoneme strings. For example, underlying word-final /s/, which is common in /t/-final loanwords, is neutralized to an unreleased /t/ on the surface (e.g., 로봇 /lopos/ "robot" is phonetically realized as [lopot]). Some of Naver's phonetizations, however, contained errors such as this (e.g., "robot" being phonetized as 로봇 [lopos] instead of 로봗 [lopot]), and so in order to correct them we ran another script that checked for any anomalous phonetizations and applied an appropriate patch. We corrected 98 errors using this procedure. Because this method could not catch any errors that did not result in an illegal phoneme string, we also hand-checked a random subset of 1000 words to gauge whether there may be more errors, but did not find any.

In addition to outright errors, however, since even the modern pronunciations provided by the Naver dictionary contained some contrasts that are not maintained by most speakers of modern Seoul Korean, we applied several phonetic neutralizations to reflect the pronunciations that modern Seoul Korean speakers actually produce. The first neutralization comprised, broadly, the widely reported loss of contrast between ㅔ /e/ and ㅐ /ɛ/, as noted earlier. This neutralization included the realization of both ㅔ /e/ and ㅐ /ɛ/ as /ɛ/, both ㅞ /we/ and ㅙ /wɛ/ as /wɛ/, and both ㅖ /je/ and ㅒ /jɛ/ as /jɛ/. Note that because the realization of ㅚ /ø/ as /wɛ/ was already reflected in the Naver modern pronunciations, this step effectively neutralized the conservative three-way contrast among ㅚ /ø/, ㅞ /we/, and ㅙ /wɛ/. Finally, we neutralized vowel length distinctions since this feature appears to play a marginal role, if any at all, in the phonology of contemporary Seoul Korean (Lee, & Ramsey, 2011, 296-297; Shin et al., 2013, 153; Sohn, 1999, 14). In summary, the list of processes which was applied to the modern forms is given in Table 1.

A few representative examples, drawn from the final database, are provided in Table 2. These examples demonstrate the orthographic representation in Hangeul, the conservative pronunciation provided by Naver, and the modern

[7]The Python script provided with the database automatically extracts the list of discarded words.

Table 1 Processes applied to obtain the modern pronunciation variants

Process	Illustration		
length neutralisation	ka:kʌnmul	→	kakʌnmul
place assimilation	kakɛtp*aŋ	→	kakɛpp*aŋ
post-velar glide deletion	kakje	→	kake
ɯi ∼ i neutralization	katɕoktɕ*ɯɯi	→	katɕoktɕ*ui
/y/ → wi	sypt*a	→	swipt*a
/ø ∼ wɛ/ neutralization	kjʌŋø	→	kjʌŋwɛ
/e ∼ ɛ/ neutralization	metɕu	→	mɛtɕu
/j/ insertion	miʌtɕita	→	mijʌtɕita

Note that a given form may undergo several processes. Therefore, the form resulting from the application of the process in the illustrations does not necessarily reflect the final modern form

[a]Note that by extension this process also resulted in the neutralization of /je ∼ jɛ/ and /we ∼ wɛ/

pronunciation provided by Naver and subsequently updated based on the mergers described above. In the final database, both the conservative and modern pronunciations were rendered in Worldbet (Hieronymus, 1994), an ASCII encoding scheme for the International Phonetic Alphabet. In addition, in order to facilitate the calculation of (possibly novel) lexical metrics, we also rendered these pronunciations using a simple encoding scheme which maps each segment (vowel, consonant or diphthong) to a single ASCII character. (This scheme is described in the documentation provided with the database).

Calculation of ND measures

Neighborhood density was calculated in several different ways. First, we calculated a set of *segment-based* ND measures following Luce (1986) and Pisoni, Nusbaum, Luce, & Slowiaczek (1985). Two words were considered neighbors if they differed by the deletion, addition, or substitution of one and only one segment—i.e., an edit distance of one. The neighborhood relation is therefore symmetric (e.g., if /mak/ is a neighbor of /hak/, then /hak/ is a neighbor of /mak/), intransitive (e.g., although /mak/ is a neighbor of /hak/, and /hak/ is a neighbor of /han/, /mak/ and /han/ are not necessarily neighbors), and anti-reflexive (i.e., a word is not a neighbor of itself). We calculated ND using three different representations. The first two representations were the *modern* and *conservative* surface phonetic forms described above. The third representation will be referred to as *orthographic*, in which we treated the orthographic representation (in Hangeul) of each word as a linear string of jamo, instead of as arranged into syllable blocks.[8] Then, ND was calculated based on an edit distance of one jamo. Note that words that differ by only one jamo may not differ phonetically. For example, 박 ‘gourd’ /pak/ and 밖 ‘outside’ /pak*/ differ orthographically (and in their underlying phonological representation) in coda position, with the former having a lax velar stop and the latter having a tense velar stop. But because homorganic Korean stops are neutralized in coda position, both words have the same surface phonetic representation of [pak]. Thus, the modern and conservative forms can be thought of as phonetic forms, whereas the orthographic form corresponds more closely to a phonological form.

Second, we calculated a set of *syllable-based* ND measures in which two words were considered neighbors if they differed by the substitution of one and only one syllable. The syllable-based measures were also calculated based on the three representations discussed above: modern, conservative, and orthographic. For example, consider the word 나무 ‘tree’ /namu/. This word has two syllables, 나 /na/ and 무 /mu/. Its syllable-based neighbors would be all *bisyllabic* words whose first syllable is 나 /na/ *or* whose second syllable is 무 /mu/. In this case, although no phonological processes would be applied to obtain the modern (/namu/) and conservative (/namu/) representations, syllable-based ND would still differ among them. For the modern and conservative syllable-based ND, the word 낙인 ‘brand’ /nak.in/ would be considered a neighbor, since the word-medial /k/, which is a coda of the first syllable, is resyllabified as the onset of the second syllable, as in [na.kin]. For the orthographic representation, however, /nak.in/ would *not* be considered a neighbor, because the first syllable is represented orthographically as 낙 /nak/, whereas in the target word, 나무 ‘tree’ /namu/, it is 나 /na/. Note that unlike in segment-based ND, syllable-based neighbors did not include words that differed by deletion or addition. Thus, 나무꾼 ‘lumberjack’ /namuk*un/ would not be considered a syllable-based neighbor of 나무 ‘tree’ /namu/ in any of the three representations.

Table 2 Examples illustrating the result of the phonetization process

orthography	conservative	modern
가겟방	ka:ketp*aŋ	kakɛp*aŋ
간행되다	kanhɛŋtøta	kanhɛŋtwɛta
시계	si:kje	sikɛ
전의	t͡ɕʌ:nɯi	t͡ɕʌni
최고가	t͡ɕʰø:kok*a	t͡ɕʰwɛkok*a
밤일	pamnil	pamnil

[8]For this segment-based orthographic measure, the null onset ㅇ was discarded since it has no phonological status. Furthermore, note that diphthongs were treated as a single segment. This means that a sequence of vowels such as 오이, which was linearized as "ㅗㅣ", was always kept distinct from the diphthong 외, which was linearized as "ㅚ"

Results

The resulting database, K-SPAN, which includes the surface phonetic forms and accompanying ND measures, is available in Appendix. In this section, we summarize some of the salient trends in segment frequencies and ND measures calculated from the database. The trends in segment frequencies will be compared to those reported in Shin et al. (2013), who reported frequency trends from the *Yonsei Korean Language Dictionary* and the Spoken Language Information Lab Corpus (Shin, 2008).

It should first be noted that the K-SPAN database differs from both the *Yonsei Korean Language Dictionary* and the Spoken Language Information Lab Corpus in several important ways. The lexical entries in K-SPAN were taken from the MKUFS2 (Kim, 2005), which listed words in their dictionary form (i.e., stripped of any morphology), in the same way as the *Yonsei Korean Language Dictionary*. On the other hand, the entries in the MKUFS2 were gathered from a variety of sources, such as textbooks, novels, screenplays, and spoken dialogue, among others, and thus reflect actual usage. The *Yonsei Korean Language Dictionary* is an actual dictionary, however, and thus may include some very low-frequency words that could be absent from the MKUFS2. The Spoken Language Information Lab Corpus, of course, reflects actual usage, but is different from K-SPAN in that it was gathered entirely from speech and contains morphological markers that are absent in K-SPAN (e.g., the topic marker 는 or the future and conditional modals 겠 /kes*/ and 면 /mjʌn/).

Table 3 Type and token frequencies of segment type in the modern (abbreviated *m*) and conservative (abbreviated *c*) forms, calculated over all 63,836 word types (186,239 syllables)

	Type-m	Token-m	Type-c	Token-c
Vowel	186,239	6,820,133	186,239	6,820,133
	(43.0 %)	(45.8 %)	(42.7 %)	(45.2 %)
Consonant	246,800	8,077,292	250,285	8,270,010
	(57.0 %)	(54.2 %)	(57.3 %)	(54.8 %)
Consonant onset	169,889	5,939,524	169,889	5,939,524
	(91.2 %)	(87.1 %)	(91.2 %)	(87.1 %)
Empty onset	16,350	880,609	16,350	880,609
	(8.8 %)	(12.9 %)	(8.8 %)	(12.9 %)
Consonant coda	76,911	2,137,768	80,396	2,330,486
	(41.3 %)	(31.3 %)	(43.2 %)	(34.2 %)
Empty coda	109,328	4,682,365	105,843	4,489,647
	(58.7 %)	(68.7 %)	(56.8 %)	(65.8 %)

The percentages for vowels and consonants are calculated over the total number of segments. The percentages for onset and coda are calculated over the total number of syllables

Turning back to K-SPAN, the type and token frequencies of vowels and consonants in both the modern and conservative forms are given in Table 3. It can be seen that there are overall slightly more consonants than vowels, which is expected, given that a syllable may contain up to two consonants but necessarily has only one vowel. The bottom two rows of Table 3 show the number and percentage of consonants that are in onset or coda position. As expected, syllable onsets are more common than syllable codas. In addition, while syllables with a consonant onset are far more common than syllables with an empty onset, open syllables are more common than closed syllables. These results are comparable to those calculated from the corpora reported in Shin et al. (2013).

Frequency counts for individual consonants and vowels are given in Tables 4 and 5, which are sorted according to the modern form type frequency. Several trends are apparent. First, although the tense and aspirated consonants have lower type frequencies than the lax obstruents, nasals, and liquid, there are a few consonants whose type and token frequency rankings diverge markedly. Among these are the alveolar stops /t, t^h, t*/, which all have a much higher relative token frequency than type frequency. We attribute this partly to the fact that the dictionary form of all verbs and adjectives ends with /ta/, resulting in /t/ being over-represented among high-frequency words.

Table 4 Consonant type and token frequencies for the modern (m), conservative (c), and orthographic (o) forms

	Type-m	Token-m	Type-c	Token-c	Type-o	Token-o
k	30,760	1,006,400	31,838	1,024,795	36,552	1,119,477
n	27,189	926,497	27,189	926,497	26,045	914,662
ŋ	23,811	538,231	23,811	538,231	45,861	1,560,904
t	23,615	1,297,106	24,796	1,433,941	25,300	1,502,841
l	23,193	741,731	23,193	741,731	23,826	753,240
s	18,351	519,091	18,351	519,091	23,531	708,182
t͡ɕ	18,003	558,108	18,003	558,108	21,383	630,652
m	17,112	539,901	17,112	539,901	16,891	536,602
p	14,255	421,997	14,443	423,602	16,177	436,829
h	13,377	464,248	14,415	500,131	17,534	641,156
$t͡ɕ^h$	7,493	174,721	7,493	174,721	7,646	178,919
p^h	4,988	98,188	4,988	98,188	4,771	103,683
t^h	4,384	152,459	4,384	152,459	4,431	109,952
s*	3,945	80,218	3,945	80,218	691	102,139
k*	3,927	83,546	3,927	83,546	1,588	46,061
t͡ɕ*	3,909	72,499	3,909	72,499	761	17,489
t*	3,525	301,062	3,525	301,062	1,100	71,331
k^h	3,149	73,176	3,149	73,176	1,623	31,213
p*	1,814	28,113	1,814	28,113	533	14,183
Total	246,800	8,077,292	250,285	8,270,010	276,244	9,479,515

Table 5 Vowel type and token frequencies for the modern (m), conservative (c), and orthographic (o) forms

	Type-m	Token-m	Type-c	Token-c	Type-o	Token-o
a	53,043	2,514,092	53,043	2,514,092	53,049	2,513,004
i	26,745	1,050,562	26,198	1,046,243	26,019	1,042,479
ʌ	19,668	682,359	19,705	683,804	19,668	682,042
o	18,643	521,282	18,633	521,204	18,632	521,201
u	17,851	495,438	17,863	495,522	17,863	495,522
ɛ	14,384	437,187	9,535	290,076	9,537	290,085
ɨ	10,167	408,848	10,167	408,848	10,160	408,833
jʌ	8,812	262,448	8,775	261,003	8,810	262,763
wa	4,000	88,169	4,000	88,169	4,000	88,169
wɛ	3,597	105,433	194	5,582	194	5,582
jo	2,092	61,043	2,090	61,037	2,090	61,037
ju	2,040	37,412	2,040	37,412	2,040	37,412
ja	1,683	35,963	1,683	35,963	1,683	35,963
wʌ	1,406	42,432	1,406	42,432	1,406	42,432
wi	1,399	44,257	1,399	44,257	1,399	44,257
jɛ	450	17,929	14	3,223	14	3,223
ɨi	259	15,279	926	27,368	1,093	31,094
e	0	0	3,920	115,328	3,918	115,319
ø	0	0	3,305	98,308	3,307	98,329
je	0	0	1,245	38,719	1,245	38,719
we	0	0	98	1,543	94	1,502
Total	186,239	6,820,133	186,239	6,820,133	186,221	6,818,967

Depending on the coda of the preceding syllable, this /ta/ can also surface as [t*a] (when preceded by an obstruent, as in 먹다 /mʌkta/‘eat’ surfacing as [mʌkt*a]) or as [t^{h}a] (when preceded by /h/, as in 않다 /anhta/ ‘do not’ and 좋다 /tɕohta/‘good’ surfacing as [ant^{h}a] and [tɕot^{h}a]).

Second, the frequencies of the lax obstruents /k/, /t/, /p/, /tɕ/, and /s/ are all lower in the modern and conservative forms than in the orthographic forms, likely reflecting the several processes that phonetically neutralize them. For example, coda lax obstruents surface as homorganic nasals when followed by a nasal or liquid, and onset lax obstruents surface as tense when preceded by an obstruent coda. The wide application of these processes should result in a decrease in the frequency of lax obstruents and an increase in the frequency of nasals and tense obstruents when comparing orthographic forms to phonetic surface forms, and that is exactly what we see in Table 4.

Lastly, it should be noted that the difference between the modern and conservative forms does not substantially impact consonant frequencies. The only consonants whose modern and conservative frequencies differ at all are /k/, /t/, /p/, and /h/. The most common process affecting consonants was same-place deletion, in which a lax stop in a coda-onset sequence of /kk*/, /tt*/, or /pp*/ was deleted.[9] Another example was the deletion of /h/ between /n/ and /j/, such as in 균형 /kjunhjʌŋ/ ‘balance’ surfacing as [kjunjʌŋ] in the modern pronunciation.

Turning next to the vowel frequencies in Table 5, we see that the frequencies are heavily skewed, with only a few vowels accounting for the majority of counts. The most common vowel across the board is /a/, representing approximately 28 % of the type counts and 37 % of the token counts. The next most frequent vowel, /i/, is at most half as frequent. Regardless of the representation used, /a/, /i/, and /ʌ/ account for over half of the type counts, and /a/ and /i/ alone account for over half of the token counts.

Another obvious pattern in the vowel frequencies is the total absence of certain vowels in the modern forms. Specifically, the absence of /e/, /ø/, /je/, and /we/ in the modern forms reflects their neutralization with /ɛ/, /wɛ/, /jɛ/, and /wɛ/, respectively. Conversely, these neutralizations are reflected in the frequencies of /ɛ/, /wɛ/, and /jɛ/, which are comparably *higher* in the modern forms than in the conservative forms. Some of these vowels, /wɛ/ in particular, are in fact quite rare underlyingly.

Overall, the modern form type frequencies of individual consonants and vowels in the current database closely mirror those of the *Yonsei Korean Language Dictionary* as reported in Shin et al. (2013). With the exception of /t/, discussed above, the most frequent consonants and vowels are also the same, and the tense and aspirated consonants are also the least frequent across both corpora.

Finally, some summary statistics of the ND calculations are presented in Tables 6, 7 and 8. First, summary statistics of segment-based ND are given in Table 6. The first column contains the statistics for the entire database, and the columns to the right contain statistics for just the words with each corresponding number of syllables. For each of the three representations (modern, conservative, and orthographic), the range, mean, and median ND are provided, along with the percentage of words that have no neighbors (“% 0”). It can be seen that the maximum, mean, and median number of neighbors decreases with increasing syllable count, with the exception of the comparison between three- and four-syllable words. We presume this discrepancy is due to the fact that two-syllable nouns are so frequent, and many of them can take a two-syllable light verb to become a four-syllable verb or adjective (e.g., the noun 행복 /hɛŋpok/ ‘happiness’ can combine with the light verb 하다 /hata/ to become the adjective 행복하다 /hɛŋpokhata/ ‘happy’). Thus, the fact that many four-syllable words already share two of their syllables with many other words serves to counteract the general trend

[9] For example, the conservative pronunciation [tɕak.k*a] is realized as [tɕa.k*a] in the modern pronunciation

Table 6 Segment-based neighborhood density summary statistics

			Syllable count					
		Total	1	2	3	4	5	6+
	Count	63,836	1,964	24,335	19,494	14,282	2,665	1,096
Modern	Range	0–234	2–234	0–181	0–58	0–21	0–6	0–2
	Mean	9.4	95.0	15.0	1.3	1.4	0.3	0.1
	Median	2	89	10	0	0	0	0
	% 0	38.0	0	3.9	63.0	54.7	81.2	93.7
Conservative	Range	0–173	1–173	0–141	0–43	0–19	0–6	0–2
	Mean	5.7	61.3	8.7	0.9	0.9	0.3	0.1
	Median	1	56	6	0	0	0	0
	% 0	42.8	0	7.3	70.4	60.3	82.6	93.7
Orthographic	Range	0–181	1–181	0–94	0–34	0–21	0–6	0–2
	Mean	7.4	74.2	12.0	0.8	1.3	0.2	0.1
	Median	1	71	8	0	0	0	0
	% 0	41.1	0	5.5	69.7	56.4	83.5	94.5

of longer words having fewer neighbors. Nevertheless, an important conclusion to be drawn from Table 6 is that the possible range of ND can vary greatly depending on the number of syllables in the word.

It can also be seen that every one-syllable word has at least one neighbor, and more than half of all words with three or more syllables have no neighbors. Thus, it is only the set of two-syllable words that contains some words with no neighbors while the majority of words still has some neighbors. Among words with five or more syllables, having even one neighbor at all seems to be more of an exception than the rule, which suggests that research on the effects of ND in Korean may not be applicable to longer words.

Table 7 contains the same statistics calculated for syllable-based ND. Across the board, syllable-based ND tends to be higher than segment-based ND, which is expected given that syllable-based neighbors can differ by more segments than segment-based neighbors can. One result of this trend is that there exists much greater variation in ND within different syllable counts. Almost all two-syllable words, and most three- and four-syllable words, have at least one neighbor. On the other hand, there is very little variation in ND among monosyllabic words. Because syllable-based neighbors are defined as words that differ by the substitution of exactly one syllable, all monosyllablic words should be neighbors of each other. The reason ND is

Table 7 Syllable-based neighborhood density summary statistics

			Syllable count					
		Total	1	2	3	4	5	6+
	Count	63,836	1,964	24,335	19,494	14,282	2,665	1,096
Modern	Range	0–1963	1945–1963	0–747	0–183	0–138	0–15	0–4
	Mean	143.2	1959	201.2	9.4	15.0	1.3	0.2
	Median	18	1960	180	3	4	0	0
	% 0	15.8	0	0.1	14.9	33.0	54.6	88.7
Conservative	Range	0–1963	1945–1963	0–738	0–197	0–140	0–15	0–4
	Mean	138.3	1959	187.7	9.4	16.0	1.3	0.2
	Median	18	1960	165	3	4	0	0
	% 0	16.3	0	0.1	16.1	33.6	54.4	88.5
Orthographic	Range	0–1972	1954–1972	0–1972	0–270	0–179	0–17	0–4
	Mean	152.6	1968	217.4	12.3	24.1	1.6	0.2
	Median	27	1969	190	4	7	1	0
	% 0	14.5	0	0.2	12.2	31.8	49.6	87.0

Table 8 Spearman's rho correlations among the different ND metrics. *Mod*, *Cons*, and *Ortho* refer to the modern, conservative, and orthographic representations, and *Seg* and *Syll* refer to the segment- and syllable-based measures, respectively

	Mod-Seg	Cons-Seg	Orth-Seg	Mod-Syll	Cons-Syll	Orth-Syll
Entire database						
Frequency	.182	.172	.175	.201	.201	.206
Mod-Seg		.950	.948	.886	.882	.874
Cons-Seg			.927	.844	.848	.842
Orth-Seg				.880	.885	.890
Mod-Syll					.992	.969
Cons-Syll						.975
2-syllable words only						
Frequency	.062	.049	.046	.063	.056	.072
Mod-Seg		.880	.892	.790	.760	.719
Cons-Seg			.845	.670	.690	.683
Orth-Seg				.745	.772	.816
Mod-Syll					.953	.811
Cons-Syll						.850

not uniform among them is that homophones are technically not neighbors of each other, and so a word's ND is reduced by the number of homophones it has.

Finally, Table 8 reports the Spearman's rho correlation among lexical frequency and the six ND measures reported in the database. The top panel reports the correlations for the entire database. It can be seen that all of the ND measures are only weakly correlated with lexical frequency, and all of the ND measures are strongly correlated with each other. Because only the two-syllable words showed substantial variation in ND according to both the segment- and syllable-based measures, the bottom panel reports the correlations among the measures when only the two-syllable words are considered. The overall trends are similar, with frequency even more weakly correlated with ND, and the various ND measures only slightly less strongly correlated with each other.

This is not to say, of course, that these different ND measures will always pattern similarly. For example, 살해 /sal.hɛ/ 'murder' has 109 orthographic syllable neighbors but 505 modern syllable neighbors, as the modern surface form is [sa.lɛ]. On the other hand, 빻다 /p*ah.ta/ 'crush' has 556 orthographic syllable neighbors but only 58 modern syllable neighbors, as the modern surface form is [p*a.t^{h}a/]. Disparities such as these can arise when an orthographically uncommon syllable undergoes some phonological process (e.g., neutralization, resyllabification) that renders its surface form the same as a frequent syllable. Alternatively, very common orthographic syllables (such as 다 /ta/, the marker for all verbs and adjectives) can undergo some process (e.g., aspiration, in this case), that renders its surface form something far less frequent (e.g., /t^{h}a/).

Conclusions

Despite the large body of research on Korean language processing, there has been no publicly available phonetized lexical database of Korean until now. The database presented here, K-SPAN, provides surface phonetic forms derived in two different ways for 63,836 Korean words. When combined with the lexical frequencies and part of speech information provided in the MKUFS2 corpus (Kim, 2005), a wide range of useful statistics may be computed. Among these, K-SPAN itself includes six different measures of neighborhood density: both segment- and syllable-based ND calculated from modern surface phonetic forms, conservative surface phonetic forms, and orthographic representations. The availability of K-SPAN opens several avenues for future research.

First, the surface phonetic forms, instantiated here as "conservative" and "modern" pronunciations, may be used to look up the pronunciation of Korean word forms without having to consult a Korean-language dictionary. Although there exist several freely available Korean corpora, including the MKUFS2, all of them are rendered orthographically (in Hangeul). K-SPAN therefore simplifies the calculation of various statistics over the Korean lexicon, such as n-gram phoneme frequencies, since the surface phonetic forms are rendered in an ASCII scheme. Such queries would be impossible in an orthographically rendered corpus. For

example, several studies have examined the potential role of functional load, a measure of the strength of a phonological contrast, in phoneme mergers and neutralizations in Korean (Eychenne and Jang, 2015; Silverman, 2010) and across languages, including Korean (Oh, Coupé, Marsico, & Pellegrino, 2015; Wedel, Jackson, & Kaplan, 2013a; Wedel, Kaplan, & Jackson, 2013b). However, the Korean data used in these studies were "phonological" forms similar to our orthographic forms and/or forms phonetized by rules, which as we have seen often do not reflect the actual pronunciation. K-SPAN now offers a more reliable database that can be used to calculate such metrics.

Second, the ND statistics provided in K-SPAN may be used to extend studies of ND effects to Korean. For example, it remains unknown whether or how ND affects spoken language production or perception in Korean. Previous studies have suggested that the eumjeol (or syllable) may play an important role in visual word recognition, but a proper comparison between the effects of segment- versus syllable-based ND has not been possible. Similarly, it has also not been explored whether there might be any meaningful difference between ND calculated on surface phonetic forms or orthographic forms (which, in Korean, more closely reflect underlying forms). Future work may also explore the usefulness of other types of ND measures, for instance position-sensitive ND, such as the first-syllable frequency metric used in Kwon et al. (2011), or ND measures calculated within a given syntactic category rather than across the lexicon, since it has been suggested that words compete more strongly when they can be substituted for one another in the speech stream (Wedel et al. 2013a).

The current database will therefore help researchers, including those who may not be literate in Korean, to explore the Korean lexicon in greater depth, thereby widening the empirical scaffolding upon which theories of the lexicon are built.

Acknowledgments Julien Eychenne's work was supported by the Hankuk University of Foreign Studies Research Fund 2016.

Appendix

The original NIKL corpus, which is distributed under an open share-alike license, is available at the following address:
http://korean.go.kr/front/reportData/reportDataView.do?mn_id=45&report_seq=1&pageIndex=1

The raw corpus is available as a ZIP archive entitled 현대 국어 사용 빈도 조사 2.zip. Uncompressing the ZIP file will create a directory entitled 현대+국어+사용+빈도+조사+2, which contains several files in TXT, Excel, and PDF format. The relevant file, which contains the full list of lexical items from the corpus, is entitled 일반어휘통계.txt. Note that on Linux, Mac, and Windows systems with a non-Korean locale, file names may not be displayed properly. Should that be the case, the file can still be identified thanks to its size: it is the largest TXT file in the directory, about 2 megabytes. We suggest renaming this file to `nikl_original.txt`. It contains the following columns: rank, word frequency, word form, disambiguation and part of speech.

The phonetized database that was derived from the NIKL corpus (as explained in Section 3) is available from the TROLLing open data archive, at the following address:
https://opendata.uit.no/dataset.xhtml?persistentId=doi:10.18710/TWM79F

This resource contains three files. The file named `kspan_doc.pdf` describes the content of the database and provides instructions to merge the K-SPAN database with the original NIKL corpus. The file entitled `hte_base.csv` is the K-SPAN database itself, and contains the following columns: word number; the modern, conservative and orthographic forms as keystrokes; the number of neighbors and mean neighbor frequency for the orthographic, modern and conservative forms, respectively. The file is made available as a UTF-8 encoded CSV file, using the tabulation character as a field delimiter. This file does not contain any information from the original NIKL corpus; however, the number in the first column corresponds to the word number in the NIKL corpus. For example, the word number for the 8th word in this file is 10, which means that this line corresponds to the 10th word in the NIKL corpus, which is the word 가가호호. The last file, which is named `merge_corpus.py`, is the Python script that can be used to merge the NIKL and K-SPAN corpora (see instructions in `kspan_doc.pdf`).

References

Ahn, S C. (1998). *An Introduction to Korean Phonology*. Seoul: Hansin Munhwasa.

Carreiras, M, Alvarez, C J, & de Vega, M (1993). Syllable frequency and visual word recognition in Spanish. *Journal of Memory and Language*, *32*, 766–780.

Coady, J A, & Aslin, R N (2003). Phonological neighbourhoods in the developing lexicon. *Journal of Child Language*, *30*, 441–469.

Cock, P J A, Antao, T, Chang, J T, Chapman, B A, Cox, C J, Dalke, A., ... Hoon, M (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423.

Cutler, A, Mehler, J, Norris, D, & Segui, J (1986). The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language*, *25*, 385–400.

Eychenne, J, & Jang, T Y (2015). On the merger of Korean mid front vowels. *Phonetics and Speech Sciences (Journal of the Korean Society of Speech Sciences)*, *7*(2), 119–129.

Hieronymus, J L. (1994). *ASCII Phonetic symbols for the world's languages: Worldbet*: Tech. rep. AT&T Bell Laboratories.

Holliday, J J, & Turnbull, R (2015). Effects of phonological neighborhood density on word production in Korean. In *Proceedings of the Eighteenth International Congress of the Phonetic Sciences*.

Hong, Y. (1988). *A sociolinguistic study of Seoul Korean*. Seoul: Hanshin Publishing Co.

Kim, H. (2005). *Hyeondae Gugeo Sayong Bindo Josa 2*. Seoul: National Institute of the Korean Language.

Kim, H (2006). Korean national corpus in the 21st century Sejong project. In *Proceedings of the 13th National Institute of Japanese Literature (NIJL) International Symposium*, (pp. 49–54).

Kwon, Y (2014). The syllable type and token frequency effect in naming task. *Korean Journal of Cognitive Science*, *25*, 91–107.

Kwon, Y, & Nam, K (2011). The relationship between morphological family size and syllabic neighborhoods density in Korean visual word recognition. *The Korean Journal of Cognitive and Biological Psychology*, *23*, 301–319.

Kwon, Y, Lee, C, Lee, K, & Nam, K (2011). The inhibitory effect of phonological syllables, rather than orthographic syllables, as evidenced in Korean lexical decision tasks. *Psychologia*, *54*, 1–14.

Lee KM, & Ramsey SR. (2011). *A history of the Korean language*: Cambridge University Press.

Luce, P A. (1986). *Neighborhoods of words in the mental lexicon*: PhD thesis, Indiana University.

Luce, P A, & Pisoni, D B (1998). Recognizing spoken words: the neighborhood activation model. *Ear & Hearing*, *19*(1), 1–36.

Mehler, J, Dommergues, J Y, Frauenfelder, U, & Segui, J (1981). The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior*, *20*, 298–305.

Munson, B, & Solomon, N P (2004). The effect of phonological neighborhood density on vowel articulation. *Journal of Speech, Language, and Hearing Research*, *47*, 1048–1058.

Oh, Y M, Coupé, C, Marsico, E, & Pellegrino, F (2015). Bridging phonological system and lexicon: insights from a corpus study of functional load. *Journal of Phonetics*, *53*, 153–176.

Perea, M, & Carreiras, M (1998). Effects of syllable frequency and syllable neighborhood frequency in visual word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 134–144.

Pisoni, D B, Nusbaum, H C, Luce, P A, & Slowiaczek, L M (1985). Speech perception, word recognition and the structure of the lexicon. *Speech Communication*, *4*, 75–95.

Scarborough, R. (2004). *Coarticulation and the structure of the lexicon*. Los Angeles: PhD thesis, UCLA.

Shin, J (2008). Phoneme and syllable frequencies of Korean based on the analysis of spontaneous speech data. *Korean Journal of Communication Disorders*, *13*(2), 193–215.

Shin, J, Kiaer, J, & Cha, J. (2013). *The sounds of Korean*. Cambridge: Cambridge University Press.

Silverman, D (2010). Neutralization and anti-homophony in Korean. *Journal of Linguistics*, *46*(02), 453–482.

Sohn, H M. (1999). *The Korean language*: Cambridge University Press.

Song, J, Nam, K, & Koo, M (2012). The effect of word frequency and neighborhood density on spoken word segmentation in Korean. *Journal of the Korean Society of Speech Sciences*, *4*(2), 3–20.

Stokes, S F (2010). Neighborhood density and word frequency predict vocabulary size in toddlers. *Journal of Speech, Language, and Hearing Research*, *53*, 670–683.

The Unicode Consortium (2015). The Unicode Standard, Version 8.0.0. The Unicode Consortium, http://www.unicode.org/versions/Unicode8.0.0/.

Vitevitch, M S, & Stamer, M K (2006). The curious case of competition in Spanish speech production. *Language and Cognitive Processes*, *21*, 760–770.

Wedel, A, Jackson, S, & Kaplan, A (2013a). Functional load and the lexicon: evidence that syntactic category and frequency relationships in minimal lemma pairs predict the loss of phoneme contrasts in language change. *Language and Speech*, *56*(3), 395–417.

Wedel, A, Kaplan, A, & Jackson, S (2013b). High functional load inhibits phonological contrast loss: a corpus study. *Cognition*, *128*(2), 179–186.

Wright, R (2004). Factors of lexical competition in vowel articulation. In Local, J, & Ogden, R (Eds.) *Papers in Laboratory Phonology*, (Vol. 6, pp. 26–50). Cambridge: Cambridge University Press.